SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(131th Edition)

NOVEMBER 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 Thirty First** 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 **November, 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. 2D router chip design, analysis, and simulation for effective communication, Prateek Agarwal, ***7.1.Tanuj Kumar Garg*** and Adesh Kumar, Electronics, DTU

2. A comparative study of combined cycles for concentrated solar power for efficient power generation using low Global Warming Potential (GWP) fluids to reduce environmental effects, ***6.Yunis Khan***, Hakan Caliskan & Hiki Hong, Mechanical, DTU

3. A Comprehensive Review on Permissions-Based Android Malware Detection, ***8.Yash Sharma*** and ***3.Anshul Arora***, Applied Mathematics, DTU

4. A Decade of Exploration of Transition-Metal-Catalyzed Cross-Coupling Reactions: An Overview, ***8.Saurav Kumar***, ***8.Jyoti***, ***8.Deepak Gupta***, Gajendra Singh and ***6.Anil Kumar***, Applied Chemistry, DTU

5. A novel approach for forecasting PM2.5 pollution in Delhi using CATALYST, ***6.Abhishek Verma***, ***3.Virender Ranga*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

6. A Novel Real-Time Parametric Tracking Approach for Robust Microwave Filter Tuning, Even Sekhri, Mart Tamre, ***3.Rajiv Kapoor*** and ***8.Rahul Kumar***, ECE, DTU

7. A recurrent neural network-based identification of complex nonlinear dynamical systems: a novel structure, stability analysis and a comparative study, ***6.R. Shobana***, Rajesh Kumar and ***3.Bhavnesh Jaint***, Electrical, DTU

8. A survey on intelligent human action recognition techniques, ***6.Rahul Kumar*** and ***3.Shailender Kumar***, CSE, DTU

9. Analytical Modelling of Inhomogeneous Energy Density Driven (IEDD) Instability in a Magnetized Dusty Plasma Cylinder, ***6.Anshu***, Jyotsna Sharma and ***3.Suresh C. Sharma***, Applied Physics, DTU

10. Application of Gene Expression Programming in Computation of Flow Resistance in Compound Channel with Converging Floodplains, ***6.Rahul Kumar***, ***6.Vijay Kaushik*** and ***3.Munendra Kumar***, Civil, DTU

11. Application of Response Surface Method and Genetic Algorithm in the Design of High-Efficiency Prototype Vehicle, ***8.Paras Singh***, ***8.Harshit Gupta***, ***8.Ojas Vinayak*** and ***8.Aryan Tyagi***, Mechanical, DTU

12. Bibliometric review of application of blockchain in supply chain management, ***8.Akash Mishra***, ***8.Ayush Saini***, ***8.Ansh Jini*** and ***3.N. Yuvraj***, Mechanical, DTU

13. Bit-plane based Reversible Data Hiding in Encrypted Images using Multi-Level Blocking with Quad-Tree, ***6.Ankur***, ***3.Rajeev Kumar*** and Ajay K Sharma, CSE, DTU

14. Comparative Study on Traditional Timber Bridges of Bhutan with Timber Bridges of Other Nations, ***3.Raju Sarkar***, ***3.T. Vijaya Kumar***, Kiran Gurung, Dipti Sharma, Tashi Choki, Kezang Wangmo, Kezang Tshomo, Sonam Penjor, Abhirup Dikshit and Ritesh Kurar, Civil, DTU

15. ConvST-LSTM-Net: convolutional spatiotemporal LSTM networks for skeleton-based human action recognition, ***3.Abhilasha Sharma*** and ***6.Roshni Singh***, Software Engineering, DTU

16. Development and assessment of a novel integrated system powered by parabolic trough collectors for combined power, heating and freshwater production, ***6.Mohd Asjad Siddiqui*** and Eydhah Almatraf, Mechanical, DTU

17. Dielectric behavior and impedance spectroscopy of Niobium substituted Lanthanum based orthovanadates at high temperatures, ***6.Ashok Kumar***, Vikas N. Thakur, Ajay Kumar, ***3.Vinod Singh***, Anita Dhaka and Rajendra S. Dhaka, Applied Physics, DTU

18. DO Modeling for the Yamuna River, Delhi, ***6.Nibedita Verma***, ***3.Dr. Geeta Singh*** and Prof. Naved Ahsan, Environmental, DTU

19. Dual metal dual layer GAA NW–FET (DMDL–GAA–NW–FET) biosensor for label free SARS-CoV-2 detection, ***6.Shivani Yadav*** and ***3.Sonam Rewari***, ECE, DTU

20. eDRAM-CIM: Reconfigurable Charge Domain Compute-In-Memory Design With Embedded Dynamic Random Access Memory Array Realizing Adaptive Data Converters, Shanshan Xie, Can Ni, ***8.1.Aseem Sayal***, Pulkit Jain, Fatih Hamzaoglu, and Jaydeep P. Kulkarni, Electronics, DTU

21. Efect of varying ambient conditions on the performance of air conditioner using evaporative cooler, ***6.S. K. Gupta***, ***3.B. B. Arora*** and ***3.A. Arora***, Mechanical, DTU

22. Efficient Design and Implementation of Matrix Multiplication, ***6.Shivam Kumar*** and ***3.Jeebananda Panda***, Electronics, DTU

23. Emerging Trends in Microfluidics Based Devices, ***6.Shipra Solanki***, ***3.Chandra M Pandey***, ***3.Rajinder K Gupta***, ***3.Bansi D Malhotra***, Biotechnology and Applied Chemistry, DTU

24. Energy transfer mechanisms and color-tunable luminescence of Tm3+/Tb3+/Eu3+ co-doped Sr4Nb2O9 phosphors for high-quality white lightemitting diodes, Ravina, Kanishk Poria, Mukesh K. Sahu, A. Kumar, ***6.Anu***, Sajjan Dahiya, Nisha Deopa and ***3.A. S. Rao***, Applied Physics, DTU

25. Evaluating public service value within collaborative e-governance: a study in the Indian context, ***3.Jitendra Kumar Pandey***, DSM, DTU

26. Fabrication and evaluation of stimuli-sensitive xanthan gum-based hydrogel as a potential carrier for a hydrophobic drug ibuprofen, ***6.Priyanka Meena***, ***3.Poonam Singh*** and ***3.Sudhir G. Warkar***, Applied Chemistry, DTU

27. Frugal Design: An Empirical Conceptualization and Measurement, ***6.Anuradha Kumari***, ***3.Dr. Ravindra Singh*** and ***3.Prof. Lalit Kumar Das***, Design, DTU

28. Growth responses of Chlorella sp. to some selected variants of culture medium and in effluents of a Brewery, Debadyuti Das, Sukhamay Maity, Debajyoti Kundu, Bana B. Jana, Suman Pandey, Jatindra N. Bhakta & ***3.Jaigopal Sharma***, Biotechnology, DTU

29. HateDetector: Multilingual technique for the analysis and detection of online hate speech in social networks, ***7.Anjum*** and ***3.Rahul Katarya***, CSE, DTU

30. Hybrid ABC and black hole algorithm with genetic operators optimized SVM ensemble based diagnosis of breast cancer, ***3.Indu Singh***, K. G. Srinivasa, ***8.Mridul Maurya***, Aditya Aggarwal, Himanshu Sheokand, Harsh Gunwant and ***8.Mohit Dhalwal***, CSE, DTU

31. Improving speech command recognition through decision-level fusion of deep filtered speech cues, ***6.Sunakshi Mehra***, ***3.Virender Ranga*** and ***3.Ritu Agarwal***, IT, DTU

32. Limited Data, Unlimited Potential: A Study on ViTs Augmented by Masked Autoencoders, Srijan Das, ***8.Tanmay Jain***, Dominick Reilly, Pranav Balaji, Soumyajit Karmakar, Shyam Marjit, Xiang Li, Abhijit Das & Michael Ryoo, CSE, DTU

33. Microbial inoculants improve growth in Zea mays L. under drought stress by up-regulating antioxidant, mineral acquisition, and ultrastructure modulations, ***6.Madhulika Singh***, ***3.Jai Gopal Sharma*** and Bhoopander Giri, Biotechnology, DTU

34. Morphology of Finished Brass Surface by Thermal Additive Centrifugal Abrasive Flow Machining Process Using Novel Electrode, ***6.ANANT BHARDWAJ***, ***3.KROVVIDI SRINIVAS*** and ***3.RAJIV CHAUDHARY***, Mechanical, DTU

35. Novel orange-red-emitting Li2Ba5(WO5)3:Eu3+ phosphor for w-LEDs and non-contact thermometry applications, ***6.Anu*** and ***3.A. S. Rao***, A Physics, DTU

36. Novel role of folate (vitamin B9) released by fermenting bacteria under Human Intestine like environment, ***6.Sharda Nara***, Gulshan Parasher, ***3.Bansi Dhar Malhotra*** and Manmeet Rawat, Biotechnology, DTU

37. Numerical simulation of free and submerged hydraulic jump over trapezoidal and triangular macroroughness, ***6.Harshit Kumar Jayant*** & ***3.Bharat Jhamnani***, Civil, DTU

38. Numerical study of submerged hydraulic jumps over triangular macroroughnesses, ***6.Harshit Kumar Jayant*** and ***3.Bharat Jhamnani***, Civil, DTU

39. On a Novel Model for ECG Signals and Its Statistical Properties, AYAN KUMAR DUTTA, SHIV DUTT JOSHI AND ***7.1.BREJESH LALL***, Electronics, DTU

40. On Transfer of Adversarial Robustness from Pretraining to Downstream Tasks, Laura F. Nern, ***8.Harsh Raj***, Maurice Georgi and Yash Sharma, IT, DTU

41. Optimization of Inverted Double-Element Airfoil in Ground Effect using Improved HHO and Kriging Surrogate Model, ***8.Paras Singh***, ***8.Arun Ravindranath***, ***8.Aryan Tyagi***, ***8.Aryaman Rao*** and ***3.Raj Kumar Singh***, Mechanical and Electrical, DTU

42. Optimizing Construction Waste Recycling: Strategies, Technologies, and Environmental Impacts, ***6.Yonatan Ayele Abera***, Civil, DTU

43. Optimizing Electric Vehicle Efficiency with Real-Time Telemetry using Machine Learning, ***8.Aryaman Rao***, ***8.Harshit Gupta***, ***8.Parth Singh***, ***8.Shivam Mittal***, ***8.Utkarsh Singh*** and ***3.Dinesh Kumar Vishwakarma***, Electrical, Mechanical, Mathematics and IT, DTU

44. Optimizing Sentiment Analysis: A Cognitive Approach with Negation Handling via Mathematical Modelling, ***6.Neha Punetha*** and ***3.Goonjan Jain***, A Mathematics, DTU

45. Organic Farming in India and its Way Forward, ***8.N. Krithika***, ***7.Rishabh Jain***, ***6.Monika Sharma*** and ***3.Rajeev Kumar Mishra***, Environmental, DTU

46. Parametrical Analysis and Thermal Sensitivity Study of 8-bit Vedic Multiplier, ***6.Hemanshi Chugh*** and ***3.Sonal Singh***, Electronics, DTU

47. Parametrization of the deceleration parameter in a flat FLRW universe: constraints and comparative analysis with the 3CDM paradigm, ***7.Himanshu Chaudhary***, Saadia Mumtaz, Amine Bouali, Ujjal Debnath and G. Mustafa, A Mathematical, DTU

48. Performance enhancement and environmental analysis of vapor compression refrigeration system with dedicated mechanical subcooling, Naveen Solanki, ***3.Akhilesh Arora*** and ***3.Raj Kumar Singh***, Mechanical, DTU

49. PM2.5 Observations of Outdoor and Indoor Environment at Dwarka, New Delhi, India, ***6.S. Malik*** and ***3.A. S. Rao*** and S. K. Dhaka, Applied Physics, DTU

50. Potential assessment of solar industrial process heating and CO2 emission reduction for Indian cement industry, ***6.Niranjan Sahoo*** and ***3.Anil Kumar***, Mechanical and Environment, DTU

51. Recent advances in the discipline of text based affect recognition, ***3.Rajiv Kapoor***, ***7.Manali Bhat***, ***8.Nikhil Singh*** and ***8.Aarchishya Kapoor***, Electronics, DTU

52. Reliability analysis of a CNT-TF-FinFET for hostile temperature, Praween Kumar Srivastava, Atul Kumar and ***6.1.Ajay Kumar***, Electrical, DTU

53. Secure chaotic image encryption method using random graph traversal and three step diffusion, ***8.1.Varun Agarwal*** and ***3.Dhirendra Kumar***, A Mathematics, DTU

54. Study of vapor compression refrigeration system with suspended nanoparticles in the low GWP refrigerant, Uma Shankar Prasad, ***3.Radhey Shyam Mishra*** and Ranadip Kumar Das, Mechanical, DTU

55. Synergistic effect of Al2O3–40%TiO2 coating on thermal conductivity and corrosion rate of SS 304 substrate, ***6.DEEPAK KUMAR***, ***3.PUSHPENDRA SINGH***, ***3.QASIM MURTAZA*** and R S WALIA, Mechanical, DTU

56. Synthesis, characterization and application of Lignosulphonate-gpoly(sodium acrylate) hydrogel, ***6.Manu***, ***3.Rajinder K. Gupta*** & ***3.Devendra Kumar***, A Chemistry, DTU

57. Thermally stable multi-color emitting Dy3+/Eu3+ co-doped BaO–ZnO–Li2O–P2O5 glasses for w-LEDs, ***6.Kartika Maheshwari***, Aman Prasad, Ravita Pilania, Shaik. Mahamuda, Yasha Tayal and ***3.A. S. Rao***, Applied Physics, DTU

58. Thermodynamic Analysis of Single-effect LiBrH2O Absorption Refrigeration System, ***3.Rajesh Kumar*** and ***6.Ravindra Kannojiya***, Mechanical, DTU

59. Uncovering visual attention-based multi-level tampering traces for face forgery detection, ***8.Ankit Yadav***, ***8.Dhruv Gupta*** and ***3.Dinesh Kumar Vishwakarma***, IT, 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*

International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 12, No. 3, December 2023, pp. 225~235
ISSN: 2252-8776, DOI: 10.11591/ijict.v12i3.pp225-235

2D router chip design, analysis, and simulation for effective communication

Prateek Agarwal[1], Tanuj Kumar Garg[1], Adesh Kumar[2]
[1]Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, Gurukula Kangri (Deemed to be University), Haridwar, India
[2]Department of Electrical and Electronics Engineering, School of Engineering, University of Petroleum and Energy Studies, Dehradun, India

Article Info

Article history:

Received Oct 2, 2022
Revised Nov 28, 2022
Accepted Mar 4, 2023

Keywords:

Data packets
Field programmable gate array
Network on chip router
Simulation
Xilinx integrated system environment

ABSTRACT

The router is a network device that is used to connect subnetwork and packet-switched networking by directing the data packets to the intended IP addresses. It succeeds the traffic between different systems and allows several devices to share the internet connection. The router is applicable for the effective commutation in system on chip (SoC) modules for network on chip (NoC) communication. The research paper emphasizes the design of the two dimensional (2D) router hardware chip in the Xilinx integrated system environment (ISE) 14.7 software and further logic verification using the data packets transmitted from all input/output ports. The design evaluation is done based on the pre-synthesis device utilization summary relating to different field programmable gate array (FPGA) boards such as Spartan-3E (XC3S500E), Spartan-6 (XC6SLX45), Virtex-4 (XC4VFX12), Virtex-5 (XC5VSX50T), and Virtex-7 (XC7VX550T). The 64-bit data logic is verified on the different ports of the router configuration in the Xilinx and Modelsim waveform simulator. The Virtex-7 has proven the fast-switching speed and optimal hardware parameters in comparison to other FPGAs.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Adesh Kumar
Department of Electrical and Electronics Engineering, School of Engineering
University of Petroleum and Energy Studies
Dehradun, India
Email: adeshmanav@gmail.com

1. INTRODUCTION

Network on chip (NoC) [1] has grown in prominence as various communication infrastructures have expanded. Other communication designs had been used in the past, but as core counts increased, these architectures turned out to be a bottleneck for upcoming systems on chips (SoCs) because they limited scalability, reusability, and difficulty. In the age of technology, the very large scale integration (VLSI) industry found a model known as SoC [2] and later uncovered many flaws in it, including underutilization of cores, low reusability, enormous complexity, and restricted scalability have been discovered a novel model termed NoC addresses the shortcomings of SoC. Individual effective on-chip communication design for SoC is called NoC. On a single chip, integrated circuits (IC) house multiple central processing unit (CPU) cores, memory, hardware cores, and analog components. In great capacity and high-end uses like aerospace, multimedia and defense, wireless, and wired communication schemes, these SoC are frequently employed [3]. Additional processors are uniting on a particular die as an IC technology scales and creating multiprocessor systems on chips (MPSoC).

The MPSoC technology is used to integrate complicated heterogeneous systems into a distinct chip and to allow on-chip communication. The processor elements of the MPSoC communicate using a variety of bus structures and communication methods. In terms of cost, performance, frequency of operation, and power requirements, the bus-based communication system is not a scalable architecture or a solution for future SoC design engineers [4]. The answer to the issues facing the semiconductor industry is NoC, which can meet the increased transistor packing density, cost, greater operating frequency, performance, and shorter time to market.

NoC is the system version of MPSoC that has been confirmed to be a unique and effective method for utilizing interconnections, and it inter-exchanges data among numerous nodes unified on a distinct chip. In the design of an NoC architecture, topology, switching, and routing techniques are critical. The NoC can be built and organized using different topologies in terms of the core and router computer architecture, as well as the processes required to implement the flow management, switching, and routing strategy. Data flow control is concerned with the quantity of data transportation or strength in the channels and routers. Routing is a methodology for defining the greatest pathway for the data and messages from the communicating device to the proposed end or the receiver. The two dimensional (2D) mesh NoC is a typical NoC architecture due to system scalability and the use of a simple routing approach. Long-distance traffic, on the other hand, may have severe broadcast dormancy. According to Pang *et al.* [5], star-type NoC has been used for long-distance transportation to pass through a second-level network. In the design of a NoC, the routing procedure is the content of the network layer. It is the most important factor that affects NoC network communications.

Small area communication or local area networks (LAN) are found to be suited for 2D-NoC architecture, whereas big area networks are not. The higher number of nodes can be considered while conducting a study as part of this ongoing endeavor. When transferring data across nodes, additional features like data security through encryption and decryption can be used. If we can combine system safety measures with via-node data transmission, it will be appropriate for broad communication systems and wireless data transfer. Every electronic device nowadays, from a little mobile phone to a satellite, contains SoC. The SoC has evolved quickly throughout time and is still processing information quickly. However, the devices are scaling down quickly because of the tremendous growth of the semiconductor industry. As a result, today's SoC is communication-centric. However, due to their inability to keep up with the rate at which devices are being scaled down, the present bus designs, which consist of wires for worldwide interconnection in SoC scheme, are experiencing proposal emergencies. NoC is a developing pattern that is quickly rising to the top of the competition to replace traditional bus designs as a reaction to this crisis.

The difficulty of SoC communication integration on-chip die is one that the NoC design and field programmable gate array (FPGA) implementation must overcome. With the continual reduction in chip dies size, the likelihood of failure in NoC is rising [6]. A fixed NoC can become irregular as a result of arbitrary placements, which may also impair the uniformity of actual topological design. Since throughput, area, power, and latency are the four most important factors in chip design [7], the intercommunication between the many cores and intellectual property (IP) modules on a single chip may have an impact on the device's communication and system performance [8].

The problem statement of the work is to design the hardware chip for the 2D router for communication using different ports and analyze the performance on different FPGAs for effective communication. Most of the work is done to synthesize the chip parameters on the FPGA board. The research work examines the performance of various FPGAs to assist chip designers in selecting the FPGA that has the best hardware and timing performance. The organization of the manuscript is as follows: section 2 details the related work, and section 3 presents the research method. The results and discussions are given in section 4, followed by the conclusions in section 5.

2. RELATED WORK

The work on the 2D NoC router has been focused on mesh topological NoC (4×4). The NoC nodes are addressed using the XY routing approach [9]. It is one of the most essential strategies for addressing nodes and routing data packets between nodes or processing elements inside NoC. The NoC design was formed with the very high-speed IC hardware description language (VHDL) programming language and simulated with the Xilinx 14.2 and Modelsim software. The logic routing and synthesis were done on the Virtex-5 FPGA [10] for 8-bit, 16-bit, 32-bit, 64-bit, and 128-bit data transmission across nodes in the NoC which was properly configured and implemented. FPGA synthesis settings were assessed by the hardware parameters and completely reported using the hardware report synthesized for the NoC chip. The 5-port optical router was detailed [11]. The chip design was focused on mesh and ring topological network modeling, simulation, and synthesis. The cluster size was regarded as (2×2), (4×4), (8×8), (16×16), (32×32), (64×64), (128×128), and (256×256) for 2D NoC router design [12]. The projected design decreased the

quantities of data frames navigated for long-distance communication, according to simulation results. The performance was gained of 17.2% and 10.3% respectively for standard 2D mesh and level 2 mesh topologies and provided by (12×12) star type NoC. When the network grows larger, the star-type NoC's performance improves still more while consuming less space and power. The star-type NoC future work will be included the development of novel routing algorithms and system designs for adaptive routing in arrange to better balance traffic loads and get better overall performance.

Different studies have been done related to NoC design for different routing approaches, as well as a system-based NoC simulator [13]. The core concept of the NoC simulator, design, and settings was detailed for the test strategy of NoC system planning. The investigation was done for the additional on-chip routing strategies, such as a novel linear programming (LP) routing strategy. The local traffic showed a significant improvement in performance for NoC systems, according to the findings. It was identified that LP routing can handle locality traffic without causing conflict in nearly a third of the cases. If the network system has excessed resources for sharing, such as connections, and channels, and the immediate rate of flit production is less than or equal to the volume for accepting flits, LP routing is more successful. The simulations of application-specific NoC topologies for audio video (AV) benchmark, MPEG-4 decoder, and applications demonstrated that regular NoC outperforms irregular NoC systems in the terms of performance. The simulation was following eight digital signal processing (DSP) connectivity architectures in a chip topology. The fundamental NoC concept and architecture are used to configure the topology structure in which distinct routing node architecture was followed for arbitration, routing, and data packet format. Moreover, (2×4) interconnection architecture was designed based on routing node coding, routing algorithm, and node degree routing direction. The work was carried out using active-HDL software programming and simulation. The data can continue to function without interruption to increase data transfer efficiency. After synthesis, the maximum frequency can reach 268 MHz, which can be utilized as a benchmark for the next research. Kumar *et al.* [14] focused on the architecture and network interface (NI) design performance examination and evaluation. This research is based on a 2D 4×4 topology 2D mesh. On the NIRGAM simulation, we compare the packet latency and throughput of the two algorithms using the XY and odd-even (OE) routing algorithms, respectively. NIRGAM is a software tool based on system 'C' and a cycle-accurate behavior simulator that provides the platform to realize the routing algorithms and NoC applications for different topologies. The findings show that for the XY routing algorithm, the packet delay in the X-dimensional direction is greater than the packet delay in the Y-dimensional direction and the X-dimensional path has higher average throughput. The packet delay in the X-dimensional direction is lower than in the Y-dimensional direction and the X-dimensional direction's average throughput is roughly equivalent to the Y-dimensional direction. To determine if the algorithm is good or terrible, the average throughput to average packet latency ratio is used. The better the performance, the higher the ratio. The results show that the OE routing method has a ratio of 2.5358, whereas the XY routing algorithm has a ratio of 2.1126. As a result, the OE routing method performs better in the 2D mesh topology [15].

Hardware chips are used to implement the 2D, 8×8, and 16×16 mesh topological networks. The ability to spot faulty nodes and replace them is the programmable structure's main benefit. The X and Y axis crossbar address generation mechanism can be used to identify problematic nodes. The device utilization findings indicate a 2% change in the number of slices, a 1% change in the number of slice flip-flops, a 1% change in the number of 4 input look-up tables (LUTs), and a 1% change in the number of bonded input/output blocks (IoBs) when the NoC cluster configuration changes from an (8×8) to a (16×16) NoC. For both networks, the amount of memory used is 58,820 kB and 71,466 kB, respectively. Small-area communication or LAN networks are found to be suited for 2D-NoC architecture, whereas big-area networks are not. The higher number of nodes can be considered while conducting a study as part of this ongoing endeavor. When transferring data across nodes, additional features like data security through encryption and decryption can be used. If we can combine network security methods with node data transfer, it will be suitable for wireless data networks and long-distance communications [16]. In response to this dilemma, the emergent paradigm of NoC is rapidly winning the race to supplant conventional bus designs. To address varied chip design requirements and routing techniques, numerous NoC topologies have been created. Regular-size processing devices can be included on a single chip with a mesh topology. Additionally, compared to other routing protocols, it is much easier to create and implement different routing protocols into it. Therefore, crossbar architecture has been created in VHDL and simulated in Xilinx ISE 14.1 to use the Xilinx XC5VLX30-3 FPGA to evaluate the operation of NoC on hardware [17]. Routers are based on the wormhole switching principle. Round-robin routing is used to create routers on FPGA platforms. The suggested router confronts the FIFO status of input ports during each clock cycle and the priority of each input port is constantly adjusted. The architecture described above makes sure that each input port is treated fairly. Utilizing Xilinx tools, the proposed architecture was created in VHDL. The design's functionality has been confirmed and its implementation in ISE design suite 14.1 has been completed.

The flat tree NoC architecture, supported by eight nodes at the root, was reported in the study. The performance of the system was assessed using the Virtex-5 FPGA, xc5vlx20t-2-ff323 device. Several test cases are used to experience the design. The data flow for the same cluster setup is seen on the Modelsim waveform. Additionally, the FPGA's timing and hardware support settings are examined. The architecture is scalable and modular in design. It can be used to create a massive tree in which several nodes can communicate with each other through their parents and root nodes [18]. The upcoming study may focus on network security techniques and their integration with NoC configurations for big clusters [19]. When the rotator on chip (RoC) or ring NoC chip has been created and modeled in Xilinx 14.2 software, Modelsim 10.1 b software is used to functionally simulate it. A data packet's intercommunication is confirmed when it reaches its destination and the NoC chip can link to up to 6,553 nodes. With FIFO logic integrated, a token ring concept is used to indicate the priority of communicating nodes at the same instance. Node addresses are utilized to identify the nodes and Xilinx synthesis results are used to determine hardware attributes such as the number of logic gates, LUTs, memory use, and the minimum-maximum times required to route packets [20]. The Virtex-5 FPGA is used to validate the results through data transfer between nodes. The outcomes show a superior option to currently used ring setup up structures. To create a ring topological network's programmable and adaptable structure, a large amount of work has been put into NoC design and FPGA synthesis. Future configurations of the same structure could include more nodes. Additionally, the concept of cryptographic techniques for encryption and decryption can be used to move data between nodes. Modelsim 10.0 software with 8, 16, 32, 64, and 128-bit data transfer successfully replicates the functionality of the 3D multilayer mesh NoC design produced in Xilinx 14.2 using VHDL. The performance hardware and time parameters are also assessed for mesh nodes with cluster sizes of (2×2×2), (3×3×3), and (4×4×4). The selection of the node effectively utilizes XYZ routing and addressing. The created chip works at a 735.00 MHz frequency. For larger networks like wireless sensor networks, the scalable design establishes inter-node communication. It is also concentrated on the integration of network security into hierarchical NoC architectures with multilayer systems environments.

Machine learning models [21] for multiple NoC topologies synthesized on Vitex-5 FPGA for various cluster configurations are constructed and validated. The described NoC structure may be advantageous for future MPSoC that seek high performance and fault tolerance from their communication topology. The programmable structure is a reconfigurable MPSoC solution that permits online configuration and guarantees that minute faults do not affect system performance. Mesh NoC is scalable and based on XY routing [22]. When comparing the two, the ring NoC has wider coverage. The parent-child (root) memory architecture and optimal routing technique form the basis of the tree NoC architecture. The design is more reliable for NoC and reconfigurable telecommunication networks and more suited for mobile signaling technologies like SS7. Modelsim 10.1b is used to verify data transfer while the resulting design is tested on an FPGA for the various test scenarios. Data on hardware usage, such as the number of slices, flip-flops, input LUTs, bounded IoBs, and gated clocks (GCLKs) used to implement the design, are included in the synthesis report that is generated. In Xilinx ISE 14.7 and Modelsim 10.0, the study project effectively builds 2D mesh (4×4) and 3D mesh (3×3×3) NoCs. The nodes in the 2D and 3D mesh NoC are addressed using the XY and XYZ routing. FPGA is a scalable device that offers a synthesis environment with various NoC cluster sizes. The FPGA is set up with a mesh topology structure that allows 2D mesh (16×16) and three dimensional (3D) mesh (8×8×8) designs with 256 nodes capable of communication (N=2, 4, 8, 16, 32, 64, and 128, respectively). Timing parameters and hardware resource usage are crucial factors for semiconductor businesses. The simulation work presented in the manuscript, will help NoC designer to estimate the resources utilization before developing the actual device by taking into consideration the known hardware design aspects, memory utilization, and timing factors to produce large-scale NoC.

3. RESEARCH METHOD

The data packets are transferred and communicated by the router. It is happening using the networking device that performs the traffic direction over the communication network. The data packets are sent over the networks to get effective communication such as in optical fibers, telecommunication, and high-speed switching [23]. The design of a 2D network route is shown in Figure 1 in which the 2D router is configured using the East_data_input, West_data_input, North_data_input, South_data_input, and Local_data_input. The corresponding outputs are East_data_output, West_data_output, North_data_output, South_data_output, and Local_data_output respectively for the inputs.

The data flow in packets is shown in Figure 2. This is bidirectional communication and the data communication mechanism is based on packet switching [24] and the flow control for NoC applications [25]. The data input to the crossbar switch is through all 5 ports defined for east, west, north, south, and local direction and corresponding outputs using the control module. The objective of the flow control is to use the

effective reasons for the maximum throughput and good performance that depends on the channel bandwidth and buffering capacity for the allocated network. In switching the flow control is maintained as bufferless using several ports. The quality of the source is maintained by the throughput of the network. The arbiter is also used to decide which switch port needs the process in the crossbar switch [26]. The router is used in different sensor networks [27] and wireless communication technologies such as ZigBee [28] technology.

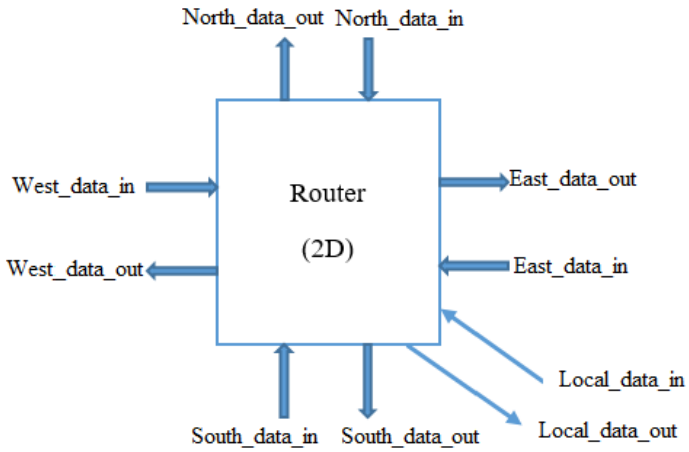


Figure 1. 2D NoC router [23]

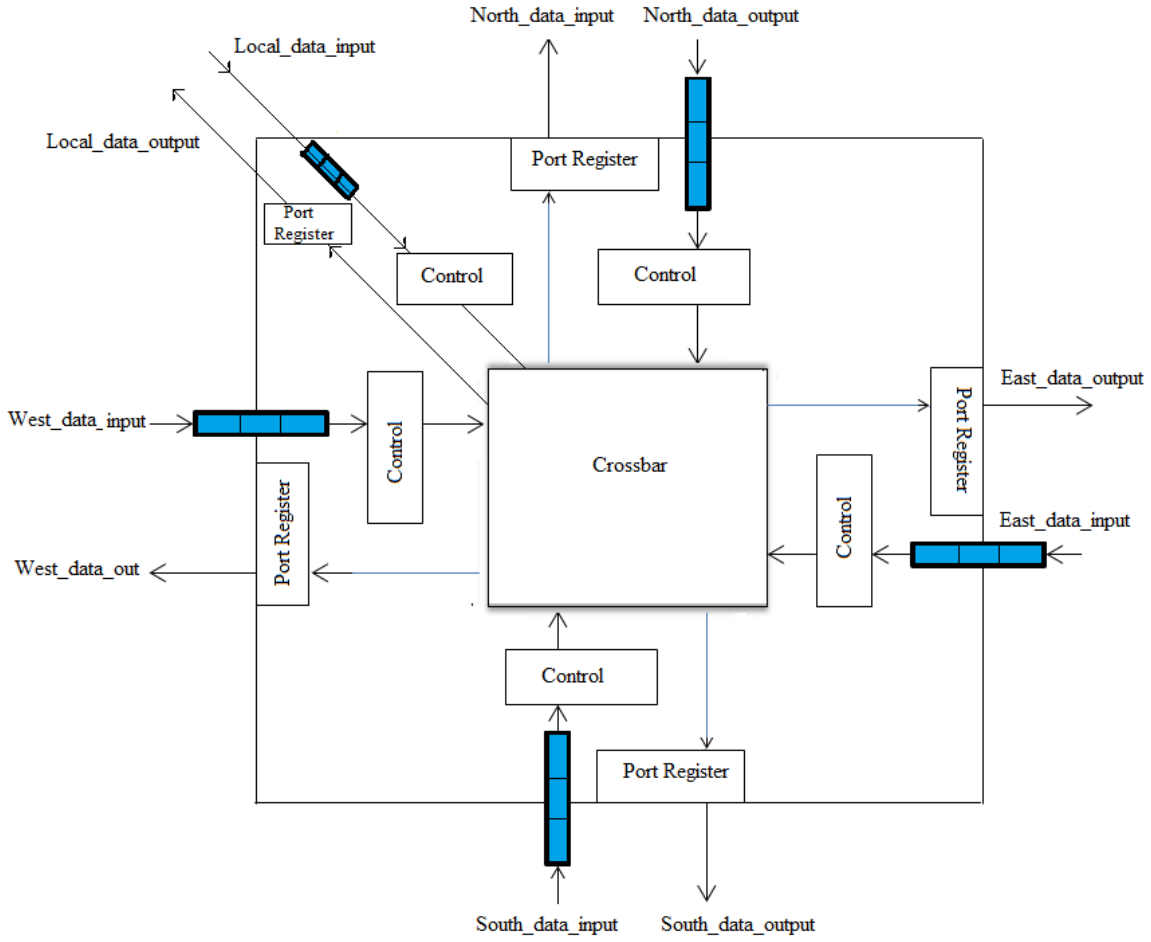


Figure 2. NoC router data packets transmission

4. RESULTS AND DISCUSSION

The chip design of the 2D-NoC router is done in Xilinx ISE 14.7 and depicted in Figure 3. The register transfer level (RTL) detail and pins functionality is given in Table 1 which provides the compressive use and application in the design strategy. The internal schematics of the design are also given in Figure 4. The Modelsim simulation waveform for the router functionality of 64-bit data communication is verified for the test inputs given. Figure 5 shows the router's waveform simulation using different binary ports. Figure 6 shows the waveform simulation of the router using different ports in hexadecimal. Figure 7 shows the waveform simulation of the router using different ports in the American standard code for information interchange (ASCII).

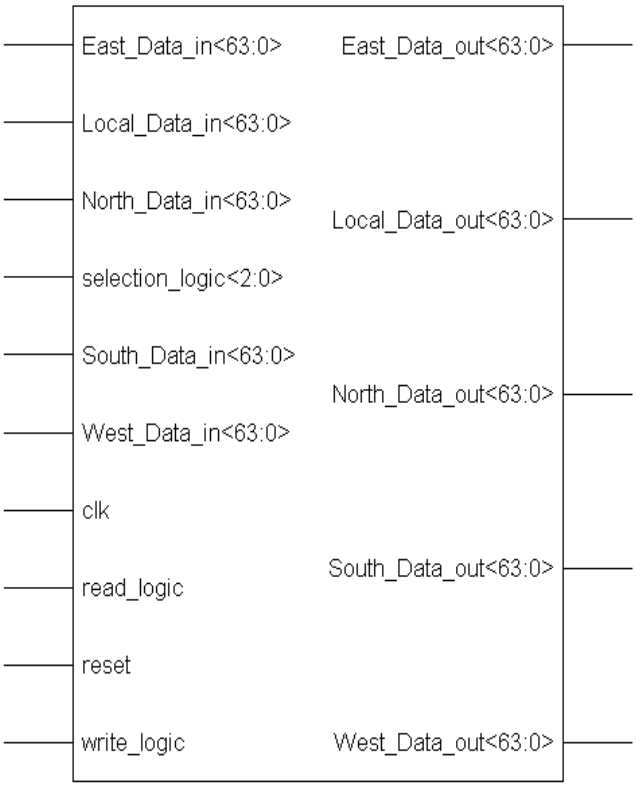


Figure 3. RTL of the chip design

Table 1. Pin functions of the chip design

Pins	Function
East_data_in <63:0>	It uses the input port to offer 64-bit data packets input (east direction)
West_data_in <63:0>	It uses the input port to offer 64-bit data packets input (west direction)
North_data_in <63:0>	It uses the input port to offer 64-bit data packets input (north direction)
South_data_in <63:0>	It uses the input port to offer 64-bit data packets input (south direction)
Local_data_in <63:0>	It uses the input port to offer 64-bit data packets input (local direction)
Clk (1-bit)	It uses the 50% duty cycle clock input signal to provide the rising edge of the clock
reset (1-bit)	It provides the input signal to reset the data input/output port values and corresponding latches
East_data_out <63:0>	It uses the output port to offer 64-bit data packets output (east direction)
West_data_out <63:0>	It uses the output port to offer 64-bit data packets output (west direction)
North_data_out <63:0>	It uses the output port to offer 64-bit data packets output (north direction)
South_data_out <63:0>	It uses the output port to offer 64-bit data packets output (south direction)
Local_data_out <63:0>	It uses the output port to offer 64-bit data packets output (local direction)
Read_logic (1-bit)	It is the control input used to write the contents from different ports into registers
Write_logic (1-bit)	It is the control input used to read the contents of different registers into ports

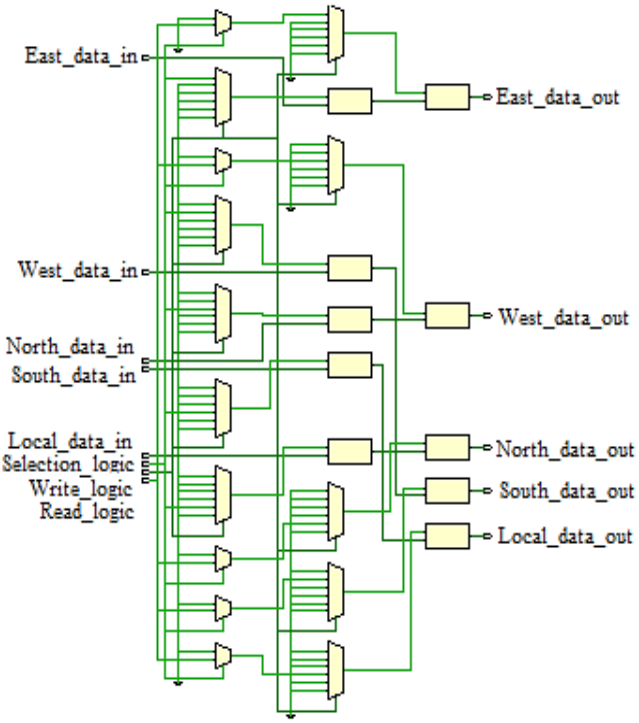


Figure 4. The internal schematic in Xilinx of the 2D router design

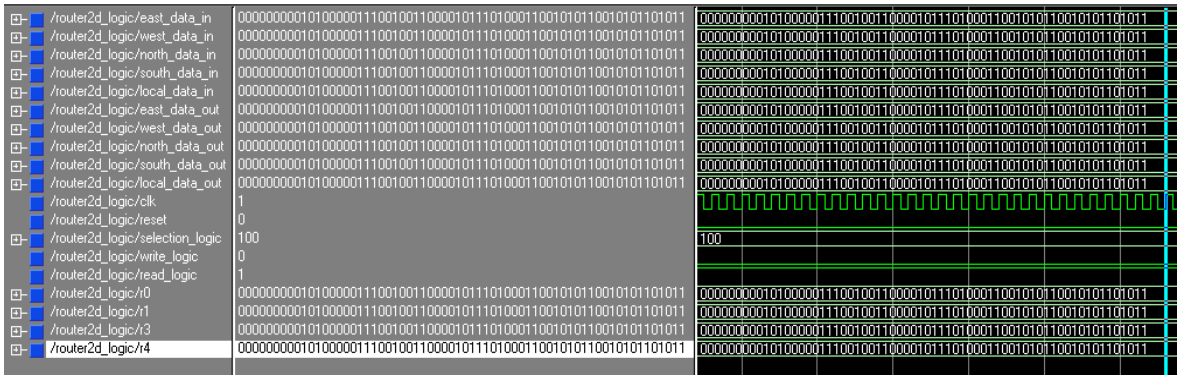


Figure 5. Waveform simulation of the router using different ports (in binary)

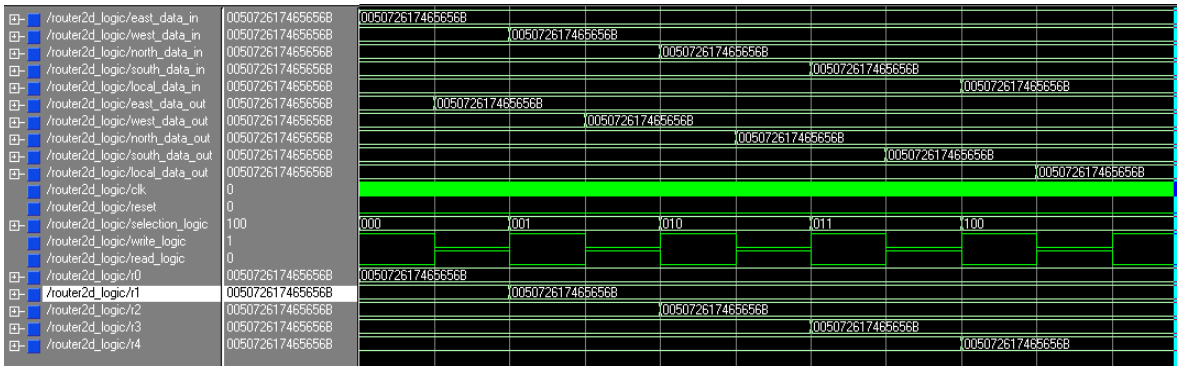


Figure 6. Waveform simulation of the router using different ports (in hexadecimal)

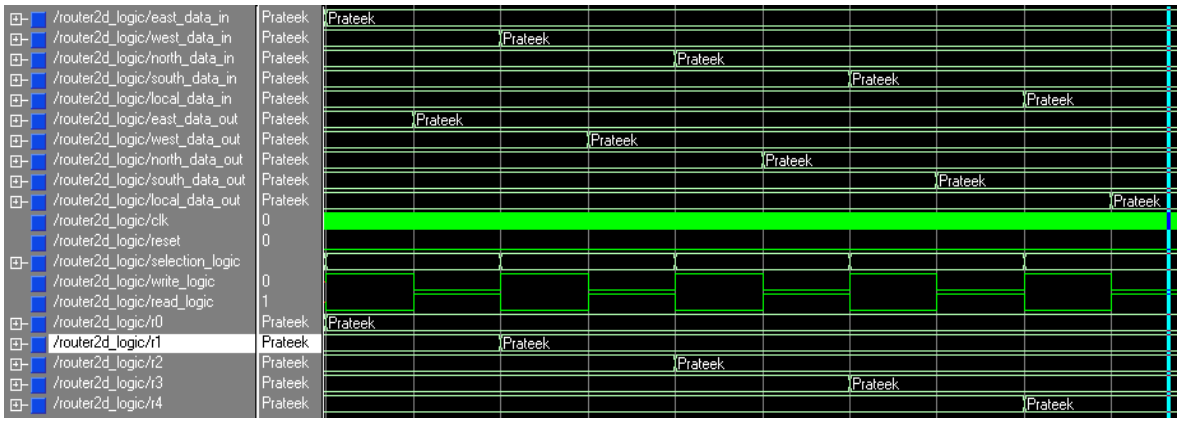


Figure 7. Waveform simulation of the router using different ports (in ASCII)

The simulation input and outputs in the waveforms are East_data_in <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101 (binary)=1'h5072617465656B (hex), West_data_in <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101 (binary)=1'h5072617465656B (hex), North_data_in <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101 (binary)=1'h5072617465656B (hex), South_data_in <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101(binary)=1'h5072617465656B (hex), Local_data_in <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101(binary)=1'h5072617465656B (hex), write_logic='1', read_logic='0', the selection_logic changes from "000", "001", "010", "011", and "100" for different ports respectively. The output of the ports is obtained when write_logic='0', and read_logic='1'. The data on the ports is verified as East_data_out <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101 (binary)=1'h5072617465656B (hex), West_data_out <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101 (binary)=1'h5072617465656B (hex), North_data_out <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101 (binary)=1'h5072617465656B (hex), South_data_out <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101(binary)=1'h5072617465656B (hex), Local_data_out <63:0>=01010000 01110010 01100001 01110100 01100101 01100101 01101101(binary)=1'h5072617465656B (hex).

The hardware parameters summary on different FPGAs is listed in Table 2. The timing values of the design on different FPGAs are listed in Table 3. Figures 8 and 9 show the variations against the obtained results of the FPGA resources for hardware and timing utilization.

Table 2. Different FPGA resources as hardware parameters summary

Hardware parameter	Spartan-3E (XC3S500E)	Spartan-6 (XC6SLX45)	Virtex-4 (XC4VFX12)	Virtex-5 (XC5VSX50T)	Virtex-7 (XC7VX550T)
Slices	368	184	368	184	184
Slice flip flops	640	320	640	320	320
LUTs	326	10	326	326	10
IOBs	647	647	647	647	647
GCLKs	5	0	5	5	0

Table 3. Different FPGA summary in terms of timing parameters

Timing parameter	Spartan-3E (XC3S500E)	Spartan-6 (XC6SLX45)	Virtex-4 (XC4VFX12)	Virtex-5 (XC5VSX50T)	Virtex-7 (XC7VX550T)
Minimum period (ns)	1.894	0.951	0.971	1.187	0.170
Minimum input arrival time before clock (ns)	4.502	1.838	4.471	1.969	0.339
Maximum output required time after clock (ns)	4.137	3.648	3.892	3.044	0.669
Maximum combinational path delay (ns)	6.031	4.600	4.863	4.231	1.178
Maximum frequency (MHz)	50.00	275.00	400.00	290.00	515.00

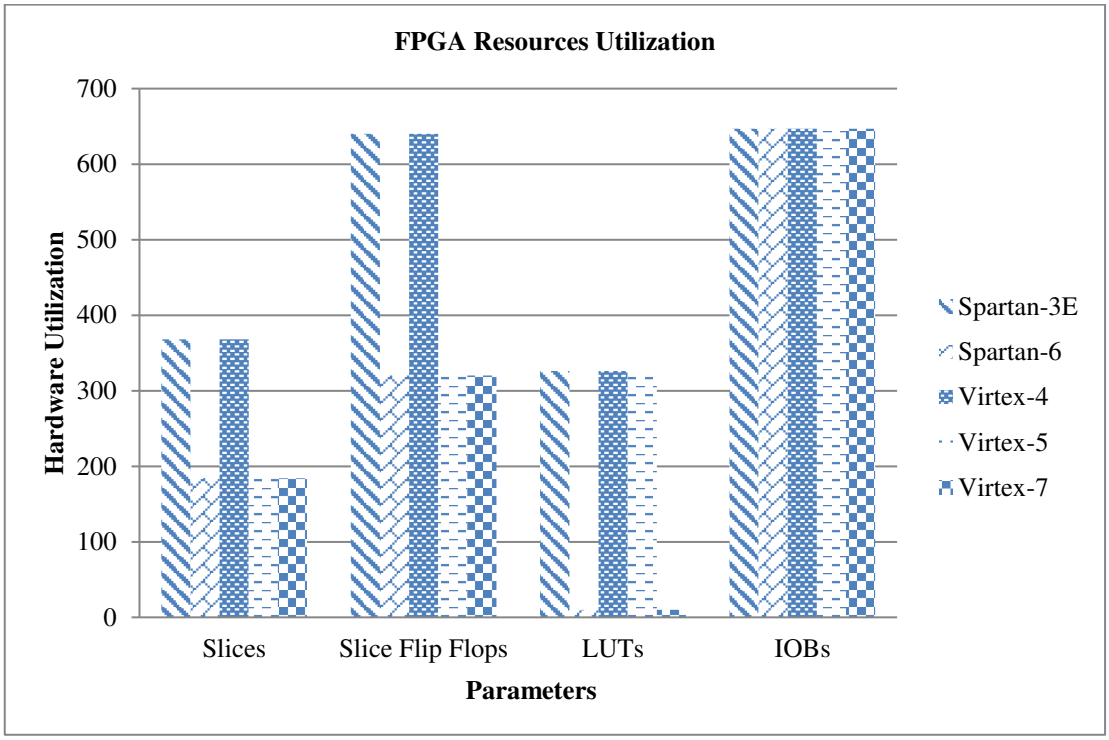


Figure 8. FPGA resources utilization summary

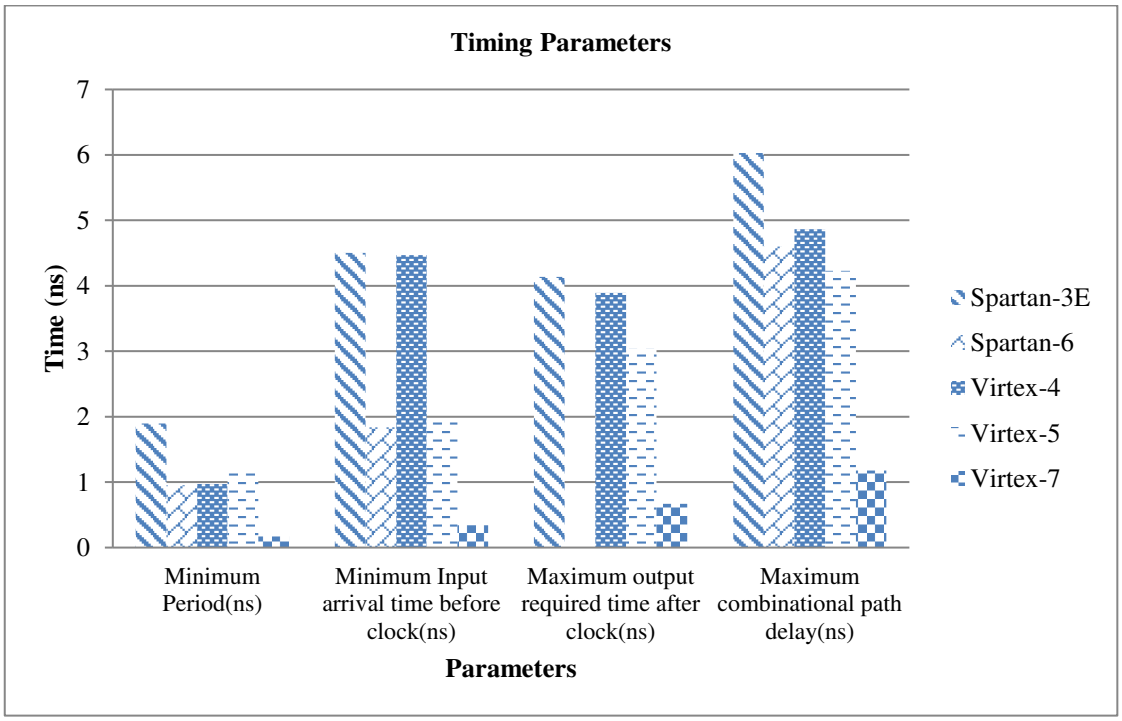


Figure 9. FPGA timing parameters utilization summary

5. CONCLUSION

The router transmits the data over the network. Routers use data packets, which can contain a variety of messages, files, data types, and simple transfers like online engagement, to route and guide network data. The hardware chip design and simulation of the 2D router are done successfully in Xilinx ISE 14.7 and Modelsim software. The chip design is verified using the different test inputs in which the 64-bit and

communication are verified using test cases on different input/output ports of the router. The behavior is verified on the different FPGA hardware such as Spartan-3E (XC3S500E), Spartan-6 (XC6SLX45), Virtex-4 (XC4VFX12), Virtex-5 (XC5VSX50T), and Virtex-7 (XC7VX550T). The comparative study presents that slices are 368, 184, 368, 184, and 184 respectively for different FPGAs. The slice flip-flops are 640, 320, 640, 320, and 320 respectively for different FPGAs. The number of LUTs is 360, 10, 326, 326, and 10 respectively for different FPGAs. It has been identified that the frequency support for different FPGAs is Spartan-3E (50.00 MHz), Spartan-6 (275.00 MHz), Virtex-4 (400.00 MHz), Virtex-5 (290.00 MHz), and Virtex-7 (515.00 MHz). The mining values indicate that the Virtex-7 FPGA has the optimal timing values in comparison to other FPGAs and Virtex-7 shows maximum frequency support for the router chip design.

ACKNOWLEDGEMENTS

Thankful to the VLSI Design Lab, University of Petroleum and Energy Studies, Dehradun, India for providing the simulation facility.

REFERENCES

[1] D. Bertozzi and L. Benini, "Xpipes: a network-on-chip architecture for gigascale systems-on-chip," *IEEE Circuits and Systems Magazine*, vol. 4, no. 2, pp. 18–31, 2004, doi: 10.1109/MCAS.2004.1330747.
[2] V. F. Pavlidis and E. G. Friedman, "3-D topologies for networks-on-chip," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 15, no. 10, pp. 1081–1090, 2007, doi: 10.1109/TVLSI.2007.893649.
[3] J. D. Owens, W. J. Dally, R. Ho, D. N. Jayashima, S. W. Keckler, and L. S. Peh, "Research challenges for on-chip interconnection networks," *IEEE Micro*, vol. 27, no. 5, pp. 96–108, 2007, doi: 10.1109/MM.2007.4378787.
[4] G. Ascia, V. Catania, M. Palesi, and D. Patti, "Implementation and analysis of a new selection strategy for adaptive routing in networks-on-chip," *IEEE Transactions on Computers*, vol. 57, no. 6, pp. 809–820, 2008, doi: 10.1109/TC.2008.38.
[5] K. Pang, V. Fresse, S. Yao, and O. A. D. Lima, "Task mapping and mesh topology exploration for an FPGA-based network on chip," *Microprocessors and Microsystems*, vol. 39, no. 3, pp. 189–199, 2015, doi: 10.1016/j.micpro.2015.03.006.
[6] R. Mishra, P. Kuchhal, and A. Kumar, "Effect of height of the substrate and width of the patch on the performance characteristics of microstrip antenna," *International Journal of Electrical and Computer Engineering*, vol. 5, no. 6, pp. 1441–1445, 2015, doi: 10.11591/ijece.v5i6.pp1441-1445.
[7] A. S. Rawat, A. Rana, A. Kumar, and A. Bagwari, "Application of multi layer artificial neural network in the diagnosis system: a systematic review," *IAES International Journal of Artificial Intelligence*, vol. 7, no. 3, pp. 138–142, 2018, doi: 10.11591/ijai.v7.i3.pp138-142.
[8] K. Suriyan, N. Ramaingam, S. Rajagopal, J. Sakkarai, B. Asokan, and M. Alagarsamy, "Performance analysis of peak signal-to-noise ratio and multipath source routing using different denoising method," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 1, pp. 286–292, 2022, doi: 10.11591/eei.v11i1.3332.
[9] N. Agrawal, A. Jain, and A. Agarwal, "Simulation of network on chip for 3D router architecture," *International Journal of Recent Technology and Engineering*, vol. 8, no. 1, pp. 58–62, 2019.
[10] X. Ju and L. Yang, "NoC research and practice: design and implementation of 2×4 2D-torus topology," *International Journal of Information Technology and Computer Science*, vol. 3, no. 4, pp. 50–56, 2011, doi: 10.5815/ijitcs.2011.04.08.
[11] R. Ji, J. Xu, and L. Yang, "Five-port optical router based on microring switches for photonic networks-on-chip," *IEEE Photonics Technology Letters*, vol. 25, no. 5, pp. 492–495, 2013, doi: 10.1109/LPT.2013.2243427.
[12] G. Verma, H. Agarwal, S. Singh, S. N. Khanam, P. K. Gupta, and V. Jain, "Design and implementation of router for NOC on FPGA," *International Journal of Future Generation Communication and Networking*, vol. 9, no. 12, pp. 263–272, 2016, doi: 10.14257/ijfgcn.2016.9.12.24.
[13] J. Y. V. M. Kumar, A. K. Swain, K. Mahapatra, and S. P. M. Verify, "Fortified-NoC: a robust approach for trojan-resilient network-on-chips to fortify multicore-based consumer electronics," *IEEE Transactions on Consumer Electronics*, vol. 68, no. 1, pp. 57–68, 2022, doi: 10.1109/TCE.2021.3129155.
[14] A. Kumar, L. Baruah, and A. Sabu, "Rotator on chip (RoC) design based on ring topological NoC," *Procedia Computer Science*, vol. 45, pp. 540–548, 2015, doi: 10.1016/j.procs.2015.03.099.
[15] J. Lai, J. Cai, and J. Chu, "A congestion-aware hybrid SRAM and STT-RAM buffer design for network-on-chip router," *IEICE Electronics Express*, vol. 20, no. 2, pp. 20220078–20220078, 2023, doi: 10.1587/elex.19.20220078.
[16] A. Kumar, G. Verma, M. K. Gupta, M. Salauddin, B. K. Rehman, and D. Kumar, "3D multilayer mesh NoC communication and FPGA synthesis," *Wireless Personal Communications*, vol. 106, no. 4, pp. 1855–1873, 2019, doi: 10.1007/s11277-018-5724-3.
[17] T. Patil, A. Sandi, D. M. D. Raj, S. Chandragandhi, and D. M. Teressa, "A minimal buffer router with level encoded dual rail-based design of network-on-chip architecture," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–6, 2022, doi: 10.1155/2022/6180153.
[18] D. Jiang, X. Wang, Z. Huang, Y. Yang, and E. Yao, "A network-on-chip-based annealing processing architecture for large-scale fully connected ising model," *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–13, 2023, doi: 10.1109/TCSI.2023.3263923.
[19] A. Jain *et al.*, "Smart communication using 2D and 3D mesh network-on-chip," *Intelligent Automation and Soft Computing*, vol. 34, no. 3, pp. 2007–2021, 2022, doi: 10.32604/iasc.2022.024770.
[20] C. Killian, C. Tanougast, F. Monteiro, and A. Dandache, "Smart reliable network-on-chip," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 2, pp. 242–255, 2014, doi: 10.1109/TVLSI.2013.2240324.
[21] N. Gupta, K. S. Vaisla, A. Jain, A. Kumar, and R. Kumar, "Performance analysis of AODV routing for wireless sensor network in FPGA hardware," *Computer Systems Science and Engineering*, vol. 40, no. 3, pp. 1073–1084, 2021, doi: 10.32604/CSSE.2022.019911.
[22] Y. Liu, R. Guo, C. Xu, X. Weng, and Y. Yang, "A Q-learning-based fault-tolerant and congestion-aware adaptive routing algorithm for networks-on-chip," *IEEE Embedded Systems Letters*, vol. 14, no. 4, pp. 203–206, 2022, doi:

10.1109/LES.2022.3176233.
[23] N. B. Jadhav and B. S. Chaudhari, "Efficient non-blocking optical router for 3D optical network-on-chip," *Optik*, vol. 266, 2022, doi: 10.1016/j.ijleo.2022.169563.
[24] R. Min, R. Ji, Q. Chen, L. Zhang, and L. Yang, "A universal method for constructing N-port nonblocking optical router for photonic networks-on-chip," *Journal of Lightwave Technology*, vol. 30, no. 23, pp. 3736–3741, 2012, doi: 10.1109/JLT.2012.2227945.
[25] F. Mehmood *et al.*, "An efficient and cost effective application mapping for network-on-chip using Andean condor algorithm," *Journal of Network and Computer Applications*, vol. 200, 2022, doi: 10.1016/j.jnca.2021.103319.
[26] S. Sikandar, N. K. Baloch, F. Hussain, W. Amin, Y. B. Zikria, and H. Yu, "An optimized nature-inspired metaheuristic algorithm for application mapping in 2D-NoC," *Sensors*, vol. 21, no. 15, pp. 1–21, 2021, doi: 10.3390/s21155102.
[27] A. N. Rao, B. R. Naik, and L. N. Devi, "An efficient coverage and maximization of network lifetime in wireless sensor networks through metaheuristics," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 10, no. 3, pp. 159–170, 2021, doi: 10.11591/ijict.v10i3.pp159-170.
[28] C. V. Nguyen *et al.*, "ZigBee based data collection in wireless sensor networks," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 10, no. 3, pp. 212–224, 2021, doi: 10.11591/ijict.v10i3.pp212-224.

BIOGRAPHIES OF AUTHORS

Mr. Prateek Agarwal is B.Tech. in Electronics and Communication Engineering from Uttar Pradesh institute of Technology, (UPTU), Lucknow India in 2006. M.E. in Electronics and Communication, Thapar University (TU), Patiala India in 2008. Pursuing Ph.D (Electronics and Communication Engineering) in the domain of VLSI design. His areas of interest are VLSI design and digital communication. He has published 5 research papers in international and national conferences. He can be contacted at email: prateek.aggrawal@gkv.ac.in.

Dr. Tanuj Kumar Garg received B.Tech. in Electronics and Communication Engineering from G.B.P.U.A.T, Pantnagar, Uttarakhand in 2000, M.E from Delhi College of Engineering, Delhi in 2001 and Ph.D. from Uttarakhand Technical University, Dehradun, Uttarakhand. He is currently working as Assistant Professor in Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, Gurukula Kangri (Deemed to be University). His research interest includes microstrip antennas, reconfigurable antennas, metamaterials, metamaterial antennas, tunable antennas, leaky wave antennas, and VLSI design. He can be contacted at email: tanujkgarg@rediffmail.com.

Dr. Adesh Kumar is working as a senior associate professor in the Department of Electrical and Electronics, Engineering ''University of Petroleum and Energy Studies, Dehradun, India. He is B.Tech. in Electronics and Communication Engineering from Uttar Pradesh Technical University Lucknow, India in 2006. M.Tech (Hons) in Embedded Systems Technology, from SRM University, Chennai in 2008. Ph.D. (Electronics Engineering) form University of Petroleum and Energy Studies (UPES), Dehradun India in 2014. He has also worked as senior engineer in TATA ELXSI LIMITED Bangalore and faculty member in ICFAI University, Dehradun. His areas of interest are VLSI design, embedded systems design, telecommunications, and signal processing, He has published more than 100+ research papers in international peer-reviewed journals (SCI/Scopus) and conferences. He is the reviewer of many SCI/SCIE and Scopus journals such as IEEE Transactions of Industrial Electronics, Wireless Personal Communications, Microsystem, Technologies, 3D research, Springer, Journal of Electronic Science and Technology, Journal of Visual Languages and Computing, Elsevier, World Journal of Engineering, and Emerald are few of them. He has supervised 7 PhD scholars and 5 candidates are doing research under his supervision. He has chaired more than 10 sessions at international conferences and is involved in the Ph.D. evaluation committee at UPES, India. Guest Editor: Special Issue-"Intelligent devices and computing applications" in Computer Systems Science & Engineering, Journal, Tech Science, 2022 (SCI Indexed). Guest Editor: Special Issue-"Intelligent communication and smart grid automation applications" in Intelligent Automation & Soft Computing, Journal, Tech Science, 2021 (SCI Indexed). Editor: Proceedings of intelligent communication, control, and devices in advances in intelligent system and computing (AISC), book series, Springer, vol. 989, 2020 (Scopus Indexed). He can be contacted at email: adeshmanav@gmail.com.

Check for updates

Received: 16 May 2023 | Revised: 13 September 2023 | Accepted: 9 October 2023

IET Renewable Power Generation

DOI: 10.1049/rpg2.12866

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

A comparative study of combined cycles for concentrated solar power for efficient power generation using low Global Warming Potential (GWP) fluids to reduce environmental effects

Yunis Khan[1] | **Hakan Caliskan**[2] | **Hiki Hong**[3]

[1]Department of Mechanical Engineering, Delhi Technological University, New Delhi, India

[2]Department of Mechanical Engineering, Faculty of Engineering and Natural Science, Usak University, Usak, Turkey

[3]Department of Mechanical Engineering, Kyung Hee University, Yongin, Republic of Korea

Correspondence
Yunis Khan, Department of Mechanical Engineering, Delhi Technological University, Bawana Road, New Delhi 110042, India.
Email: yuniskhan21@gmail.com

Hakan Caliskan, Department of Mechanical Engineering, Faculty of Engineering and Natural Science, Usak University, 64200, Usak, Turkey.
Email: hakan.caliskan@usak.edu.tr

Hiki Hong, Department of Mechanical Engineering, Kyung Hee University, Yongin 17104, Republic of Korea.
Email: hhong@khu.ac.kr

Abstract

A comparative analysis of a combined system comprising organic Rankine cycles (ORC) and supercritical CO_2 (sCO_2) cycles for concentrated solar power (CSP) applications was performed in the article. Low global warming potential (GWP) fluids were used to investigate strategies for reducing the effects on global warming. In the current study, four different combinations of the sCO_2 cycles and ORC, including basic recuperated, precompression, recompression with main compressor intercooling, and partial cooling, were taken into consideration, and their performances with regard to CSP application were compared. The integrated recompression with main compressor intercooling and ORC arrangement was seen to operate thermodynamically more efficient than the other considered configurations. This configuration's highest thermal and exergy efficiencies were discovered to be 51.9% and 55.84%, respectively, and R1243zf was found to be the best-performing fluid among other considered low GWP fluids to reduce the environmental effects.

1 | INTRODUCTION

Energy consumed for cooling and producing electricity has dramatically increased nowadays [1]. Carbon emissions rise as a result of the continual increase in the use of fossil fuels. However, fossil fuel reserves are steadily diminishing. As a result, it becomes challenging to find reliable and sustainable energy sources [2]. Biomass, wind, solar, and geothermal energy are just a few of the renewable energy sources currently being employed to produce, clean and environmentally friendly energy. Solar energy is the most suited renewable energy source for power generation since it is readily available, inexpensive, and operates quietly [3].

Sunlight is captured by solar collectors and used for heating, cooling, and electricity generation [4]. Concentrated solar collectors, such as parabolic trough collectors (PTC), heliostats, dish collectors etc. are needed for high-temperature applications. The operating temperature ranges for parabolic trough collectors (60–500°C), dish collectors (750–1000°C), and heliostats (150–1500°C) were observed by Kalogirou [5]. CSP technology is beneficial for power generation in tropical climatic conditions using solar heat that is abundantly available [6].

Of the various CSP technologies, the solar power tower (SPT) system is the latest technology. The SPT system is made up of several intricate sub-systems, including a receiver, a tower that is 75–150 meters tall, a thermal storage system that is optional, a field of heliostats with each heliostat having a 50–150 square meters surface area, and a power conversion system. A SPT is recommended for producing high-temperature heat for the

Abbreviations: CIT, Compressor inlet temperature (°C); Cond, Condenser; CR, Concentration ratio; HEX1, Heat exchanger (1); HEX2, Heat exchanger (2); HTR, High temperature recuperator; LTR, Low temperature recuperator; MC, Main compressor; MT, Main turbine; ORC, Organic Rankine cycle; OT, ORC turbine; PC, Pre-compressor; RCMCIC, Recompression with main compressor intercooling; SPT, Solar power tower; TIT, Turbine inlet temperature (°C).

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *IET Renewable Power Generation* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

generation of power through a high-temperature power cycle [7]. Previous studies based on the SPT-driven cycles such as the recompression cycle combined with the trans-critical CO_2 (tCO_2) cycle [8], cascade sCO_2 Brayton cycle and ORC [9], the recompression sCO_2 with and without main compressor intercooling [10], the sCO_2 cycle and parallel double evaporator ORC [11], and the co-generation combine cycle [12] have been carried out.

The carbon dioxide (CO_2) over its critical thermodynamic properties (i.e. 7.38 MPa, 30.9°C) is referred to as the supercritical carbon dioxide, that is, sCO_2. Additionally, the sCO_2 cycle refers to the gas cycle in which the supercritical stage of CO_2 is used as the working fluid that can make use of the heat from a variety of heat sources, including geothermal, coal, natural gas, and solar thermal energy [13]. A few scholars studied the solar PTC-operated combined cycle based on the sCO_2 cycle such as Khan and Mishra [14] published the combined partial heating sCO_2 cycle and ORC. They found that combining ORC with the partial heating cycle led to an improvement in thermal efficiency by 4.47%. The energy efficiency of the integrated system increased as solar radiation increased. They also found that based on a thermodynamic performance point view, R1233zd(E) overpowers all other working fluids, including R1243zf, R1234ze(E), R1234yf, R1234ze(Z) and R1224yd(Z). Furthermore, Khan and Mishra [15] studied the pre-compression configuration of the sCO_2 cycle combined with ORC integrated with the SPT. It was found that the standalone pre-compression sCO_2 cycle's energy efficiency and output power rose by 4.51% and 4.52%, respectively, when ORC was used as the bottoming cycle. They also discovered that, among all the working fluids, R227ea recovered the highest heat. Qin et al. [16] developed a novel combined cycle that consists of a recompression configuration of the Brayton cycle and tCO_2 refrigeration system to recover waste heat from the marine engine to develop combined power and cooling. They found the coefficient of performance (COP) and waste heat recovery efficiency to be 3.059 and 65.1%, respectively. An exergoeconomic analysis of the pre-compression configuration of the sCO_2 cycle combined with ORC for SPT application was performed by Khan and Mishra [17]. They discovered that at 950 W/m^2 of solar irradiation, R1336mzz(Z) provided the maximum exergy efficiency, energy efficiency, and power production. In another work they [18] examined the thermal performance of an SPT-driven recompression configuration of a sCO_2 cycle in which main compression intercooling was used and ORC was utilized as a waste heat recovery cycle using low GWP fluids. They observed that adding the ORC boosted the thermal efficiency by 7–8% as compared to the basic considered cycle. Kim et al. [19] conducted a comparative analysis of different sCO_2 cycle configurations, including a basic recuperated cycle, recompression, pre-compression, single heated cascade cycle, dual expansion cycle, partial heating, partial recuperated cycle, dual heated cascade cycle, inter-cooled single heated cascade cycle, dual heated and flow split, intercooled dual heated cascade cycle, triple heating cycle, as land filling gas turbine bottoming cycles. They found that compared to the other cycles under consideration, the pre-compression and re-compression cycles were more thermally efficient. It was also suggested to conduct more research on the sCO_2 as waste heat working fluid in combined cycles. Recently, Zhu et al. [20] proposed two novel combined cycles consisting sCO_2 Brayton cycle and the ORC/Kalina cycle to utilize waste heat. In comparison results they found that using R32 and ammonia, the sCO_2 Brayton cycle/ORC performed thermodynamically best as compared to the sCO_2 Brayton cycle/Kalina cycle.

In addition, ORC is a useful technology for recovering waste heat at low temperatures. At the moment, ORC is predominantly used as a bottoming cycle. Several studies back up this claim, such as Nondy and Gogoi [21] examined the performance of the different configurations of the ORC as the waste heat recovery cycle consisting of basic ORC, regenerative ORC, recuperated ORC, and regenerative-recuperative ORC. They observed that recuperative-regenerative ORC was thermodynamically superior to the other three configurations. This configuration's exergy efficiency and power output were observed as 16.19% and 15.33%, respectively, higher than that of the basic ORC. Wang et al. [22] examined a combined system of gas turbine and SOFC (solid oxide fuel cell) and ORC as a low-temperature heat recovery cycle. They performed exergy, energy, and economic analyses of the proposed system. They observed that overall system exergy efficiency was found as 92.95%. The ORC was found as beneficial for waste heat recovery. Mahmoud et al. [23] considered two configurations of the ORCs first one basic ORC and second regenerative ORC to use wasted heat from the diesel engine. After detailed thermodynamic analysis, they observed that basic ORC and the regenerative ORC enhanced the thermal performance by 7.98% and 15.31%, respectively, as compared to standalone diesel engines. Hoang [24] examined the use of the ORC as a waste heat recovery cycle to utilize the diesel engine exhaust heat. They observed that a combined system with diesel engines can achieve thermal efficiencies of up to 90%, whereas alone ORC can only reach up to 25%. Song et al. [25] conducted an optimization and performance evaluation of the combined sCO_2 cycle-ORC. The effects of the topping cycle recuperative ratio and pressure ratio on the effectiveness of the bottoming ORC were researched. Parallel double evaporator ORC (PDORC) is superior to basic ORC in terms of its ability to recover waste heat. PDORC generated more work when compared to the conventional ORC due to its two turbines [26].

The literature review revealed that numerous investigations of the sCO_2 cycles on the CSP application had been carried out. However, in previous research, very few comparative studies were performed on a combination of cycles for the application of the CSP system. It can be said no comparative study was performed for the application of solar power tower systems with new combined cycles. Therefore, four new sCO_2 cycles combined with ORC for waste heat recovery were taken into consideration for harvesting solar heat from the SPT plant in the current study. Simple recuperation, pre-compression, recompression with main compressor intercooling, and partial cooling are the four sCO_2 configurations combined with ORCs that were considered. For the CSP application, the cycle's performance was compared. Simultaneously, performances with low

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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 SPT operational and geometric parameters

Parameters	Values
Temperature approach of solar receiver (δT_R)	423.15 K [33]
Direct normal irradiation (DNI)	850 W/m^2 [8]
Thermal emittance (ζ)	0.85 [33]
View factor (f_{View})	0.8 [33]
Tower height	74.62 m [34]
Initial temperature difference	15 K [8]
Concentration ratio (CR)	900 [33]
Coefficient of convective heat loss (h_{conv})	10 W/m^2-K [33]
Number of heliostats (N_h)	141 [33]
Total mirror area of heliostat (A_h)	9.04 ×7.89 [8]
Convective heat loss factor (f_{conv})	1 [33]
Efficiency of heliostat (η_h)	58.71 % [33]
Solar multiple	2.8 [8]
Absorptance (α)	0.95 [33]
Sun temperature (T_{su})	5800 K [28]

TABLE 2 Input Information for the ORC and sCO_2 cycles

Parameters	Values
Pressure at the main turbine intake	25 MPa [28]
Temperature at the main turbine's intake	650°C [28]
Pre-compressor's inlet pressure	5.6–6.8 MPa [28]
Main compressor's inlet pressure	7–10.5 MPa [28, 33]
Temperature at the main compressor's intake	32–38°C [35, 36]
Efficiency of ORC turbines	0.8 [25]
Efficiency of the main compressor	0.85 [35, 36]
Effectiveness of the heat exchanger	0.95 [32]
Turbine isentropic efficiency	0.88 [35, 36]
Effectiveness of HTR/LTR/Heat exchangers	0.95 [37]
sCO_2 mass flow rate	1.5 kg/s
ORC fluid mass flow rate	2.5 kg/s
Incoming pressure of ORC turbine	3 MPa
Efficiency of pre-compressor	0.85 [35, 36]
Efficiency of the ORC pump	0.7 [25]

GWP fluids in ORC for all considered configurations were also compared to reduce global warming effects.

2 | SYSTEMS DESCRIPTION

In this section, all considered combined cycles were described one after the other.

2.1 | Combined basic recuperated sCO_2 cycle and ORC

SPT is the most recent technology of operating high-temperature cycles, which has a high level of maturity for commercial use. SPT can operate effectively within a temperature range of 300–2000°C. The operating parameters of SPT used in this study are given in Table 1. The current system consists of three subsystems: a basic configuration of the Brayton cycle using sCO_2, an SPT plant, and an ORC that serves as a waste heat recovery cycle. The SPT system powers the sCO_2 cycle by supplying heat via the heat exchanger-1 (HEX1) to the sCO_2 Brayton cycle, which generates electricity via a turbine. The heat that is left over after the recuperator is recovered by the ORC, where some power is also generated by the organic turbine (OT), via heat exchanger-2 (HEX2). The whole configuration is shown in Figure 1.

2.2 | Combined pre-compression sCO_2 cycle and ORC

An additional compressor known as pre-compressor is used in the pre-compression configuration between the turbine and main compressor that makes the turbine outlet pressure independent of the compressor inlet pressure. Additional compression is used for further regeneration once the temperature difference in the recuperator has been minimized [27]. The benefit of this configuration is the inclusion of an extra compressor, which gives space for more regeneration [28].The temperature of the fluid rises with more compression, and due to increased compression, the fluid's specific heat also rises. As a result, the regeneration process can occur for a long time and generates a lot more heat than the standalone Brayton cycle does. The two thermodynamic cycles that make up the current combined cycle model are the pre-compression configuration that employs the sCO_2 as a working fluid and ORC as a waste heat recovery cycle. A representation of the present configuration driven by the SPT is given in Figure 2. Heat transfer fluid (HTF) (molten salt) circulates in the SPT field. The heat exchanger-1 (HX1) (process a-b) is used to transfer heat from the HTF to the topping cycle. Table 2 lists the combined cycle's input parameters.

2.3 | Combined recompression main compressor intercooling (RCMCIC) sCO_2 cycle and PDORC

It was found that the standalone recompression cycle's thermal efficiency increased by 2.68% by integrating intercooling between its main compressors at the same base conditions [29]. The intercooler (IC) is positioned between the two main compressors, as opposed to the main compressor intercooling, which places the intercooler between the pre-compressor and main compressor. By incorporating the IC into the standalone recompression cycle, efficiency is highly improved [30]. For the analysis in this study, the RCMCIC configuration is also taken into account. A recent innovation is the PDORC.

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

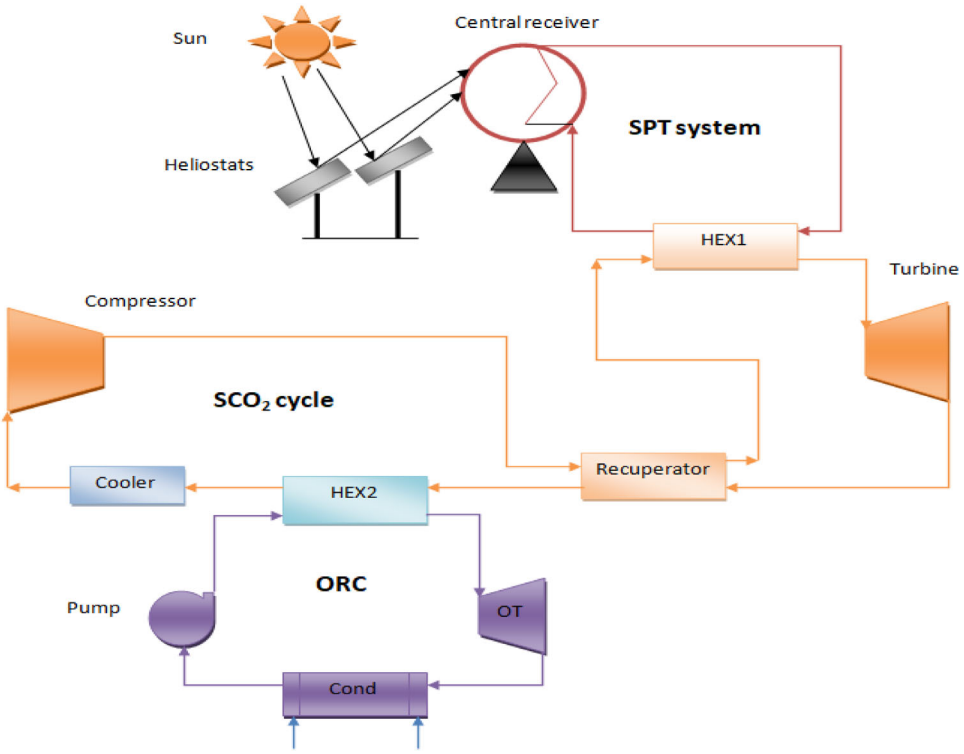


FIGURE 1 Configuration of combined basic sCO_2 cycle-ORC driven by SPT

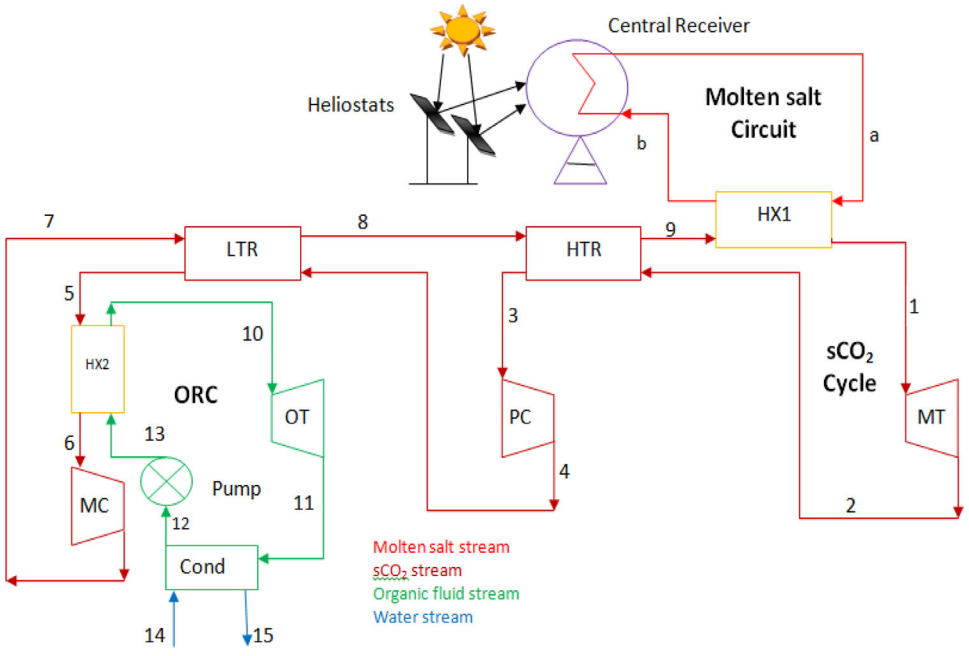


FIGURE 2 Schematic diagram of combined pre-compression sCO_2 cycle-ORC driven by SPT [17]

Due to the simultaneous operation of the two recuperators, this PDORC is better suited for waste recovery than standard ORC. Compared to the standard ORC, PDORC can generate more work output [26]. In this system, the standalone RCMCIC cycle's thermal performance was improved by the PDORC's increased heat recovery. PDORC was therefore regarded as a waste heat recovery cycle in the present setup. As can be seen in Figure 3, the present model includes three subsystems: the SPT system, the RCMCIC sCO_2 cycle, and the PDORC as the bottoming cycle. PDORC subsystem works to recover

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

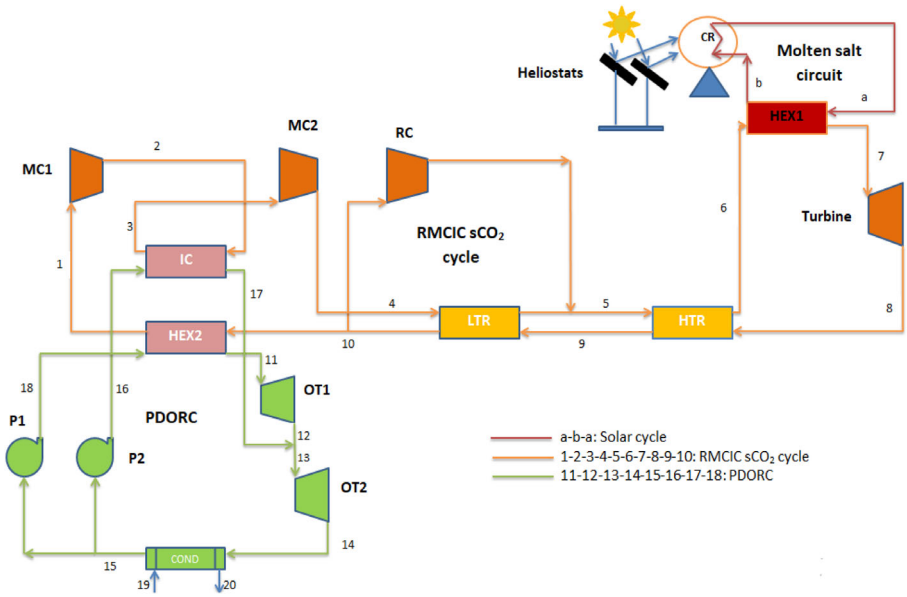


FIGURE 3 Combined RCMCIC sCO_2 cycle-PDORC driven by SPT

TABLE 3 Thermo-environmental properties of considered fluids [17, 40, 41]

Working substance	Weight (Kg/Kmole)	T_b(°C)	T_c (°C)	P_c (MPa)	Type	GWP	Lifetime (years)	ODP	Security Group
R1234ze(E)	114.043	−19.0	109.4	3.64	D	6	0.025	0	A2L
R1224yd(Z)	148.5	14	155.5	3.33	I	0.88	–	0.00023	A1
R1233zd(E)	130.5	18.32	165.5	3.57	I	1	–	0.00024	A1
R1234yf	114.04	−30	94.7	4.597	I	<1	–	0	A2L
R1336mzz(Z)	164	33.4	171.3	2.903	D	<2	0.0602	0	A1
R1234ze(Z)	114.04	9.8	150.1	3.53	I	<10	–	0	–
R1243zf	96.05	−25.41	104.4	3.518	D	<1	–	0	A2
R1225ye(Z)	130.5	−20	106.5	3.335	I	0.87	–	0.00012	–

I: Isentropic, D: dry.

waste heat through the HEX2 and intercooler at the same time. For the performance examination of PDORC the eight fluids were considered. These consist of R1234yf, R1234ze(Z), R1224yd(Z), R1225ye(Z), R1233zd(E), R1243zf, R1234ze(E) and R1336mzz(Z) as the working fluids. The thermophysical characteristics of these fluids are listed in Table 3.

2.4 | Combined partial cooling sCO_2 cycle and ORC

Three subsystems make up the current combined system, the first one is the SPT system as the heat source, the second one is the partial cooling sCO_2 ($PCSCO_2$) cycle as the topping power cycle and the last one is the bottoming waste heat recovery ORC. Similar to the recompression cycle, the $PCSCO_2$ cycle also uses an additional compressor and cooler, as seen in Figure 4. Intercooling between distinct compression stages is used to lessen the amount of work required during the compression process [31]. At high temperatures, it also outperforms the recompression design in terms of efficiency. The cycle is appropriate for warming since its pressure ratio is often higher than the recompression cycle [32]. These benefits of the partial cooling cycle have been taken into account while conducting the present study's additional investigation of this cycle. The ORC has been incorporated into this cycle as a waste heat recovery cycle. The eight low GWP fluids are considered for the analysis in ORC as listed in Table 3.

3 | THERMODYNAMIC ANALYSIS

3.1 | Assumptions

The following presumptions have been taken into consideration when conducting the thermodynamic research of the

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

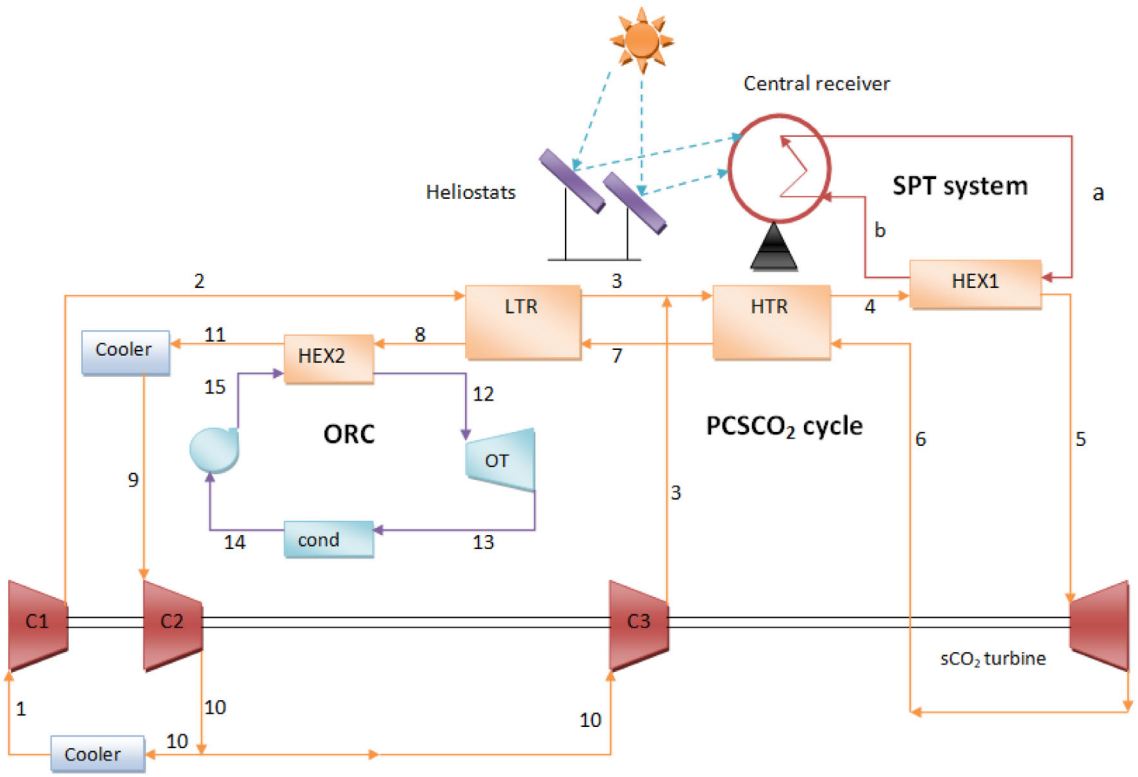


FIGURE 4 Combined partial cooling sCO_2 cycle-ORC driven by SPT

SPT-driven all four combined cycles in accordance with the simulation; (1) all components of the system are in steady condition has been assumed. (2) Pressure loss, and friction loss in connecting pipes have been neglected. (3) Each component's potential and kinetic energy are insignificant. (4) The heliostat and receiver parameters are assumed in the input data, which are displayed in Table 1 to aid simulation. (5) The HEX1's input temperature for molten salt has been measured as 700°C [36]. (6) The maximum topping cycle temperature has been calculated to be 50°C below the salt temperature at the HEX-inlet. (7) Another assumption made was that the heliostat only receives beam irradiation (DNI).

3.2 | Mathematical modeling

Modeling equations for SPT have been taken from previous studies [8]. The control volume approach of each component has also been considered.

Solar heat incidence upon the heliostat is expressed by

$$\dot{Q}_{solar} = DNI \cdot A_h \cdot N_h, \tag{1}$$

where DNI is direct normal irradiation received on heliostat, A_h is the single area of heliostat (m^2), and N_h shall be the number of the heliostats. However, part of this heat is lost in the atmosphere due to the efficiency of the heliostat. Therefore, the total heat received by the heliostats is defined as;

$$\dot{Q}_h = \dot{Q}_{solar} \cdot \eta_h, \tag{2}$$

where η_h is heliostat's efficiency. This quantity of heat travels to the receiver where the HTF is flowing to transfer the solar heat to the cycles. Here too, some part of the heat is lost from the receiver to the environment. Therefore, the net heat present at the solar central receiver is calculated as;

$$\dot{Q}_r = \dot{Q}_h \cdot \eta_r \dot{Q}_h - \dot{Q}_{loss,r}, \tag{3}$$

where η_r is the solar receiver efficiency;

$$\eta_r = \alpha - \frac{\zeta \cdot f_{view} \cdot \sigma \cdot T_R^4 + h_{conv} \cdot f_{conv} \cdot (T_R - T_{air})}{DNI \cdot \eta_h \cdot CR}, \tag{4}$$

where T_R and CR are the solar receiver surface temperature and concentrated ratio, respectively. The T_R is approximated as

$$T_R = T_1 + \delta T_R, \tag{5}$$

where the temperature of the turbine inlet is T_1, and the temperature of the solar receiver approach is δT_R. Table 1 lists the operational and geometric requirements for the heliostat field, solar receiver, and heliostats.

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

Moreover, the exergy can be defined as the maximum work that the system can achieve when the system reaches its equilibrium conditions. The equation of exergy balance can be established as control volume. Exergy destruction for each component can be calculated by the exergy balance equation. Therefore using the control volume approach the exergy balance equation can be written as [37]

$$\sum\left(1-\frac{T_0}{T_J}\right)\dot{Q}_j-\dot{W}_{cv}+\sum(\dot{m}_iE_i)-\sum(\dot{m}_eE_e)-\dot{ED}=0, \quad (6)$$

where $\dot{ED}$ refers to the destruction of exergy, and j relates in a particular state to thermal property. The suffix 'i' and 'e' refers to the inlet and exit to the control volume. Petela's formula is used to calculate the solar exergy at the inlet of combined cycles. This empirical relation is shown in Equation (7) [38];

$$\dot{E}_{solar}=\dot{Q}_{solar}\cdot\left[1+\frac{1}{3}\left(\frac{T_0}{T_{su}}\right)^4-\frac{4}{3}\left(\frac{T_0}{T_{su}}\right)\right], \quad (7)$$

where T_{su} is the temperature of the sun and T_0 is the reference temperature. Exergy that is transferred to the HTF (molten salt) in the solar circuit through the solar receiver can be calculated as;

$$\dot{E}_r=\dot{m}_{ms}\cdot Cp_{ms}\cdot\left[(T_b-T_a)-\left(T_0\cdot\ln\frac{T_b}{T_a}\right)\right]. \quad (8)$$

There is no concentration change in the working fluids therefore chemical exergy is constant throughout. Physical exergy at k^{th} thermodynamic state is defined as [35];

$$E_k=(h_k-h_0)-T_0(h_k-s_0). \quad (9)$$

The net power obtained from the combined cycle is defined as the difference between total power produced from the cycle and total power consumed from that cycle. Therefore, mathematical it can be expressed as

$$\dot{W}_{net}=\dot{W}_T+\dot{W}_{OT}-\dot{W}_{Comp}-\dot{W}_{pump}. \quad (10)$$

Thermal efficiency of the solar-driven combined cycle is defined as the ratio of output power to the input heat, that is, solar heat is expressed as;

$$\eta_{th}=\frac{\dot{W}_{net}}{\dot{Q}_{solar}}. \quad (11)$$

Exergy assessment of the integrated system has also been covered in this section. After assuming that there is no heat loss within the component, the exergy balance Equation (6) for each component is used to compute the exergy destruction in each component [37]. In order to analyze how well the SPT-powered combined cycle performs, a variety of mathematical relationships are employed based on the thermal modeling, as discussed below;

TABLE 4 Molten salt ($MgCl_2$+ KCl) thermophysical properties [8]

Parameters	Values
Solidification temperature	699 K
Stability limit	1691 K
Specific heat capacity	1.028 (kJ/kg-K)
Thermal conductivity	0.39 (W/m-K)
Density	1593 (kg/m^3)

Combined cycle exergy efficiency is defined as the ratio of useful exergy to the exergy consumed, that is, exergy of solar is expressed as [37];

$$\eta_{ex}=\frac{\dot{W}_{net}}{\dot{E}_{solar}}. \quad (12)$$

The combined cycle's thermal efficiency can also be defined by the relation between the thermal and exergy efficiency of the combined cycle [35];

$$\eta_{th}=\eta_{ex}\cdot\eta_{Carnot}. \quad (13)$$

The computational technique Engineering Equation Solver (EES) was employed to solve modeling equations [39].

3.3 | Working fluid selection

Since the working fluids are responsible for thermodynamic, economic, and environmental aspects. Therefore, choosing the working fluids carefully is necessary. As molten salt HTF in the receiver, a solution of potassium chloride (KCl) and magnesium dichloride ($MgCl_2$) with mass fractions of 68% and 32%, respectively, has been utilized [8]. The selection of this HTF was made because it is less expensive than solar salt and liquid sodium (Na) for the heliostat-driven sCO_2 cycle [8]. The thermophysical characteristics of HTF are listed in Table 4.

It is difficult to choose a suitable working fluid for ORC that gives the best thermo-physical properties at the right temperature and pressure because organic fluids lose their chemical stability at a certain temperature limit [40]. Wet, isentropic, and dry fluids are the three categories of working fluids for the ORC. To keep the high-quality vapor (which means superheated vapor) at the exit of the expander isentropic and dry type of working fluids are better options than wet type of fluids [41]. In order to choose the optimal fluids for the experiment, several elements, including essential ones like GWP, ozone depletion potential (ODP), and thermal stability were considered. The present investigation eliminated substances like CFCs (chlorofluorocarbons) and hydrofluorocarbons (HCFCs) that have high GWPs and ODPs. There should be zero ODP for the selected working fluids. However, the GWP is limited to fewer than 150, as restrictions from bodies like the

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

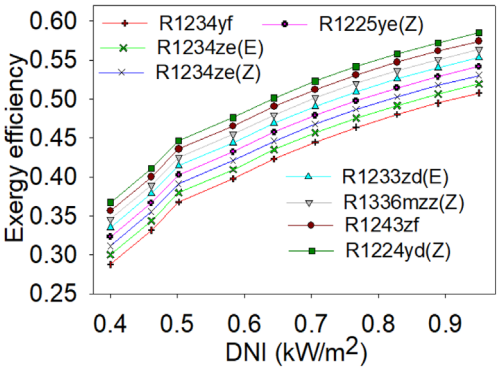


FIGURE 5 Working fluids effects on exergy efficiency of basic cycle-ORC

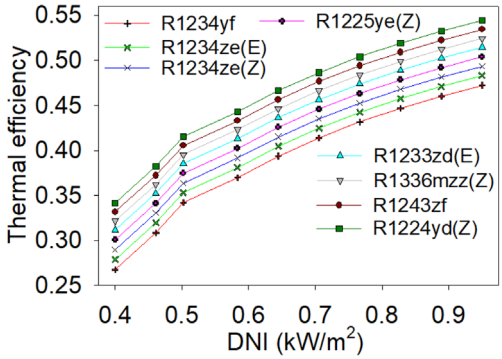


FIGURE 6 Working fluids effect on the thermal efficiency of basic cycle-ORC

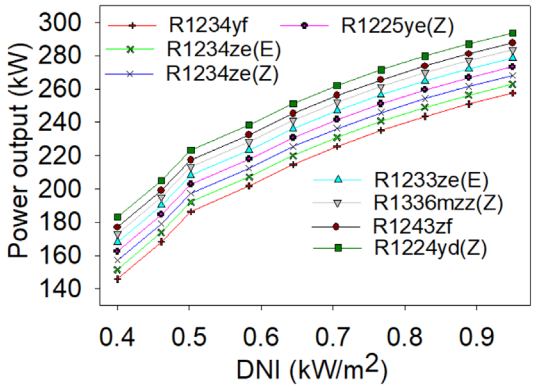


FIGURE 7 Working fluids effects on the power output of basic cycle-ORC

European Union [42]. Since the expander outlet requires high-quality vapor dry and isentropic fluids are more suited than other fluids. Therefore, this analysis was performed on the ORC analytics based on eight HFO (hydro fluoro olefins) working fluids: R1234ze(Z), R1225ye(Z), R1234yf, R1243zf, R1224yd(Z), R1336mzz, R1234ze(E) and R1233zd(E). There are two classes of toxicity: Class A, which is for lower toxicity, and Class B, which is for higher toxicity. To define the fluid's flammability intensity the four flammability classes are available: 1, 2L, 2, or 3, the higher the number, the more flammable the substance is. The character 'L' refers to lower burning velocity [43]. As a result of these factors, the working fluids indicated in Table 3 were taken into consideration in the current investigation, and their performances were compared.

4 | RESULTS AND DISCUSSION

Mathematical equations have been solved in EES software taking input parameters as listed in Tables 1 and 2. In this section, first, the working fluid selection for all configurations was performed individually then performance comparisons of all configurations on the single graph as well as effects of the key variables on the systems performances were investigated taking the same input parameters as listed in Tables 1 and 2.

4.1 | Working fluid selection analysis for the combined cycles

4.1.1 | Combined basic recuperated sCO_2 cycle and ORC

The system's output power and thermal efficiency have both increased with solar radiation. As demonstrated in Figures 5–7, R1224yd(Z) gave the highest thermal performance among the considered working fluids. The fluid R1224yd(Z) obtained the highest exergy, thermal efficiencies, and output power values at 58.52%, 54.43%, and 293.5 kW, respectively, at the 0.95 kW/m^2 of DNI. However, working fluid R1243zf was observed as the second-best-performing fluid, while R1234yf was observed as the worst-performing working fluid. All other fluids such as R1234ze(E), R1234ze(Z), R1225ye(Z), R1233zd(E) and R1336mzz(Z) performed between these two fluids. The power output and thermal efficiency curves follow a similar trend to that of the exergy efficiency curve due to exergy and thermal efficiency is directly related [37]. With R1224yd(Z), power output increased from 183 kW to 293.5 kW, as DNI rose from 0.4 kW/m^2 to 0.95 kW/m^2. Therefore in this combined cycle R1224yd(Z) has been recommended as the best-performing fluid in the ORC system in the current combined cycle.

4.1.2 | Combined pre-compression sCO_2 cycle and ORC

Figure 8 shows that the highest exergy efficiency was found with R1336mzz(Z) as a working fluid among the other considered fluids. As solar irradiation grew from 400 to 950 W/m^2, the exergy efficiency of the system utilizing R1336mzz(Z) working fluid improved from 37.8% to 59.60%. Despite the other examined fluids, R1234ze(Z) showed the lowest exergy efficiency. At 950 W/m^2 of solar radiation, the other fluids demonstrated

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

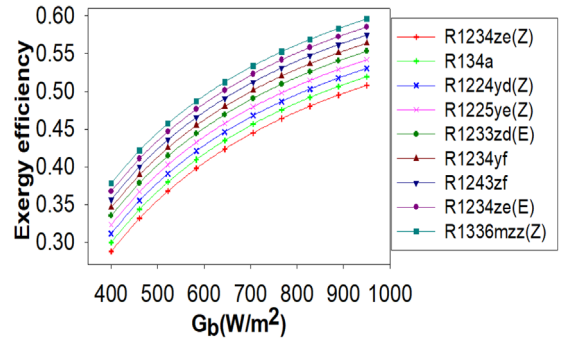


FIGURE 8 Working fluids effects of on the exergy efficiency of pre-compression cycle-ORC

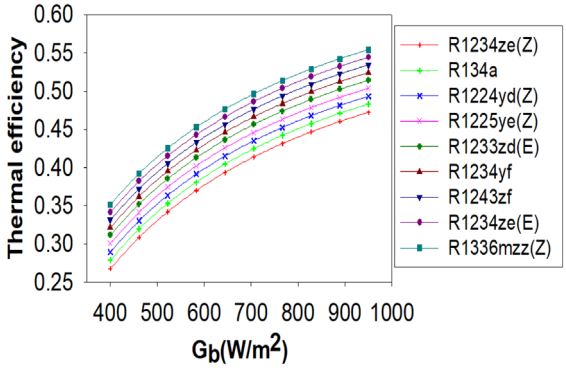


FIGURE 9 Working fluids effects on the thermal efficiency of pre-compression cycle-ORC

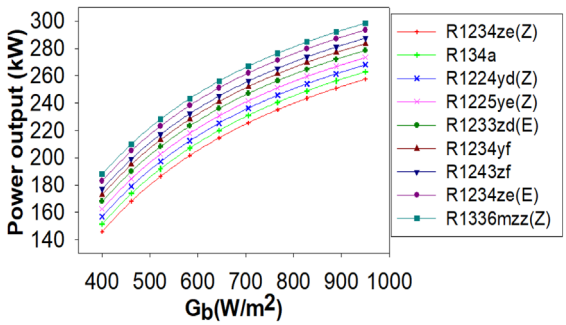


FIGURE 10 Working fluids effects on the power output of pre-compression cycle-ORC

the maximum exergy efficiency, with values of 58.79%, 57.8%, 55.98%, 54.65%, 53.87%, 52.32%, and 50.95%, respectively, for R1234ze(E), R1243zf, R1234yf, R1233zd(E), R1225ye(Z), and R1224yd(Z) and R134a. In this examination, it can be seen that all HFO working fluids, except for R1234ze(Z), performed more efficiently than HFC working fluids like R134a. Figure 9 shows that, among the various working fluids used, R1336mzz(Z) provided the maximum thermal efficiency. Based on R1336mzz(Z), the highest thermal efficiency increased by 58.77% when solar irradiation increased from 400 W/m^2 to 950 W/m^2. In addition, Figure 10 illustrates how solar radiation enhanced the power generation of the combined cycle.

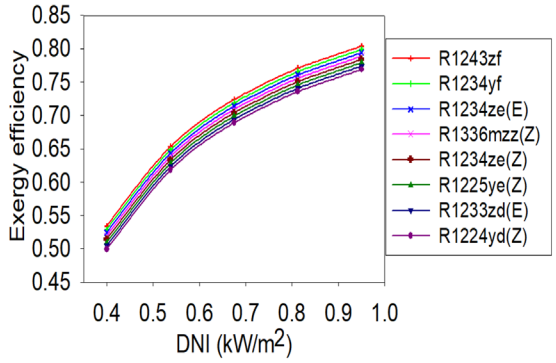


FIGURE 11 Working fluids effects on exergy efficiency of RCMCIC cycle-PDORC

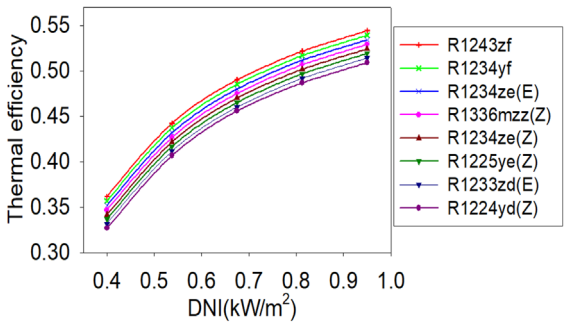


FIGURE 12 Working fluids effects on the thermal efficiency of RCMCIC cycle-PDORC

The enthalpy at the turbine inlet increased due to increased solar irradiation. This leads to improved enthalpy differences across the turbine. This enables an increase in output power overall. Maximum output power was obtained as 298.5 kW with R1336mzz(Z) at the DNI of 950 W/m^2. Therefore, R1336mzz(Z) has been recommended as the best-performing fluid in the ORC of the combined pre-compression cycle-ORC. R1336mzz(Z) fluid has zero ODP and less than 1 GWP which reduces the environmental effects. However, R1234ze(E) was observed as the second best-performing fluid among the other considered working fluids.

4.1.3 | Combined RMCIC sCO_2 cycle and PDORC

The thermal and exergy efficiency of this integrated system grew steadily as solar radiation rose with given fluids while fixing all the other remaining factors as shown in Tables 1 and 2. According to the analysis, R1243zf working fluid had the maximum thermal and exergy efficiency out of all the fluids that were taken into consideration. However, R1224yd(Z) achieved the lowest thermal and exergy efficiencies, as illustrated in Figures 11 and 12, respectively. When solar irradiation rose from 0.4 to 0.95 kW/m^2, maximum thermal and exergy efficiency increased from 36.17% to 54.42% and 53.42% to 80.39%, respectively, using R1243zf as a working fluid. On the

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

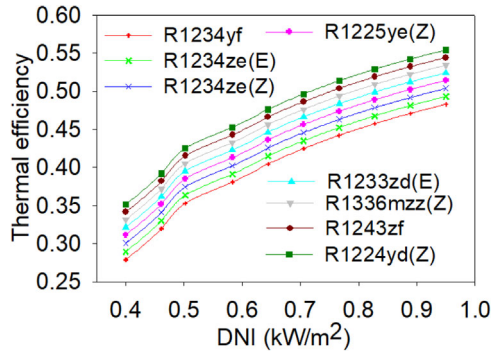


FIGURE 13 Working fluid effects on the thermal efficiency of partial cooling cycle-ORC

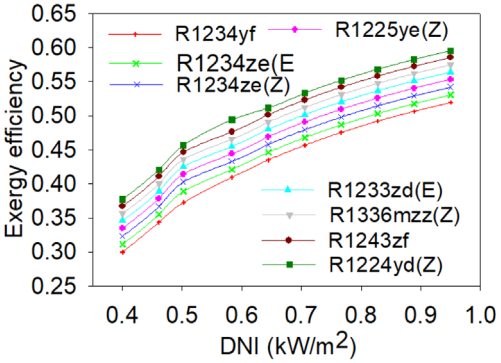


FIGURE 14 Working fluid effects on the exergy efficiency of partial cooling cycle-ORC

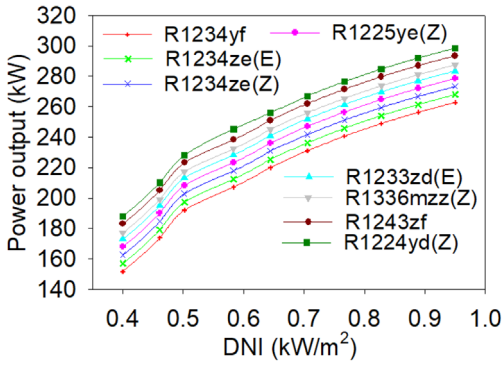


FIGURE 15 Working fluid effects on the power output of partial cooling cycle-ORC

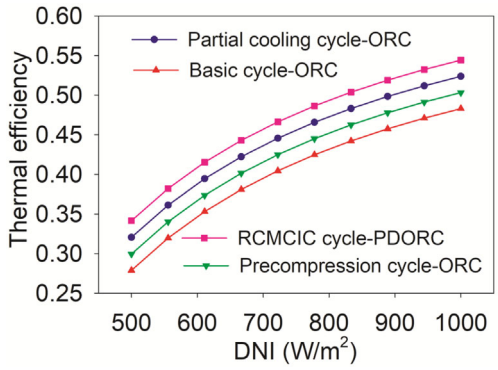


FIGURE 16 Variation and comparisons of thermal efficiencies of combined cycles with DNI

other hand, R1234yf was shown to be the second-best performing fluid. The performances of the other fluids varied between the performance of the R1243zf and R1224yd(Z) working fluids. Recommended working fluid R1243zf has zero ODP and less than 1 GWP, therefore using it could lessen its negative effects on the environment.

4.1.4 | Combined partial cooling sCO_2 cycle and ORC

In this section, the effects of various working fluids on the performance of combined partial cooling sCO_2 cycle-ORC have been investigated. Working fluid R1224yd(Z) gave the highest thermal performance among the other considered working fluids followed by the R1234yf, R1234ze(E), R1234ze(Z), R1225ye(Z), R1233zd(E), R1243zf, and R1336mzz(Z). With R1224yd(Z) working fluid, the performance of the combined partial cooling cycle-ORC was increased with solar irradiation. Based on working fluid R1224yd(Z), the thermal efficiency, exergy efficiency, and output power of the combined partial cooling sCO_2 cycle-ORC were increased from 35.16% to 55.43%, 37.73% to 59.42% and 188 kW to 298.5 kW, respectively, as DNI rose from 400 to 950 W/m^2, as indicated in Figures 13–15.

4.2 | Performance comparison of different configurations with parametric analysis

In this subsection, the performances of all four configurations were compared as well as suitable parametric analysis was also performed. Various parameters affected the performance of the combined cycles. However, here effects of only key parameters were investigated such as solar irradiation (DNI), compressor inlet temperature (CIT), and turbine inlet temperature (TIT).The performances of all combined cycles were compared taking its best recommended working fluid as observed in the above section.

From Figures 16-18, it was observed that the performances of all combined cycles increased with the solar irradiation (DNI). The reason is the same for all the cycles; it can be explained as the current integrated models are powered by an SPT plant. The exergy efficiency of the cycles rose gradually as solar radiation increased. The exergy flow rate was enhanced by the effective utilization of the increased solar irradiation by the solar field. It also increased the receiver efficiency; consequently, combined cycle performance was enhanced [28]. The highest thermal performance was obtained by the combined RCMCIC cycle-PDORC among the other considered cycles. However, the combined partial cooling cycle-ORC obtained the second highest performance followed by pre-compression

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

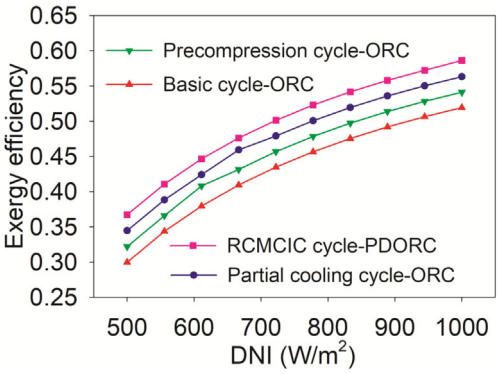


FIGURE 17 Variation and comparisons of exergy efficiencies of combined cycles with DNI

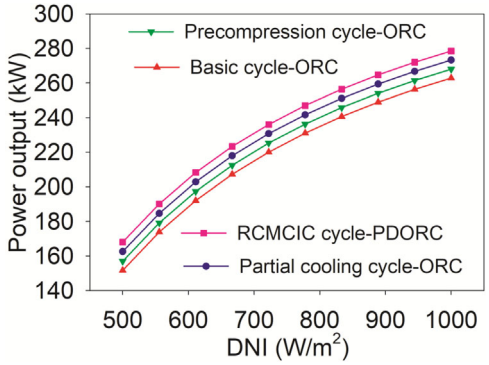


FIGURE 18 Variation and comparisons of the power output of combined cycles with DNI

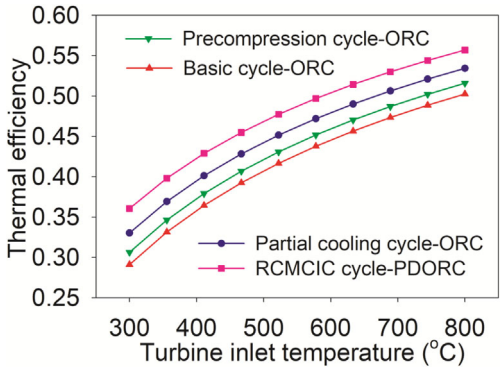


FIGURE 19 Variation and comparisons of thermal efficiencies of combined cycles with TIT

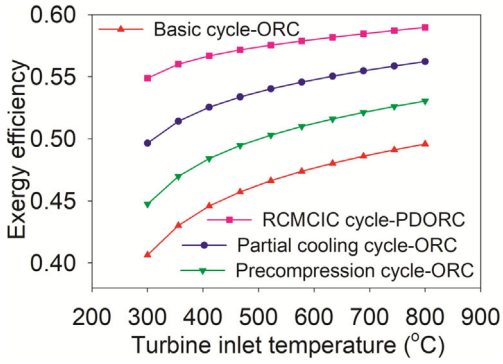


FIGURE 20 Variation and comparisons of exergy efficiencies of combined cycles with TIT

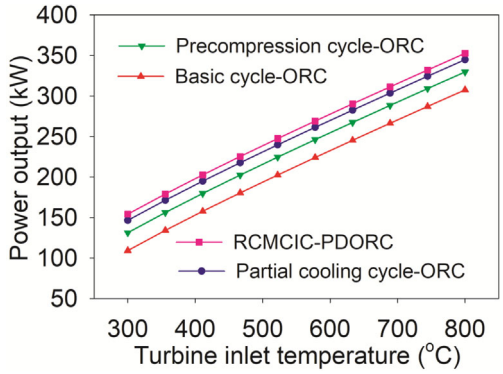


FIGURE 21 Variation and comparisons of output power combined cycles with TIT

cycle-ORC and lowest by basic cycle-ORC. The reason is that PDORC has recovered more waste heat from the RCMCIC topping cycle due to its double evaporator. PDORC recovered heat from the intercooler and HEX2 simultaneously due to its two evaporators. It was already explained that stand stand-alone RCMCIC cycle performed better than the basic cycle. At the given operating conditions (base conditions) as listed in Tables 1 and 2, the thermal efficiency of the combined basic cycle-ORC, combined pre-compression cycle-ORC, combined partial cooling cycle-ORC and RCMCIC cycle-PDORC were obtained as 45.75%, 47.78%, 49.84% and 51.9%, respectively, and corresponding exergy efficiency were obtained as 49.19%, 51.37%, 53.59%, 55.84, respectively. Apart from this corresponding power outputs were obtained as 248.9 kW, 254.2 kW, 259.4 kW, and 264.7 kW, respectively (See Figures 18).

Figures 19–21 show the impacts of the turbine inlet temperature (TIT) on the combined cycle's performance. Performance of the cycles increased with TIT. As the TIT increased the enthalpy difference across the turbine increased, consequently the output power increased. Therefore, both the thermodynamic efficiencies increased. Thermal efficiency of the combined basic cycle-ORC, combined pre-compression cycle-ORC, combined partial cooling cycle-ORC, and RCMCIC cycle-PDORC were increased from 29.10% to 50.25%, 30.63% to 51.59%, 33.02% to 53.43% and 36.05% to 55.7%, respectively, and corresponding exergy efficiency was increased 40.64–49.58%, 44.75–53.05%, 49.65–56.22% and 54.88–58.97%, respectively, as TIT increased from 300°C to 800°C. Apart from this corresponding power outputs were increased from 109.3 kW to 307.6 kW, 131.3 kW to 329.7 kW, 146.6 kW to 344.9 kW, and 154.2 kW to 352.7 kW, respectively, as TIT increased from 300°C to 800°C. Therefore, apart from the solar heliostats field, these cycle configurations can be driven

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

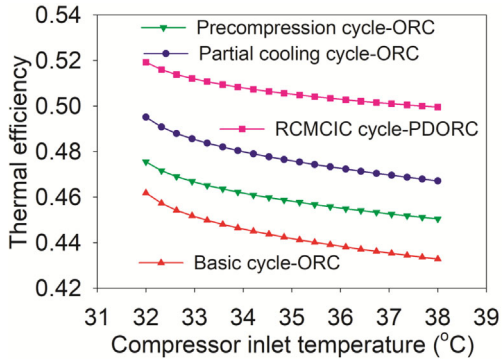


FIGURE 22 Variation and comparisons of thermal efficiencies of combined cycles with CIT

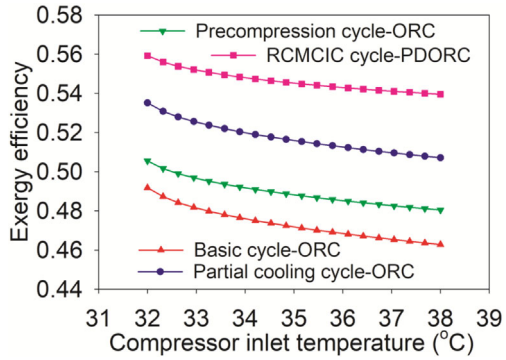


FIGURE 23 Variation and comparisons of exergy efficiencies of combined cycles with CIT

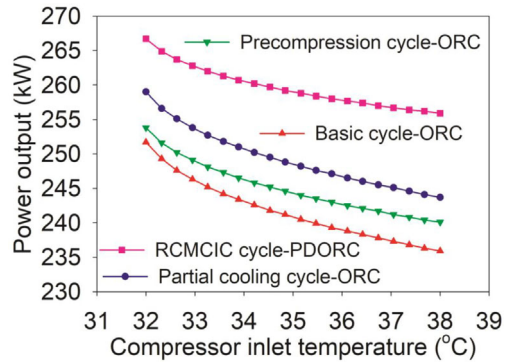


FIGURE 24 Variation and comparisons of the output power of combined cycles with CIT

by a wide range of other heat sources such as parabolic trough collectors, geothermal energy, nuclear energy etc.

Figures 22–24 discuss variations of the performance with the compressor inlet temperature (CIT). CIT much affected the performance of the cycles. Here CIT also refers to the main compressor temperature or minimum temperature of the cycles. The thermodynamic performance of the combined cycles decreases with the increase in CIT. An increase in CIT means an increase in enthalpy difference across the compressor that leads to an improvement in the compressor work. Therefore, the net power output of the cycle decreases, and consequently performance of the system decreases. Thermal efficiency of the combined basic cycle-ORC, combined pre-compression cycle-ORC, combined partial cooling cycle-ORC and RCMCIC cycle-PDORC were decreased from 46.18% to 43.28%, 47.55% to 45.04%, 49.51% to 46.71% and 51.92% to 49.95%, respectively, and corresponding exergy efficiency was decreased 49.18% to 46.28%, 50.55% to 48.04%, 53.51% to 50.71% and 55.92% to 53.95%, respectively, as CIT increased from 32°C to 38°C. Apart from this corresponding power outputs decreased from 251.7 kW to 235.9 kW, 253.8 kW to 240.1 kW, 259 kW to 243.7 kW, and 266.7 kW to 255.9 kW, respectively, as CIT increased from 32°C to 38°C. Therefore, it was seen that an improvement in the compressor temperature, and decrement in the performance of the configuration RCMCIC cycle-PDORC was less as compared to the other cycles. This is due to the double turbine in PDORC compensating for the power loss in the combined cycle. Therefore, thermodynamically this system was observed as the best-performing combined cycle for harvesting solar heat from the SPT plant.

Table 5 reveals the summary of the results of the present comparative study. All the considered combined cycle's performances at the given operating conditions have been listed in Table 5. Additionally, the best-performing fluids with their GWP values have been presented there. It can be seen the best performing combined cycle is RCMCIC cycle-PDORC for the SPT plant among other considered combined cycles. Its thermal efficiency, exergy efficiency, and power output values were observed as 51.9%, 55.84%, and 264.7 kW, respectively. Correspondingly, R1243zf was found to be the best-performing fluid with a low GWP of less than 1. However, in the case of the Basic cycle-ORC, Pre-compression cycle-ORC, and Partial cooling cycle-ORC, the best-performing fluids were observed as R1224yd(Z), R1336mzz(Z), R1224yd(Z), respectively. The RCMCIC cycle-PDORC's thermal efficiency was 13.44% better than that of the basic conventional cycle (i.e. basic cycle-ORC). However, the thermal efficiency of pre-compression cycle-ORC and partial cooling cycle-ORC were found as 4.43% and 8.93% better than that of basic cycle-ORC.

5 | CONCLUSIONS

In the present comparative study, the four configurations of the combined sCO_2 cycle and ORC were considered to utilize solar energy efficiently through the solar power tower system. A comparative analysis was performed using the EES computational technique. Simultaneously, low GWP fluids were considered to reduce global warming and environmental effects. The following conclusions were drawn from the present research;

- The combined RCMCIC sCO2 cycle-PDORC was found as the best configuration for utilizing solar energy through the CSP system among the considered configurations.
- The combined partial cooling sCO2 cycle was considered the second most efficient configuration for the utilization of solar energy from the CSP system.

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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 5 Summary of results of present comparison study of combined sCO_2 power cycles

S.N.	Combined cycles	η_{th} (%)	η_{ex} (%)	Power output (kW)	η_{th} improvement (%)	Best performing fluid	Corresponding GWP
1	Basic cycle-ORC	45.75	49.19	248.9	–	R1224yd(Z)	0.88
2	Pre-compression cycle-ORC	47.78	51.37	254.2	4.43	R1336mzz(Z)	<2
3	Partial cooling cycle-ORC	49.84	53.59	259.4	8.93	R1224yd(Z)	0.88
4	RCMCIC cycle-PDORC	51.9	55.84	264.7	13.44	R1243zf	<1

- Combined RCMCIC sCO2 cycle-PDORC was determined to have the highest exergy and thermal efficiency, measuring 55.84% and 51.9%, respectively.
- Additionally, fluid R1243zf was found as the best-performing low GWP fluid of combined RCMCIC sCO2 cycle-PDORC.
- The proposed efficient power generation system is free from any environmental effects due to using low GWP fluids.

NOMENCLATURE

A	Area (m^2)
ĖD	Exergy destruction rate (kW)
Ė	Exergy rate (kW)
f_{View}	View factor of the receiver
G_b	Solar irradiation (W/m^2)
h_{conv}	Convection heat loss coefficient (W/m^2-K)
$\dot{m}$	Mass flow rate (kg/s)
N_h	Number of heliostat
$\dot{Q}$	Heat rate (kW)
T_R	Receiver surface temperature (K)
$\dot{W}$	Power (kW)
Cp	Constant pressure specific heat (kJ/kg-K)
H	Specific enthalpy (kJ/kg)
S	Specific entropy (kJ/kg-K)
sCO_2	Supercritical CO_2
T	Temperature (K)

Greek letters

β	Half angle of cone subtended by sun (rad)
δ	Change in property
ζ	Thermal emittance
α	Solar absorbance
η	Efficiency
σ	Stephen Boltzmann constant (W/m)
ε	Effectiveness

Subscript

0	Environmental conditions
b	Boiling
c	Critical
e	Exit
ex	Exergy
h	Heliostat
i	Inlet/state
ms	Molten salt
r	Receiver
th	Thermal

AUTHOR CONTRIBUTIONS

Yunis Khan: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; validation; visualization; writing—original draft; writing—review & editing. **Hakan Caliskan**: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; validation; visualization; writing—original draft; writing—review & editing. **Hiki Hong**: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing—original draft; Writing—review & editing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Hakan Caliskan https://orcid.org/0000-0002-6571-0965

REFERENCES

1. Zhou, J., Ali, M.A., Zeki, F.M., et al.: Thermoeconomic investigation and multi-objective optimization of a novel efficient solar tower power plant based on supercritical Brayton cycle with inlet cooling. Therm. Sci. Eng. Prog. 39, 101679 (2013)
2. Khan, Y., Raman, R., Rashidi, M.M., et al.: Thermodynamic analysis and experimental investigation of the water spray cooling of photovoltaic solar panels. J. Therm. Anal. Calorimetry 148, 5591–5602 (2023). https://doi.org/10.1007/s10973-023-12119-y
3. Khan, Y., Mishra, R.S., Singh, A.P.: Performance comparison of organic Rankine cycles integrated with solar-based combined cycle: A thermodynamic and exergoenvironmental analysis. Proc. Inst. Mech. Eng. Part C:J. Mech. Eng. Sci. 1–16 (2023). https://doi.org/10.1177/09544062231167069

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

4. Goyal, A., Sherwani, A.F.: Analysis of solar energy driven organic Rankine cycle-vapor compression refrigeration system. Therm. Sci. Eng. Prog. 35, 101477 (2022)
5. Kalogirou, S.A.: Solar thermal collectors and applications. Prog. Energy Combust. Sci. 30(3), 231–295 (2004)
6. Fuqiang, W., Ziming, C., Jianyu, T., et al.: Progress in concentrated solar power technology with parabolic trough collector system: A comprehensive review. Renew. Sust. Energ. Rev. 79, 1314–1328 (2017)
7. Farges, O., Bézian, J., El-Hafi, M.: Global optimization of solar power tower systems using a Monte Carlo algorithm: Application to a redesign of the PS10 solar thermal power plant. Renew. Energy 119, 345–353 (2018)
8. Khatoon, S., Kim, M.: Performance analysis of carbon dioxide based combined power cycle for concentrating solar power. Energy Convers. Manage. 205, 112416 ((2020)). https://doi.org/10.1016/j.enconman.2019.112416
9. Khan, Y., Mishra, R.S.: Performance investigation of the solar power tower-driven combined cascade supercritical CO_2 cycle and organic Rankine cycle using HFO fluids. Aust. J. Mech. Eng. 1–15 (2022). https://doi.org/10.1080/14484846.2022.2030087
10. Ma, Y., Morozyuk, T., Liu, M., et al.: Optimal integration of recompression supercritical CO_2 Brayton cycle with main compression intercooling in solar power tower system based on exergoeconomic approach. Appl. Energy 242, 1134–1154 (2019)
11. Khan, Y., Mishra, R.S.: Performance comparison of basic and parallel double evaporator Organic Rankine Cycle integrated with solar-based supercritical CO_2 cycle. J. Therm. Eng. 9(3), 565–579 (2023)
12. Adnan, M., Zaman, M., Ullah, A., et al.: Thermo-economic analysis of integrated gasification combined cycle co-generation system hybridized with concentrated solar power tower. Renew. Energy 198, 654–666 (2022)
13. Ahn, Y., Bae, S.J., Kim, M., et al.: Review of supercritical CO2 power cycle technology and current status of research and development. Nucl. Eng. Technol. 47, 647–661 (2015)
14. Khan, Y., Mishra, R.S.: Parametric (exergy-energy) analysis of parabolic trough solar collector-driven combined partial heating supercritical CO_2 cycle and organic Rankine cycle. Energy Sources Part A 1–27 (2020). https://doi.org/10.1080/15567036.2020.1788676
15. Khan, Y., Mishra, R.S.: Performance evaluation of solar-based combined pre-compression supercritical CO_2 cycle and organic Rankine cycle. Int. J. Green Energy 18(2), 172–186 (2021)
16. Qin, L., Xie, G., Ma, Y., et al.: Thermodynamic analysis and multi-objective optimization of a waste heat recovery system with a combined supercritical/transcritical CO_2 cycle. Energy 265, 126332 (2023)
17. Khan, Y., Mishra, R.S.: Thermo-economic analysis of the combined solar-based pre-compression supercritical CO_2 cycle and organic Rankine cycle using ultra-low GWP fluids. Therm. Sci. Eng. Prog. 23, 100925 (2021)
18. Khan, Y., Mishra, R.S.: Performance analysis of solar-driven combined recompression main compressor intercooling supercritical CO_2 cycle and organic Rankine cycle using low GWP fluids. Energy Built Environ. 3(4), 496–507 (2022)
19. Kim, M.S., Ahn, Y., Kim, B., et al.: Study on the supercritical CO_2 power cycles for landfill gas firing gas turbine bottoming cycle. Energy 111, 893–909 (2016)
20. Zhu, H., Xie, G., Yuan, H., et al.: Thermodynamic assessment of combined supercritical CO_2 cycle power systems with organic Rankine cycle or Kalina cycle. Sustainable Energy Technol. Assess. 52, 102166 (2022)
21. Nondy, J., Gogoi, T.K.: Exergoeconomic investigation and multi-objective optimization of different ORC configurations for waste heat recovery: A comparative study. Energy Convers. Manage. 245, 114593 (2021)
22. Wang, Z., Chen, H., Xia, R., et al.: Energy, exergy and economy (3E) investigation of a SOFC-GT-ORC waste heat recovery system for green power ships. Therm. Sci. Eng. Prog. 32, 101342 (2022)
23. Mahmoud, M., Naher, S., Ramadan, M., et al.: Investigation of a ground-cooled organic Rankine cycle for waste heat recovery. Int. J. Thermofluids 18, 100348 (2023)
24. Hoang, A.T.: Waste heat recovery from diesel engines based on Organic Rankine Cycle. Appl. Energy 231, 138–166 (2018)
25. Song, J., Li, X., Ren, X., et al.: Performance analysis and parametric optimization of supercritical carbon dioxide (S-CO2) cycle with bottoming Organic Rankine Cycle (ORC). Energy 143, 406–416 (2018)
26. Dai, Y., Hu, D., Wu, Y., et al.: Comparison of a basic organic rankine cycle and a parallel double-evaporator organic rankine cycle. Heat Transfer Eng. 38, 990–999 (2017)
27. Kulhanek, M., Dostal, V.: Supercritical carbon dioxide cycle thermodynamic analysis and comparison. In: Proceedings of Scco2 Power Cycle Symposium (2011)
28. Al-Sulaiman, F.A., Atif, M.: Performance comparison of different supercritical carbon dioxide Brayton cycles integrated with a solar power tower. Energy 82, 61–71 (2015)
29. Ma, Y., Liu, M., Yan, J., et al.: Thermodynamic study of main compression intercooling effects on supercritical CO_2 recompression Brayton cycle. Energy 140, 746–756 (2017)
30. Wang, K., He, Y.L., Zhu, H.H.: Integration between supercritical CO_2 Brayton cycles and molten salt solar power towers: A review and a comprehensive comparison of different cycle layouts. Appl. Energy 195, 819–836 (2017)
31. Padilla, R.V., Too, Y.C.S., Benito, R., et al.: Exergetic analysis of supercritical CO_2 Brayton cycles integrated with solar central receivers. Appl. Energy 148, 348–365 (2015)
32. Besarati, S.M., Goswami, D.Y.: Analysis of advanced supercritical carbon dioxide power cycles with a bottoming cycle for concentrating solar power applications. J. Solar Energy Eng. 136, 010904 (2014)
33. Ho, C.K., Iverson, B.D.: Review of high-temperature central receiver designs for concentrating solar power. Renew. Sust. Energ. Rev. 29, 835–846 (2014)
34. Wang, X., Liu, Q., Lei, J., et al.: Investigation of thermodynamic performances for two-stage recompression supercritical CO_2 Brayton cycle with high-temperature thermal energy storage system. Energy Convers. Manag. 165, 477–487 (2018)
35. Neises, T., Turchi, C.: A comparison of supercritical carbon dioxide power cycle configurations with an emphasis on CSP applications. Energy Procedia 49, 1187–1196 (2014)
36. Reyes-Belmonte, M.A., Sebastian, A., Romero, M., et al.: Optimization of a recompression supercritical carbon dioxide cycle for an innovative central receiver solar power plant. Energy 112, 17–27 (2016)
37. Cengel, Y.A., Boles M.A. (eds.): Thermodynamics: An Engineering Approach, 5th ed. McGraw-Hill publication, New York (2004)
38. Petela, R.: Exergy analysis of the solar cylindrical-parabolic cooker. Sol. Energy 79, 221–233 (2005)
39. Klein, S.A.: Engineering equation solver (EES), Academic Commercial V7.714. F-Chart Software (2022). www.fChart.com
40. Yan, Y., Yang, F., Zhang, H., et al.: Study on performance evaluation framework and design/selection guidelines of working fluids for subcritical organic Rankine cycle from molecular structure perspective. Energy 282, 128582 (2023)
41. Bahrami, M., Pourfayaz, F., Kasaeian, A.: Low global warming potential (GWP) working fluids (WFs) for Organic Rankine Cycle (ORC) applications. Energy Rep. 8, 2976–2988 (2022)
42. Moloney, F., Almatrafi, E., Goswami, D.Y.: Working fluids parametric analysis for the regenerative supercritical organic Rankine cycle for medium geothermal reservoir temperatures. Energy Procedia 129, 599–606 (2017)
43. Calm, J.M.: Referegerant safety. ASHRAE J. 36(7), 17–26 (1994)

How to cite this article: Khan, Y., Caliskan, H., Hong, H.: A comparative study of combined cycles for concentrated solar power for efficient power generation using low Global Warming Potential (GWP) fluids to reduce environmental effects. IET Renew. Power Gener. 1–14 (2023). https://doi.org/10.1049/rpg2.12866

17521424, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/rpg2.12866 by Delhi Technological University, Wiley Online Library on [06/12/2023]. 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

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Comprehensive Review on Permissions-Based Android Malware Detection

Yash Sharma (ysharma2098@gmail.com)
Delhi Technological University
Anshul Arora
Delhi Technological University

Research Article

Keywords: Android Security, Android Malware, Permissions Based Detection, Static Detection, Mobile Security, Literature Review.

Posted Date: November 29th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3657506/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Comprehensive Review on Permissions-Based Android Malware Detection

Yash Sharma and Anshul Arora†

[1]Department of Applied Mathematics, Delhi Technological University, Delhi, -110042, India.

*Corresponding author(s). E-mail(s): ysharma2098@gmail.com;
Contributing authors: anshul15arora@gmail.com;
†These authors contributed equally to this work.

Abstract

The first Android-ready "G1" phone debuted in late October 2008. Since then, the growth of Android malware has been explosive analogous to the rise in the popularity of Android. The major positive aspect of Android has been its open-source nature, which empowers app developers to expand their work. But at the same time, authors with malicious intentions pose grave threats to users. In the presence of such threats, Android malware detection is the need of an hour. Consequently, researchers have proposed various techniques involving static, dynamic, and hybrid analysis to address such threats using numerous features in the last decade. But the feature that most researchers have extensively used to perform malware analysis and detection in Android security is Android permission. Hence, to provide a clarified overview of the latest and past work done in Android malware analysis and detection, we perform a comprehensive literature review using permissions as a central feature or in combination with other components by collecting and analyzing 200 studies from January 2009 to February 2023. We extracted information such as the choice opted by researchers between analysis or detection, techniques used to select or rank the permissions feature set, features used along with permissions, detection models employed, the malware datasets used by researchers, and lastly, the limitations and challenges in the field of Android malware detection to propose some future research directions. Additionally, based on the information extracted, we answer the six research questions designed considering the above factors.

Keywords: Android Security, Android Malware, Permissions Based Detection, Static Detection, Mobile Security, Literature Review.

1 Introduction

In the last decade, we have witnessed the exponential growth of the Android operating system in the mobile market. According to a recent report, the Android system constitutes more than 80% of the whole market of smart phones[1]. The main reason behind Android's success is its free, open-source code, which empowers smartphone manufacturers to transform their devices with pre-installed applications and customized user interfaces for a beautiful customer experience. But Android's

[1]www.tenda.com.cn

open-source nature is a boon and a bane. On the one hand, where it brings the benefits of technological broadband and update, it also allows criminals to use it for ill practices. These days mobile phones are not only used for communication purposes, but gradually they have become a crucial part of our lives containing the smallest to the most critical and private user data. In such a situation, robust and effective Android detection mechanisms are the need of the hour.

Android OS was released in 2008, and just then, the first Android malware was spotted in 2010, which targeted users by subscribing to premium SMS services. Since then, malware attacks have been on the rise, and at the same time, security attempts have been fighting to keep up with the ever-increasing and constantly changing malicious attacks. The total number of Android malware worldwide have already increased from 22,088 in 2012 to 33,237,653 in Jan 2023[2]. Looking closer at the real-time threat analysis and statistics of Android malware worldwide, we will understand how desperately the Android Market needs Android security and malware detection systems.

Researchers and practitioners use various analysis and detection measures to address the above-mentioned concerning numbers. Based on the standard research techniques in the literature, Android malware detection is usually carried out using three analysis techniques: static, dynamic, and hybrid. The static analysis aims to investigate malware without executing the actual code but by collecting basic information about the app's functionality. In contrast, dynamic analysis performs and monitors an application to track its nature to find traces of malicious behavior. The working of hybrid analysis works in a way such that it combines the advantages of both static and dynamic analysis. Although static analysis poses some limitations while dealing with advanced malicious deformation techniques such as dynamic code loading, static analysis still proves to be quite efficient, usable, and scalable in prohibiting malware before execution. Amongst all the static features, the most popular and commonly used static feature is permissions. Google introduced a permission system for the Android OS, making it mandatory for all developers to define the necessary permissions required for the functionality of their product. It is up to the user to grant or deny access to the requested permissions during the installation. Hence, monitoring the usage of permissions before installing any application can prevent the spread of malware. It may seem challenging to a standard user, but the researchers have used the meaning, frequency, and combination of permissions requested over the past 13 years to build robust Android malware detection models.

Table I shows that this review offers much more than most previous works in this area. This review is not based on a general topic like [5], but instead focuses systematically on applying for permissions as a feature in Android Malware analysis/detection. Some surveys and reviews exist in the literature, such as [13] and [12], to analyze the work done in Android malware detection using analysis techniques, features, and machine learning models. A systematic literature review focusing on the trend of static analysis and the usage of machine learning models, respectively, was covered in the two surveys mentioned above. However, there has been a gap in the Android malware detection investigation in recent years. More specifically, a survey highlighting the popularity of Android permissions as a feature in Android malware analysis/detection is missing. Therefore, it is indispensable to summarize the related work in the field of Android malware detection using permissions from the advent of the first proposed Android malware-related model.

To have a clear view of the usage and popularity of permissions as a feature in Android malware analysis and detection over almost 14 years (2009-2023), we conducted this Comprehensive Literature Review (CLR) after analyzing the related work thoroughly. The main contributions of this work are highlighted below :

- We perform this CLR using a vast dataset of 200 research papers that aim to use permissions for Android malware analysis/detection, almost covering the advent of Android OS [14] and the first malware in 2009 to the current research scenario in the second month of 2023 [15].
- Apart from obtaining information about the usage and permission-based techniques, we

[2]https://portal.av-atlas.org/malware

Table 1: Comparison of recent reviews having similar coverage with this article.
(✓= Having content, x = having no or little content)

Reference	Feature Analysis	Technique used	Model used (ML/DL)	Analysis/ Detection	Methodology Comparison	Dataset used	Discussion on Limitations and Future Directions
[1]	✓	x	✓	x	x	x	✓
[2]	✓	x	✓	x	x	x	x
[3]	✓	x	✓	x	x	✓	x
[4]	✓	x	✓	x	x	✓	x
[5]	x	x	x	x	x	x	x
[6]	✓	x	x	x	x	x	x
[7]	x	x	✓	x	x	x	x
[8]	✓	x	✓	x	x	x	✓
[9]	x	x	✓	x	x	✓	✓
[10]	x	x	x	x	x	x	x
[11]	✓	x	x	x	✓	x	✓
[12]	✓	✓	✓	x	x	✓	✓
This article	✓	✓	✓	✓	✓	✓	✓

comprehensively analyzed other features combined with permissions, detection models, and datasets used.
- According to the results of empirical evidence, permissions prove to be quite informative and efficient as a means of detecting malware in Android smartphones.
- Finally, we make the discussions about the results, limitations and possible future work directions using permissions as a feature.

1.1 Review Protocol

In this section, we discuss in detail the main steps involved in conducting this CLR :

1. Research questions: After reviewing the related work and locating the different queries that need to be analyzed in this review, we put forward six research questions and their answers that further perform the basis of the following sections. Table 2 presents six research questions about the usage of permissions in Android malware detection.
2. Search Strategy: Our first step while building a comprehensive literature survey is collecting related work by various authors in Android security, mainly based on permissions. Our work revolves around the usage and impact of permissions in the area of Android malware detection; hence we chose the related work that has used permissions alone or combined with other static/dynamic features to analyze or detect malware in Android smartphones. We identified several search sources and search items to cover the related research between the period of 2009 to 2023 and, in the end, selected seven electronic databases, which include main journals and conferences, namely -
   - IEEE Xplore Digital Library
   - ScienceDirect
   - ACM Digital Library
   - Wiley Online Library
   - Google Scholar
   - SpringerLink
   - Web of Science
3. Data extraction and synthesis: The last step of our survey, which is the data extraction and synthesis, is directly related to the first step, i.e., the research questions. According to the information extracted from various research papers, we can find the answers to the research

Table 2: Research Questions

ID	Research Questions	Motivation
RQ1	What was the primary underlying purpose of using permissions defined by the researchers- behavioral analysis or malware detection?	Assess the goals of researchers and practitioners.
RQ2	What feature ranking, selection, or other techniques are used to build an Android malware analysis/detection system considering permissions?	Identify the commonly used method for selecting and extracting relevant features.
RQ3	Which features are primarily used in combination with permissions for Android malware analysis/detection?	Identify the commonly used features along with permissions.
RQ4	Which ML/DL/other models are used for Android malware analysis/detection?	Compare the popularity of ML and DL models.
RQ5	Which datasets are used for malware analysis/detection?	Identify the most famous experimental datasets.
RQ6	What are the limitations, challenges, and future directions for permissions-based Android malware analysis or detection?	Assess the limitations or challenges and consequently, propose future research directions.

questions and queries. We extracted the following information from all the research papers used for this review -

- Purpose of research - Choice between creating an Android malware detection model or analyzing benign and malware apps.
- Technique used - As described above, our focus relies upon permissions; hence we obtained information on how authors utilized permissions to build their malware analysis or detection model.
- Features used - Most common features chosen and utilized by authors while carrying out their research work in combination with permissions.
- Type of model used - Choice of model and classifiers preferred between Machine learning (ML) and Deep learning (DL) or any other kind if used for analysis or detection.
- Malware dataset used - What malware datasets are used by the authors in their corresponding works?
- Limitations - Limitations of detection/analysis models mainly associated with permissions as a feature in the field of Android security?
- Future Directions - Assessing the limitations and challenges faced by the permissions-based Android malware analysis/detection models, we propose some future research directions.

2 Purpose of Research

In response to RQ1, we present the data summarizing the choice preferred by authors in Table 3, between only analyzing the permissions in malware/benign apps or detecting malicious behavior in applications by proposing full-fledged malware detection using a dataset comprising of both benign and malware applications. The analysis process closely observes the behavioral pattern of features requested by applications. It could be performed using both applications classes [16] or even individual

types too [17]. On the other hand, a research paper falls under the detection category if the corresponding authors have performed detection using a mixed or unlabelled dataset after completing the analysis or training phase.

Table 3 indicates that over the span of around 14 years, almost all the researchers and practitioners aimed to detect malware in Android smartphones using the datasets available in the market comprising both normal and malware applications. But some authors chose to carry out only the analysis of feature behavior; for instance, the authors in [18] analyzed descriptions of apps downloaded from the Google Play Store to predict the requested permissions but didn't use any malware applications. If we try to come to a conclusion using the data highlighted by Table 3 to understand the situation better, we can observe that a little over 93% of the work mentioned in the table revolved around malware detection using permissions and not only analyzing apps' permissions. **Hence, based on the results presented above, we answer the first Research Question that the majority of researchers have chosen the path of building permissions-based Android malware detection models instead of merely analyzing the permissions of malware.**

Table 3: Purpose of Research (Analysis or Detection)

Related works	Analysis		Malware Detection
	Normal	Malware	
Zhang et al. [19]	✓	✓	✓
Li et al. [20]	✓	✓	✓
Sahin et al. [21]	✓	✓	✓
Talha et al. [22]	✓	✓	✓
Varma et al. [23]	✓	✓	✓
Mahindru et al. [24]	✓	✓	✓
Dogru et al. [25]	✓	✓	✓
Rathore et al. [26]	✓	✓	✓
Shang et al. [27]	✓	✓	✓
Tchakounte et al. [28]	✓	✓	✓
Ju et al. [16]	✓	✓	
Ilham et al. [29]	✓	✓	✓
Sahin et al. [30]	✓	✓	✓
Angelo et al. [31]	✓	✓	✓
Xiong et al. [32]	✓	✓	✓
Lu et al. [33]	✓	✓	✓
Kavitha et al. [34]	-	-	-
E. Amer [35]	✓	✓	✓
Chakravarty et al. [36]	✓	✓	✓
Pondugula et al. [37]	✓	✓	✓
Sahal et al. [38]	✓	✓	✓
Tuan Mat et al. [39]	✓	✓	✓
Wang et al. [40]	✓	✓	✓
Park et al. [41]	✓	✓	✓
Liang et al. [42]	✓	✓	✓
Enck et al. [14]	✓	-	-
Enck et al. [17]	✓	-	-
Wang et al. [43]	✓	✓	-
Peng et al. [44]	✓	✓	-
Pandita et al. [45]	✓	-	-
Samra et al. [46]	✓	-	-
Yerima et al. [47]	✓	✓	✓
Aung et al. [48]	✓	✓	✓
Yerima et al. [49]	✓	✓	✓
Sanz et al. [50]	✓	✓	✓
Moonsamy et al. [51]	✓	✓	✓
Backes et al. [52]	✓	-	-
Wu et al. [53]	✓	✓	✓
Kato et al. [54]	✓	✓	✓
Arora et al. [55]	✓	✓	✓
Alsoghyer et al. [56]	✓	✓	✓
Saleem et al. [57]	✓	✓	-
Ghasempour et al. [58]	✓	✓	✓
Shrivastava et al. [59]	✓	✓	✓
Upadhayay et al. [60]	✓	✓	✓
Lee et al. [61]			
Surendran et al. [62]	✓	✓	✓
A. T. Kabakus [63]	✓	✓	✓
Wang et al. [64]	✓	✓	✓
Akbar et al. [65]	✓	✓	✓
Zhu et al. [66]	✓	✓	✓
Wang et al. [67]	✓	✓	✓
N. McLauglin [68]	✓	✓	✓
Wang et al. [69]	✓	✓	✓
Grace et al. [70]	✓	✓	✓
Liu et al. [71]	✓	✓	✓
Bayazit et al. [72]	✓	✓	✓
Lee et al. [73]	✓	✓	✓
Zhu et al. [74]	✓	✓	✓

3 Techniques used

In response to RQ2, this section aims to present the various feature selection or ranking or other similar techniques used by researchers to exploit permissions while building an analysis/detection model. Be it feature selection or feature ranking, researchers aim to select features highly dependent on the response. Various feature reduction techniques have been used in primary studies to determine and choose significant features in Android malware detection. But some authors have approached the detection process in different ways too. Hence, we divided all the techniques into five categories: feature ranking, frequent patterns-based, graph-based approach, feature selection, and others. Some of the most commonly used methods are discussed in brief below.

(a) Frequency-based techniques - Frequency-based techniques generally fall under the

Related works	Analysis		Malware Detection
	Normal	Malware	
Almahmoud et al. [75]	✓	✓	✓
Feng et al. [76]	✓	✓	✓
Kandukuru et al. [77]	✓	✓	✓
Arora et al. [78]	✓	✓	✓
Ding et al. [79]	✓	✓	✓
Sahin et al. [80]	✓	✓	✓
Idrees et al. [81]	✓	✓	✓
Khariwal et al. [82]	✓	✓	✓
Idrees et al. [83]	✓	✓	✓
Zhu et al. [15]	✓	✓	✓
Bai et al. [84]	✓	✓	✓
Taheri et al. [85]	✓	✓	✓
Alazab et al. [86]	✓	✓	✓
Mathur et al. [87]	✓	✓	✓
Imtiaz et al. [88]	✓	✓	✓
Liu et al. [89]	✓	✓	✓
Chen et al. [90]	✓	✓	✓
Guan et al. [91]	✓	✓	✓
Mohamed et al. [92]	✓	✓	✓
Varma et al. [93]	✓	✓	✓
Gyunka et al.[94]	✓	✓	✓
Taha et al. [95]	✓	✓	✓
Peng et al. [96]	✓	✓	✓
Ashwini et al. [97]	✓	✓	✓
Jiang et al. [98]	✓	✓	✓
Wang et al. [99]	✓	✓	✓
Rana et al. [100]	✓	✓	✓
Lu et al. [101]	✓	✓	✓
Millar et al. [102]	✓	✓	✓
Barrera et al. [103]	✓	-	-
Shabtai et al. [104]	✓	-	-
Felt et al. [105]	✓	-	-
Erickson et al. [106]	✓	-	-
Sarma et al. [107]	✓	✓	✓
Frank et al. [108]	✓	-	-
Jhu et al. [109]	✓	✓	✓
Peiravian et al. [110]	✓	✓	✓
Sanz et al. [111]	✓	✓	✓
Feldman et al. [112]	✓	✓	✓
Pehlivan et al. [113]	✓	✓	✓
Rahman et al. [114]	✓	✓	✓
Rovelli et al. [115]	✓	✓	✓
Arp et al. [116]	✓	✓	✓
Yerima et al. [117]	✓	✓	✓
Kang et al. [118]	✓	✓	✓
Zhao et al. [119]	✓	✓	✓
Qiao et al. [120]	✓	✓	✓
Chen et al. [121]	✓	✓	✓
Demertzis et al. [122]	✓	✓	✓
Verma et al. [123]	✓	✓	✓
Wang et al. [124]	✓	✓	✓
Tangil et al. [125]	✓	✓	✓
Wang et al. [126]	✓	✓	✓
Li et al. [127]	✓	✓	✓
Bhattacharya et al. [128]	✓	✓	✓
Xie et al. [129]	✓	✓	✓
Xie et al. [130]	✓	✓	✓
Ren et al. [131]	✓	✓	✓
Tao et al. [132]	✓	✓	✓
Namrud et al. [133]	✓	✓	✓
Alswaina et al. [134]	✓	✓	✓
Qiu et al. [135]	✓	✓	✓
Zhu et al. [136]	✓	✓	✓
Feng et al. [137]	✓	-	-

Related works	Analysis		Malware Detection
	Normal	Malware	
Aonzo et al. [138]	✓	✓	✓
Urooj et al. [139]	✓	✓	✓
Wang et al. [140]	✓	-	-
Wang et al. [141]	✓	✓	✓
Zhang et al. [142]	✓	-	-
Kesswani et al. [143]	✓	-	-
Ibrahim et al. [144]	✓	✓	✓
Arshad et al. [145]	✓	✓	✓
Yuan et al. [146]	✓	✓	✓
Zhou et al. [147]	✓	✓	✓
Cilleruelo et al. [148]	✓	✓	✓
Firdaus et al. [149]	✓	✓	✓
Wang et al. [150]	✓	✓	✓
Singh et al. [151]	✓	✓	✓
Rafiq et al. [152]	✓	✓	✓
Mahdavifar et al. [153]	✓	✓	✓
Seraj et al. [154]	✓	✓	✓
Mahindru et al. [155]	✓	✓	✓
Sahin et al. [21]	✓	✓	✓
Anupama et al. [156]	✓	✓	✓
Chen et al. [157]	✓	✓	✓
Mahindru et al. [158]	✓	✓	✓
Tchakounté et al. [159]	✓	✓	✓
Nissim et al. [160]	✓	✓	✓
Peynirci et al. [161]	✓	✓	✓
Nauman et al. [162]	✓	✓	✓
Bhattacharya et al. [163]	✓	✓	✓
Bao et al.[164]	✓	-	-
Medrano et al. [165]	✓	-	-
Mat et al. [165]	✓	✓	✓
Shatnawi et al. [166]	✓	✓	✓
Smmarwar et al. [167]	✓	✓	✓
Arif et al. [168]	✓	✓	✓
Manzanares et al. [169]	✓	✓	-
Bhat et al. [170]	✓	✓	✓
Elayan et al.[171]	✓	✓	✓
Syrris et al. [172]	✓	✓	✓
Idrees et al. [173]	✓	✓	✓
Rehman et al. [174]	✓	✓	✓
Martin et al. [175]	✓	✓	✓
Navarro et al. [176]	✓	✓	✓
Milosevic et al. [177]	✓	✓	✓
Alzaylaee et al. [178]	✓	✓	✓
Cai et al. [179]	✓	✓	✓
Badhani et al. [180]	✓	✓	✓
Hijawi et al. [181]	✓	✓	✓
Sheen et al. [182]	✓	✓	✓
Nisha et al. [183]	✓	✓	✓
Song et al. [184]	✓	✓	✓
Zhang et al. [185]	✓	✓	✓
Yang et al. [186]	✓		-
Thiyagarajan et al. [187]	✓	✓	✓
Qaisar et al. [188]	✓	✓	✓
Appice et al. [189]	✓	✓	✓
Zhu et al. [190]	✓	✓	✓
A. Altaher [191]	✓	✓	✓
Su et al. [192]	✓	✓	✓
Mahindru et al. [193]	✓	✓	✓
Dehkordy et al. [194]	✓	✓	✓
Nguyen et al. [195]	✓	✓	✓
Taheri et al. [196]	✓	✓	✓

Related works	Analysis		Malware Detection
	Normal	Malware	
Mahesh et al. [197]	✓	✓	✓
Firdaus et al. [198]	✓	✓	✓
Shrivastava et al. [199]	✓	✓	✓
Varsha et al. [200]	✓	✓	✓
M. Deypir [201]	✓	✓	✓
Mahindru et al. [202]	✓	✓	✓
Keyvanpour et al. [203]	✓	✓	✓
Razak et al. [204]	✓	✓	✓
Xie et al. [205]	✓	✓	
Mahindru et al. [206]	✓	✓	✓
Alecakir et al. [18]	✓	-	-
Ali et al. [207]	✓	-	-
Sun et al. [208]	✓	✓	✓
AlJarrah et al. [209]	✓	✓	✓
Gharib et al. [210]	✓	✓	✓
Sun et al. [211]	✓	✓	✓

category of feature ranking techniques; however, they can sometimes be used as a feature selection technique too. The main underlying concepts that these techniques tend to exploit are -

- Some features are frequently requested by only one class of dataset, either benign or malware, and as the goal is generally to differentiate between normal and malicious applications, the features that are frequently requested only by the malware applications are considered dangerous, and the ones having a high frequency in the case of normal applications are considered to have low-risk factor [58].
- Some features are commonly used by both the classes, benign and malware, and hence they can be excluded to choose only the more informative features. For instance, permissions such as "INTERNET" are frequently requested by both malware and benign apps; hence the authors in [20] chose to eliminate such permissions as they might introduce ambiguity in the malware detection process. Moreover, they ranked the features based on their frequency of usage in malicious and benign apps.
- Apart from the two approaches mentioned above, the authors also focus on how often one permission or feature is repeated in the whole dataset or has a low support value. In such cases also, the frequency is used as a parameter to rank and further select only the relevant features.

(b) Information Gain (Information Gain) - Information Gain can be defined as a measure of reduction in entropy. One can understand it as a measure of reduction in the amount of information upon splitting a dataset according to a certain value of a random variable. Information Gain is inversely proportional to entropy, i.e., the higher the information gain, the lower the entropy of that particular group, as the element of surprise would be less. Mathematically, Information Gain is calculated by comparing the entropy of the dataset before and after a transformation. Common usage of Information Gain includes forming decision trees from a training dataset. Information Gain is calculated for each variable in the first step, followed by selecting the variable with the maximum Information Gain value and thus minimizing the entropy and obtaining the best splits of the dataset for efficient classification.

As feature ranking is a subset of feature selection, Information Gain can be used as both techniques, i.e., to reduce the dataset and simultaneously choose only the most informative features. This is done by calculating the gain of each variable in the context of the target variable. Information Gain value for each independent attribute is calculated and further ranked from top to bottom, the top being the most relevant with the highest Information Gain score. After this, a threshold could be decided to filter out only the features with Information Gain values above the threshold, which can be further included in the machine or deep learning classifiers.

(c) Principal Component Analysis (PCA) - Another commonly used feature selection technique in Android malware detection is Principal Component Analysis, which works on reducing the number of variables while preserving as much information as possible. PCA identifies the correlations to calculate the eigenvectors and eigenvalues, which identifies the principal components in return. After that, one can choose which

principal components to keep and which ones to discard by creating a feature vector.

(d) Chi-square - Chi-square is generally used in statistics to test the independence of two events, but quite often, it has also been used as a feature selection technique in Android malware detection to reduce the size of the feature set. Mathematically, Chi-square measures how expected count E and observed count O deviate from each other. The Chi-square formula [212] is defined in the below equation.

$$\chi_c{}^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:
c=Degrees of freedom,
O=Observed value(s), and
E=Expected value(s)

While aiming for feature selection, the features having higher chi-square values are selected for model building as a higher chi-square value depicts higher dependence over the target variable or the response.

(e) Clustering - Clustering is another technique that has been extensively used in Android malware detection first to understand a dataset and further club the similar and less informative features. In terms of machine learning, unlabelled grouping examples are called clustering.

In response to RQ2, we conclude by the information depicted in Table 4 that the most commonly used techniques involving the use of permissions are Information Gain and frequency-based, e.g., [58] and [20]. Apart from these, most researchers based their approach around utilizing either ML or DL classifiers by tuning the hyperparameters, e.g., [196] and [171]. Figure 1 gives a better and closer understanding of Table 4. As we can see, close to 48% of the total research works used in this review based their approach on either ML or DL classifiers/techniques [110]-[115]. Either they fed the extracted features directly to the machine learning classifiers to be dealt with or used techniques like gain ratio, correlation coefficient [29], mutual information, relief [36]

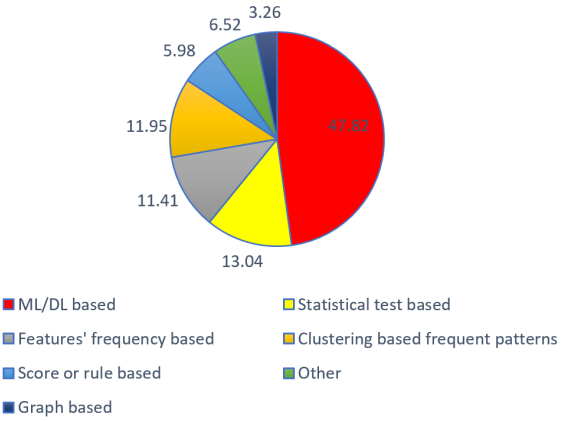


Fig. 1: Statistics depicting the most commonly used techniques to build an Android malware analysis/detection system considering permissions

etc., to compute the feature score. The second most common approach used by researchers and practitioners to reduce or choose the best set of features is utilizing the statistical tests (13.04%) such as Chi-square [119], PCA [34], Mann–Whitney test [132], variance threshold [26], ANOVA [120] and many more. Around 11% of the total work account for approaches based upon features' frequency [19]-[20] and clustering-based frequent patterns [46] each. 5.98% papers utilised some sort of similarity score [22] or rule based techniques [17] to analyse or detect Android malware. Lastly, 6.52% and 3.26% of the comprehensive studies used graph-based [185], [186], and some other techniques, respectively. For instance, in [31], the authors mapped the permissions on the x-y plane using their corresponding protection level, whereas in [45], they utilized app descriptions to deduce the permissions required by an application. **Hence, based on the results presented above, we answer the second Research Question that popular techniques to handle the extracted permissions and other features are the ML and DL classifiers as their hyper parameters can be easily tuned to reduce, rank, or select the unlabelled or unbalanced feature set.**

Table 4: Technique used

Related works	Feature Ranking	Frequent Patterns	Feature Selection	Graph based	Others
Zhang et al. [19]		Monitored the permission usage within the application and with the system			
Li et al. [20]	Ranking based on the frequency of permissions being requested.				
Sahin et al. [21]			Linear Regression		
Talha et al. [22]					Combined Risk score calculated for each app
Varma et al. [23]					Used permissions as features to study the performance of ML algorithms
Mahindru et al. [24]			Discarded the ones not installed or starting at launching stage		
Dogru et al. [25]					Permission groups score calculated, to sum up an app's Risk Score
Rathore et al. [26]			Variance threshold, autoencoders, and PCA		
Shang et al. [27]			Reduced the permission set with Pearson's Correlation Coefficient		
Tchakounte et al. [28]		Similarity score based on sequence alignment			
Ju et al. [16]		Manual pattern recognition to existing malware permissions patterns			
Ilham et al. [29]			Gain Ratio, Information Gain, Correlation Coefficient, CFS subset Evaluator		
Sahin et al. [30]					Relevance Frequency
Angelo et al. [31]					Mapped the permissions on the x-y plane using their corresponding protection level
Xiong et al. [32]		Used unique and common permissions patterns from both datasets as weak Classifier			
Lu et al. [33]			Improved RF algorithms along with introducing fuzzy sets of samples		
Kavitha et al. [34]			PCA and Sequential Forward selection, Limiting the permissions by accepting or denying each permission separately according to Dangerous level		
E. Amer [35]					Developed an ensemble comprising multiple classifiers
Chakravarty et al. [36]			Information Gain, Relief, Gain Ratio		
Pondugula et al. [37]					Deep Neural Network model

Sahal et al. [38]	Ranking based on permissions class frequency				
Tuan Mat et al. [39]			Used Bayes classifier after optimising features using Chi-Square Test		
Wang et al. [40]			Association rule Mining, PCA , Deep Cross Network		
Park et al. [41]			Reduced features by removing built-in, custom, dangerous, and permissions that are used at least once		
Liang et al. [42]		Generated k maps for permission combinations based on their usage			
Enck et al. [14]					Presented analysis of the newly launched Android OS in 2009
Enck et al. [17]		Defined rules based upon dangerous level and possible negative impact			
Wang et al. [43]		Calculated the risk score of each app using Baye's rule based upon frequency of permissions and risk levels			
Peng et al. [44]					Calculated the risk scores of applications using probabilistic methods like Naïve Bayes and its modifications
Pandita et al. [45]					Introduced a framework based upon natural language processing called "WHYPER" which used application descriptions to describe which permissions are needed and why
Samra et al. [46]		Permissions are used as features to make clusters using K-means cluster algorithm			
Yerima et al. [47]	Permissions are used as features along with others to be fed into the Bayesian Classifier				
Aung et al. [48]			K best features selected using Information Gain		
Yerima et al. [49]	Permissions are used as features along with others to be fed into the Bayesian Classifier				
Sanz et al. [50]					Work based upon permission frequency
Moonsamy et al. [51]			Biclustering method to visualize for rare, unique as well as frequent patterns, followed by reducing permissions based on their support difference between normal and malicious datasets		

Backes et al. [52]					Appguard, a powerful system capable of modifying user-defined security policies on untrusted applications and converting them into new trusted ones
Wu et al. [53]		Used permissions, and other information to form a feature set, further applied K-means and EM algorithm to form clusters for malware detection			
Kato et al. [54]					Calculated similarity score between malware and normal permission pairs after dividing them into different categories
Arora et al. [55]				Used permission pairs to construct graphs for normal and malware apps	
Alsoghyer et al. [56]					Malware detection model based on the frequency of permissions occurrences followed by using machine learning algorithm to assess the model
Saleem et al. [57]					Calculated the first four moments for permission's binary data after using Kernel Density Estimation and used the varying values as a means to distinguish between applications
Ghasempour et al. [58]	Permissions ranked by using frequency-based weighting method		PCA followed by using a statistical method based upon eigenvalues and eigenvectors		
Shrivastava et al. [59]					Risk score was calculated based on permissions frequency for each permission which in turn was used to classify applications as high risk, medium risk, and low risk
Upadhayay et al. [60]	Ranked the permissions based on the frequency of their occurrence in the datasets and combined them with the best network traffic features				
Lee et al. [61]					Classified normal and malicious apps through ML-based detection techniques based on the frequency of permission of Android apps
Surendran et al. [62]					Fed into the Logistic Regression (LR) classifier first followed by applying the naïve bayes classifier to find interdependency between features
A. T. Kabakus [63]					Fed into the Convolutional Neural network (CNN) model as one-dimensional input to form a training model

Wang et al. [64]			Decision Tree, Extra trees, Chi-square test, Genetic algorithm, SVM based on recursive feature elimination, MI		
Akbar et al. [65]			Employed Random Forest to generate the feature importance and combined it with Google's dangerous permissions and the features used in [66] to select the best set of permissions		
Zhu et al. [66]			Used TF-IDF and Cosine similarity to choose the best features		
Wang et al. [67]			Fed the permissions data combined with the API sequence data into the LR model		
N. McLauglin [68]					Used permissions as features in PerceiverIO blocks to be combined with the opcode PerceiverIO blocks
Wang et al. [69]	Ranked using Mutual Information (MI), correlation coefficient and T-test				
Grace et al. [70]		Risk level on an unknown application is decided by checking for dangerous permission requests			
Liu et al. [71]		Sensitive permission patterns are extracted	Filtered out using improved FP growth, removing clusters having same support and low JARO distance followed by performing hierarchical clustering		
Bayazit et al. [72]					Fed as input into the Recurrent neural Networks (RNN)-based classifiers
Lee et al. [73]			Information Gain and genetic algorithm		
Zhu et al. [74]					Fed into deep learning hybrid methods such as unsupervised method Merged Sparse Auto-Encoder (MSAE) and supervised method Stacked Denoising Auto-encoders (SDAE) and fed the extracted results to SVM and KNN
Almahmoud et al. [75]	Ranked on the basis of cosine similarity				
Feng et al. [76]			Chi-square test and extremely randomized tree method		
Kandukuru et al. [77]					Computed the permission's score using Jaccard-bitwise similarity technique to compare an app's risk score with the threshold

Arora et al. [78]		Combined the permissions and network traffic feature values and generated the frequent patterns using FP- Growth algorithm			
Ding et al. [79]			Chi-square test, analysis of variance (ANOVA) F-value, and MI		
Sahin et al. [80]					Fed into linear regression based classifiers
Idrees et al. [81]		Distinguishing frequency pattern ranges			
Khariwal et al. [82]	Information Gain				
Idrees et al. [83]		Constructed a detection matrix			
Zhu et al. [15]					Fed into Multi-Head Squeeze-and-Excitation Residual block (MSer), and stacked it to construct a deep network MSerNet
Bai et al. [84]			Fast Correlation-Based Filter by [213]		
Taheri et al. [85]			Random Forest Algorithm		
Alazab et al. [86]			Information Gain		
Mathur et al. [87]			Frequency counting, backward elimination and collinearity check		
Imtiaz et al. [88]					Fed into a deep learning artificial neural network classifier
Liu et al. [89]			Fed into RBM combined with subspace methods to reduce the feature dimensionality after choosing the best subspaces by clustering techniques		
Chen et al. [90]			PCA		
Guan et al. [91]			k-means clustering algorithm		Fed the results to Synthetic Minority Over-Sampling Technique (SMOTE)
Mohamed et al. [92]			Chose the most common features		
Varma et al. [93]			Bat Optimization, Cuckoo Search, and Grey Wolf Optimization wrapper feature selection techniques		
Gyunka et al.[94]			PCA		
Taha et al. [95]		Clustering similar Permissions			
Peng et al. [96]			Adaptive shrinkage CNN		
Ashwini et al. [97]					Fed the features to machine learning classifiers
Jiang et al. [98]	The dangerous features backtracked from sensitive API calls are ranked in order of Information Gain score				

Wang et al. [99]					Fed to various classifiers to create an ensemble classifier based upon selective Ensemble method and genetic algorithm
Rana et al. [100]			Selected only the most useful feature word by creating a dictionary		
Lu et al. [101]					Fed the permissions to DBN classifier
Millar et al. [102]					Fed the permissions to a CNN classifier
Barrera et al. [103]		Found permissions usage pattern			Fed to the SOM to analyze permissions usage pattern
Shabtai et al. [104]	Information Gain, Fisher Score and Chi-Square				
Felt et al. [105]					Compared permissions required to invoke API methods and actually requested permissions to check for over privilege issues
Erickson et al. [106]					Mapped permissions required to invoke API methods to check for privacy leaks
Sarma et al. [107]		Considered the permission patterns from their malware dataset and some critical most requested permissions to generate risk signals for the users			
Frank et al. [108]		To extract statistically significant permission request patterns			
Jhu et al. [109]					To compare the permissions required by an application by analyzing the description given by the developer with its actual requested permissions
Peiravian et al. [110]					Fed to various machine learning classifiers
Sanz et al. [111]					Fed to various machine learning classifiers
Feldman et al. [112]					Fed to various machine learning classifiers
Pehlivan et al. [113]			Gain Ratio Attribute Evaluator, Relief Attribute Evaluator, Cfs Subset Evaluator, Consistency Subset Evaluator		
Rahman et al. [114]					Fed to various machine learning classifiers
Rovelli et al. [115]					Fed to various machine learning classifiers on the server side component
Arp et al. [116]		Mapped to a joint vector space, where patterns were analyzed geometrically			
Yerima et al. [117]					Fed to various machine learning classifiers

Kang et al. [118]		Calculated the likelihood ratio under the given distribution of permissions and further used the Needleman-Wunsch algorithm to calculate similarity score			
Zhao et al. [119]			Chi-square and Information Gain		
Qiao et al. [120]			ANOVA and SVM—Recursive Feature Elimination		
Chen et al. [121]					Fed to various machine learning classifiers
Demertzis et al. [122]					Fed to ELM classifier and eSNN along with various hardware components
Verma et al. [123]					Extracted the functional call graph of the applications, used a procedure inspired by the neighborhood hash graph kernel (NHGK) and create a graph classification problem
Wang et al. [124]	Absolute permission rate difference				Feature vector was created composed of permissions, receiver actions, and hardware components
Tangil et al. [125]	Extra Tree algorithm and rank them by mean decrease impurity				
Wang et al. [126]					*FrequenSel* [119] and Information Gain
Li et al. [127]			PCA Method based on singular value decomposition (SVD)		
Bhattacharya et al. [128]	Information Gain				
Xie et al. [129]					Fed the results of syntax and semantic features to the machine learning classifiers
Xie et al. [130]					Feature vectors were reshaped to matrices by embedding to feed to CNN classifier
Ren et al. [131]					Fed the features to various machine learning and ensemble classifiers
Tao et al. [132]			Mann–Whitney test to analyze statistical significance of permission usage in both the datasets		
Namrud et al. [133]		Analysed the permissions usage pattern w.r.t different categories by using a combination of SOM and K-means			
Alswaina et al. [134]			Extremely Randomized Trees		

Qiu et al. [135]					Created a Feature vector table with the help of TF-IDF technique to form binary type vectors with security/privacy-related capabilities as annotations
Zhu et al. [136]					Fed the extracted features to an ensemble classifier of MLP and SVM as fusion classifier
Feng et al. [137]					App descriptions pre-processed by NLP methods are fed into the Neural network model to produce binary probability distribution for each permission
Aonzo et al. [138]					Fed the features to their own proposed linear and nonlinear classifiers
Urooj et al. [139]					Fed the features to various machine learning and ensemble-based classifiers
Wang et al. [140]					Processed the user reviews by using permissions docs, API docs, and app descriptions to infer the permissions required by an application
Wang et al. [141]					Fed the extracted features to various machine learning classifiers
Zhang et al. [142]					Utilised Whole call graphs (WCG) and parsed the mapping list to locate sensitive operations, form User-aware Call Graph, and perform static analysis
Kesswani et al. [143]		Divided the permissions under generic and privacy-invasive categories to further calculate the percentage of generic and privacy-invasive permissions of an unknown application			
Ibrahim et al. [144]					Fed to the deep learning classifier, in particular to embedding layers, followed by clustering and flattening layers to make their shape appropriate
Arshad et al. [145]					Fed the static and dynamic features to a machine learning classifier
Yuan et al. [146]					Fed the static and dynamic features to a deep learning classifier
Zhou et al. [147]					Fed the extracted features to deep learning based classifier
Cilleruelo et al. [148]			Reduce the multiple permissions that are the same but present differences based on the application package		

Firdaus et al. [149]			Generic Search		
Wang et al. [150]		Support-based permission candidate method to mine unique required or used permission patterns			
Singh et al. [151]			BI-Normal Separation (BNS), MI, Relevancy Score (RS), and the Kullback-Leibler (KL)		
Rafiq et al. [152]					Fed the extracted features to machine learning classifiers tuned by using NIAs
Mahdavifar et al. [153]					Fed the extracted features to a deep learning classifier
Seraj et al. [154]					Fed the extracted permissions to a MLP neural network classifier
Mahindru et al. [155]			Chi-Square, Gain Ratio, Filtered Subset selection, Information Gain feature, LR analysis, PCA		
Sahin et al. [21]			Relevance frequency feature selection (RFFS), Document frequency thresholding (DF), Information Gain, Chi-square, Odds ratio (OR), IDF and other filter-based methods		
Anupama et al. [156]	Fischer Score				
Chen et al. [157]					Fed the extracted permissions after removing the useless ones to RF machine learning classifier
Mahindru et al. [158]	Chi-squared test, Information Gain feature evaluation, LR analysis, Information Gain, oneR feature evaluation, PCA		T-test, Pearson's correlation Coefficient, Rough set analysis (RSA), Consistency subset evaluation approach, Filtered subset evaluation		
Tchakounté et al. [159]					Permissions and other features are utilized to form fuzzy-hashed signatures of known malware to find similarity score between them and unknown applications
Nissim et al. [160]					Fed the extracted features to SVM and a couple of proposed active learning methods for detection after processing
Peynirci et al. [161]			*Delta IDF* based upon differential inverse document frequency (IDF) values		
Nauman et al. [162]					Fed the extracted features to various Deep learning and machine learning classifiers
Bhattacharya et al. [163]			Improvised Particle Swarm optimization (PSO) algorithm for a rough set with a new random key encoding method		

Bao et al.[164]		Traced the API - permission patterns to predict the permissions for an app by using API			Fed into the naive Bayes multinomial classification model for text classification and predicting permissions for an application
Medrano et al. [165]		Taint tracking to analyze and find mappings between Android Class Function (ACFs) and the Permissions			
Mat et al. [165]			Information Gain and chi-square		
Shatnawi et al. [166]	LR model		Recursive Feature Elimination (RFE)		
Smmarwar et al. [167]			Binary Grey Wolf Optimization (BGWO)		
Arif et al. [168]			Information Gain		
Manzanares et al. [169]					Extracted permissions and other static and dynamic features to prepare a comprehensive dataset
Bhat et al. [170]	Information Gain		Deleted the features that are too infrequent and the ones that are present in almost the same number in both the datasets		
Elayan et al.[171]					Fed the extracted results to various machine classifiers
Syrris et al. [172]			Removed the features having low variance		
Idrees et al. [173]	Information Gain				Found the correlation between permissions and intents using Pearson correlation coefficient
Rehman et al. [174]					Fed the results to various machine learning classifiers to find the cosine similarity between features
Martin et al. [175]					Extracted to create a comprehensive and complete dataset, then fed the results to various machine learning classifiers for detection
Navarro et al. [176]	RF feature importance		Eliminated the linearly dependent vectors while performing the Bag of Graphs (BoG) technique	Formed Ontology-based graphs to find the relationship between permissions defined, used, interfaces protected by permissions and the resources access through them	

Milosevic et al. [177]					Fed into various machine learning and ensemble classifiers for classification and clustering
Alzaylaee et al. [178]	Information Gain				
Cai et al. [179]	Information Gain				
Badhani et al. [180]			Removed features having constant value or zero variance		Fed the features to various machine learning and ensemble classifiers. Further applied k-mode algorithm for clustering of features
Hijawi et al. [181]	Fed the features to various machine learning classifiers to rank them on the basis of their importance				
Sheen et al. [182]	Chi-Square, Relief, Information Gain				
Nisha et al. [183]			PSO, Social Spider Algorithm (SSA), and Gravitational Search Algorithm (GSA)	Represented permissions in the form of ontology graphs to understand the relationship between permissions, packages and interface classes	
Song et al. [184]		Dangerous permissions are matched with the permissions requested by an unknown application to generate a detection report and submit it to users			
Zhang et al. [185]				Identified explicit and implicit permission use points to further create permission use graphs to analyze the permissions behavioral pattern	
Yang et al. [186]				Constructed State Transition Graphs (STG) from permissions to implement breadth-first search (BFS) in the dynamic exploration phase to analyze the permission behavior	

Thiyagarajan et al. [187]	Pruning on the basis of frequency		Chi-square, Support based, association-based, PCA		
Qaisar et al. [188]			PCA, Co-relation attribute evaluation (CAE)		Constructed clusters using k-means algorithm to match up with the newly added features in the case base
Appice et al. [189]					Clustering based k- means ++ algorithm to form separate clusters for each view which were combined later using stacking-based fusion method to learn the consensus malware detection pattern
Zhu et al. [190]					Fed the results to RF-based machine learning classifier
A. Altaher [191]	Information Gain ratio				Fed the extracted permissions to an evolving fuzzy neuro inference classifier involving clustering methods
Su et al. [192]		Fed the results to a DBN learning model to obtain unique behavioral characteristics			
Mahindru et al. [193]	Chi-squared test, Information Gain feature evaluation, logistic regression analysis, Information Gain, oneR feature evaluation, PCA		T-test, Pearson's correlation Coefficient, RSA, Consistency subset evaluation approach, Filtered subset evaluation		
Dehkordy et al. [194]	Ranked features on the basis of frequency in both the datasets and removed the common or rarely used or irrelevant features				Balanced the dataset using SMOTE, Random undersampling, and Hybrid
Nguyen et al. [195]	ANOVA		DNN, SVM		
Taheri et al. [196]	Random Forest Regressor algorithm				Fed the results to a C4N and Robust-NN classifier
Mahesh et al. [197]			Minmax technique		Fed the preprocessed results to CNN-ARFO proposed classifier
Firdaus et al. [198]	Information Gain, frequency-based top range selected				Fed the results bio-inspired versions of ANN
Shrivastava et al. [199]			Defined frequency based rules defined over permissions and intents to predict malicious risk level of an application		
Varsha et al. [200]	NB, weight calculation		Entropy based Category Coverage Difference (ECCD) and Weighted Mutual Information (WI)		
M. Deypir [201]	Ranked the permission based on the entropy and Information Gain score calculated by the proposed method				Entropy-based method to calculate information gain score for permission and correspondingly risk score for an application

Mahindru et al. [202]	Chi-squared test, Information Gain feature evaluation, logistic regression analysis, Information Gain, oneR feature evaluation, PCA		T-test, Pearson's correlation Coefficient, RSA, Consistency subset evaluation approach, Filtered subset evaluation		
Keyvanpour et al. [203]			Eliminated features having frequency count and RF weigh below the standard deviation, frequency counts over groups of features		
Razak et al. [204]			Particle Swarm optimization (PSO), evolutionary computation, Information Gain		
Xie et al. [205]	Frequency based		Fisher score		Fingerprinted Android malware families based on top permission behavioral patterns
Mahindru et al. [206]			T-test, multivariate linear regression stepwise forward selection and cross-correlation		Fed the extracted features to various machine learning and ensemble classifiers
Alecakir et al. [18]					App descriptions are fed into neural network classifiers to learn and analyze the sentences and predict the permissions requested
Ali et al. [207]					Fed the permissions and other information to the SVM classifier to learn and predict preferred application permissions to its user
Sun et al. [208]					Took keywords from permissions, API calls, intents, etc as parameters and found the Keywords Correlation Distance (KCD) between them to choose the most relevant features and fed the results to SVM classifier
AlJarrah et al. [209]	Information Gain				
Gharib et al. [210]					Fed the features to Deep Auto Encoder (DAE) neural network classifier
Sun et al. [211]			Criteria based on the ratio between the number of apps in both datasets, PCA		Fed the extracted features to a Positive and Unlabelled (PU) learning classifier

```
<uses-permission android:name="android.permission.READ_PHONE_STATE"/>
<uses-permission android:name="android.permission.READ_PHONE_NUMBERS"/>
<uses-permission android:name="android.permission.RECEIVE_SMS"/>
<uses-permission android:name="android.permission.VIBRATE"/>
<uses-permission android:name="android.permission.AUTHENTICATE_ACCOUNTS"/>
```

Fig. 2: Snapshot of permissions requested by WhatsApp Messenger app

4 Features used

In response to RQ3, this section aims to present the results obtained after analyzing the types and number of features used for Android malware detection in combination with permissions. We divide the features used into two categories: Static and Dynamic.

4.1 Static features

The static analysis investigates malware without the real code or instructions being executed. It provides basic information about app functionality and collect technical indicators from the AndroidManifest.xml and other resource files. In other words, it can be defined as a source code review of an Android application file. Several reverse engineering tools like Apktool[3] or AAPT2 [4] can be used to decompile an apk and extract the features required. The features that can perform static analysis of applications are called static features. Some of the commonly used examples of static features are explained in brief below with Android permission being the most popular one -

(a) Permissions - The presence of App permissions helps the user in two ways mainly-
- By protecting access to the **Restricted data**, such as user's personal information.
- By protecting access to the **Restricted actions**, actions such as recording audio.

Because of the above two reasons, it has been made mandatory by the Android OS and the Android permission check system for all application developers to declare the list of permissions their application needs for its functionality or invoke the Android API successfully. Hence, the manifest file contains the list of all Android permissions required to run the application efficiently. Permission is declared using the <uses-permission>tag within the manifest file. For example, as shown in Figure 2, which is the snapshot of the AndroidManifest.xml file of "The WhatsApp Messenger" app, requires permissions such as "READ_PHONE_STATE", "READ_PHONE_NUMBERS", "RECEIVE_SMS", "VIBRATE" and "AUTHENTICATE_ACCOUNTS" to execute on Android smartphones. Some permissions fall under the category of install-time permissions, i.e., they are automatically granted upon the installation of the app, whereas some permissions are known as runtime permissions, which are further requested at runtime. Install-time permissions permit the app limited access to restricted data or actions that can affect the user to a minimal amount. Install-time permissions can be further divided into the following types -

- Normal permissions - These permissions present minimal risk to the user's privacy and the functionality of other apps.
- Signature permissions - These permissions are granted by the permission check system only when the requesting app is signed by the same certificate as the one that declared the permission.

Runtime permissions, often addressed as dangerous permissions, are requested at runtime by the application to ask for access to view restricted data or perform any prohibited action by presenting a runtime permission request prompt.

4. API calls- Application Programming Interfaces (APIs) act as the medium for one program to interact with another, and an API call or request can be defined as a message sent to a server asking an API to provide a service or information. After traveling from a client to an API endpoint and being received by the server, it is processed, and the request is executed in return for a response.
5. Intent: An Intent is a messaging object a developer can use to request an action from another app component. Three main fundamental use cases of intents are starting an activity, starting a service, and delivering a broadcast.
6. Opcode sequence: Opcodes can reflect the behavior pattern of an application to a certain extent using the underlying machine code, so they are often used as static features. They are extracted by decompiling the APK file and are generally obtained to facilitate the input detection model.

4.2 Dynamic features

Dynamic analysis opts for a different approach than static analysis. Instead of examining the code, it relies upon monitoring an application's behavior while it is running over any virtual or real CPU. As the name suggests, dynamic analysis is performed by analyzing the runtime behavior of applications, and the features analyzed during this process are called dynamic features, such as -

[3] https://apktool.en.lo4d.com/windows
[4] https://developer.Android.com/studio/command-line/aapt2

1. Network traffic - In simple words, network traffic is the amount of data that moves across a network during any time interval.
2. System calls - A system call provides an interface between a process and the OS. They enable the operating system's services via API for the user programs to use. Some important system calls commonly used in the OS are wait(), fork(), exec(), kill():, exit(): etc.

   Other commonly used features include hardware components, description of applications, network addresses, code-related patterns, etc.

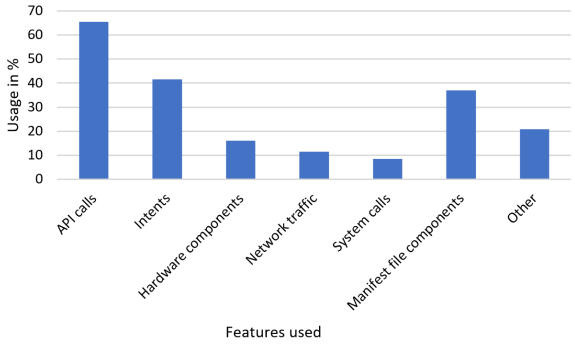


Fig. 3: Statistics depicting the usage of features in combination with permissions

Table 5 summarizes the information regarding the usage of permissions as a feature in combination with other features. Based on the information depicted by Table 5, we can conclude that in combination with permissions, API calls and intents are the most used features by researchers for Android malware analysis/detection. Figure 3 presents a statistical summary of Table 5. As already discussed above and highlighted by the figure too, API calls [125]-[136] and intents [96]-[100] account for 65.38% and 41.53% usage respectively, i.e., out of the total research papers used in this review, majority of them utilized API calls or intents in combination with permissions. The third most commonly used feature is another static feature called hardware components [194]-[195], and it accounts for 16.15% of the total. Some dynamic features such as network traffic [76]-[79] and systems calls [47]-[49] are also seldom used by researchers and practitioners in combination with permissions and their shares are 11.53% and 8.46% respectively. Apart from the static features mentioned above, many other components present in the manifest file of an apk prove to be a popular partner of permissions while building an analysis/detection model, as they account for a surprising 36.92% of the total. We note that intent and hardware components are also a part of the manifest file only, but since the researchers have extensively used them in combination with permissions, we display their usage separately instead of combining them with other rarely used Manifest file components. These features include App descriptions [109], version numbers [112], app components [116], meta information [126] and many more. Other features commonly used in combination with permissions are small file size [65], dex file [96], URLs [90], code-related information [130] etc., and they form 20.76% of the share. **Hence, based on the results presented above, we answer the third Research Question that the researchers have preferred using API calls and intents the most in combination with permissions for Android malware analysis and detection.**

5 Model used

In response to RQ4, this section describes the various ML and DL models [214] used by researchers and practitioners to conduct their research over the chosen range of reviews.

5.1 Machine Learning

Machine learning is a subset of artificial intelligence and a popular technology enabling machines to learn from past data and perform a given task automatically. Machine learning can be broadly divided into two types -

1. Supervised Machine learning
2. Unsupervised Machine learning

5.1.1 Supervised Machine learning

Just like a student who learns a concept under the supervision of a teacher, the machines are used to predict an output correctly with the help of the training data working as a supervisor that teaches the machines. The type of machine learning in which the devices are trained using well

Table 5: Features Used

Related works	Permissions	API calls	Hardware Components	Intents	Others
Zhang et al. [19]	✓	✓			
Enck et al. [14]	✓			✓	
Yerima et al. [47]	✓	✓			System calls
Yerima et al. [49]	✓	✓			System calls
Wu et al. [53]	✓	✓		✓	
Shrivastava et al. [59]	✓			✓	
Upadhayay et al. [60]	✓				Network traffic
Surendran et al. [62]	✓	✓			System calls
A. T. Kabakus [63]	✓	✓		✓	
Wang et al. [64]	✓	✓			Opcode sequences
Akbar et al. [65]	✓				Permission rate, smali file size
Zhu et al. [66]	✓	✓			Permission rate, system events.
Wang et al. [67]	✓	✓			
N. McLauglin [68]	✓				Opcode sequences
Grace et al. [70]	✓				Log file for app activities.
Liu et al. [71]	✓	✓			
Bayazit et al. [72]	✓			✓	
Lee et al. [73]	✓				
Zhu et al. [74]	✓	✓			
Almahmoud et al. [75]	✓	✓			Permission rates, system calls
Feng et al. [76]	✓			✓	Network traffic
Kandukuru et al. [77]	✓				Network traffic
Arora et al. [78]	✓				Network traffic
Ding et al. [79]	✓			✓	Network traffic
Idrees et al. [81]	✓			✓	
Khariwal et al. [82]	✓			✓	
Idrees et al. [83]	✓			✓	
Zhu et al. [15]	✓	✓	✓		
Bai et al. [84]	✓				Opcode sequences
Taheri et al. [85]	✓	✓		✓	
Alazab et al. [86]	✓	✓			
Imtiaz et al. [88]	✓	✓		✓	Network traffic
Liu et al. [89]	✓	✓		✓	System commands, opcodes, Package and FlowDroid's features [175], Network traffic
Chen et al. [90]	✓			✓	URL's, data flow features
Guan et al. [91]	✓	✓			
Mohamed et al. [92]	✓	✓			
Peng et al. [96]	✓			✓	Dex file headers, power spectrum density information of dex file structure entropy
Ashwini et al. [97]	✓	✓	✓	✓	
Jiang et al. [98]	✓			✓	
Wang et al. [99]	✓	✓		✓	
Rana et al. [100]	✓		✓	✓	
Lu et al. [101]	✓	✓			Resource, semantic and dynamic features
Millar et al. [102]	✓	✓			Opcode sequences
Shabtai et al. [104]	✓				Dex features, Xml features and apk features
Felt et al. [105]	✓	✓			
Erickson et al. [106]	✓	✓			
Jhu et al. [109]	✓				App Descriptions
Peiravian et al. [110]	✓	✓			
Sanz et al. [111]	✓		✓		
Feldman et al. [112]	✓			✓	Low version numbers, network traffic
Arp et al. [116]	✓	✓	✓	✓	App components, Network addresses
Yerima et al. [117]	✓	✓			Linux/Android commands
Kang et al. [118]	✓	✓		✓	File hash, serial number, System commands
Zhao et al. [119]	✓	✓			Action features and IP, URL features
Qiao et al. [120]	✓	✓			
Chen et al. [121]	✓	✓			Network traffic, newly defined feature *Sequence* based upon API calls
Demertzis et al. [122]	✓				Battery, Memory, CPU, Network

Verma et al. [123]	✓	✓	✓	✓	App components, network addresses
Wang et al. [124]	✓		✓		Receiver actions
Tangil et al. [125]	✓	✓		✓	Meta information, new features from app's resource files
Wang et al. [126]	✓	✓	✓	✓	Components, protected strings, IP addresses, and URL, commands
Li et al. [127]	✓	✓		✓	App components
Xie et al. [129]	✓	✓			
Xie et al. [130]	✓	✓	✓	✓	Code related patterns
Ren et al. [131]	✓	✓			Opcode sequence, hardware components
Tao et al. [132]	✓	✓			
Qiu et al. [135]	✓	✓	✓	✓	Network addresses
Zhu et al. [136]	✓	✓			Permission rate, monitoring system events, data flows
Feng et al. [137]	✓				App description
Aonzo et al. [138]	✓	✓			
Urooj et al. [139]	✓	✓		✓	App components, Packages, Receivers, services
Wang et al. [140]	✓				API docs, user reviews, app descriptions
Wang et al. [141]	✓	✓	✓	✓	Code patterns, functional call graphs
Ibrahim et al. [144]	✓	✓			Opcode sequences, application size, services, receivers, fuzzy hash
Arshad et al. [145]	✓	✓	✓	✓	Application components, system call logs, network addresses
Yuan et al. [146]	✓	✓			Dynamic App actions
Zhou et al. [147]	✓	✓			Network addresses
Cilleruelo et al. [148]	✓		✓		Information published on the Google Play Store
Firdaus et al. [149]	✓			✓	Code based features, system commands, directory path features
Wang et al. [150]	✓				
Singh et al. [151]	✓	✓			
Rafiq et al. [152]	✓	✓		✓	
Mahdavifar et al. [153]	✓	✓		✓	Packages, receivers, system calls, basic binders, composite behavior
Mahindru et al. [155]	✓	✓			User rating, number of user download apps
Anupama et al. [156]	✓				System calls
Chen et al. [157]	✓	✓			
Mahindru et al. [158]	✓	✓			
Tchakounté et al. [159]	✓				Resource names, .dex codes, source codes, package name certificate
Nissim et al. [160]	✓				Metadataa, number of activities, services, receivers, permissions, providers
Peynirci et al. [161]	✓	✓			Strings
Bao et al.[164]	✓	✓			App description
Shatnawi et al. [166]	✓	✓			
Smmarwar et al. [167]	✓	✓		✓	
Manzanares et al. [169]	✓			✓	System calls, hardware-software features etc.
Bhat et al. [170]	✓	✓	✓	✓	Network addresses
Elayan et al.[171]	✓	✓			
Syrris et al. [172]	✓	✓		✓	Actions, services, Broadcast, categories
Idrees et al. [173]	✓			✓	
Rehman et al. [174]	✓			✓	Process name
Martin et al. [175]	✓	✓		✓	Opcode sequences, system commands, Network data etc.
Navarro et al. [176]	✓				App components
Milosevic et al. [177]	✓				Source code
Alzaylaee et al. [178]	✓	✓		✓	
Cai et al. [179]	✓	✓		✓	App components, shell commands
Badhani et al. [180]	✓	✓			
Hijawi et al. [181]	✓		✓		Software components, URL, broadcast receivers features

Sheen et al. [182]	✓	✓			
Nisha et al. [183]	✓				Components and interfaces
Song et al. [184]	✓				MD5 blacklist database, permission intention
Zhang et al. [185]	✓	✓		✓	
Yang et al. [186]	✓	✓			
Qaisar et al. [188]	✓			✓	Services provider, new APK attributes
Appice et al. [189]	✓	✓		✓	Network addresses
Zhu et al. [190]	✓	✓			Permission rate, system events
Su et al. [192]	✓	✓	✓	✓	Strings, certificate-payload info, code pattern
Mahindru et al. [193]	✓	✓			Number of user download an app, rating of an app
Dehkordy et al. [194]	✓	✓	✓	✓	URLs, activity, service receiver, provider
Nguyen et al. [195]	✓	✓	✓	✓	Provider, activity, service, URLs
Taheri et al. [196]	✓	✓			
Mahesh et al. [197]	✓			✓	
Firdaus et al. [198]	✓				Directory path, telephony
Shrivastava et al. [199]	✓			✓	
Varsha et al. [200]	✓		✓	✓	Opcodes, strings, app components
Mahindru et al. [202]	✓	✓			Number of user download an app, rating of an app
Keyvanpour et al. [203]	✓	✓	✓	✓	
Razak et al. [204]	✓				
Xie et al. [205]	✓	✓	✓		
Mahindru et al. [206]	✓	✓			Number of user download an app, rating of an app
Alecakir et al. [18]	✓				App description
Ali et al. [207]	✓				Rating, comments and number of downloads related to an app
Sun et al. [208]	✓	✓		✓	Package Name
AlJarrah et al. [209]	✓	✓			Contextual information
Gharib et al. [210]	✓	✓			Logos, strings of notification messages, system call sequences
Sun et al. [211]	✓	✓			IP address, requested URLs

"labeled" training data, i.e., some input data is already tagged with the correct output, to find a mapping function between the input variable with the output variable is called supervised machine learning. Practical usage of supervised machine learning includes Risk Assessment, Image classification, Fraud Detection, and spam filtering. Supervised learning can be divided into two types -

1. Classification
2. Regression

Classification - This type of learning is used when the output variable is categorical, i.e., Yes-No, Male-Female, etc. type. Some commonly used classification algorithms are discussed in detail below -

1. Decision Tree (DT) Algorithm - Generally considered a classification problem; however, a Decision tree (DT) is a supervised learning technique that can be used both for classification and regression problems. As the name suggests, it is a tree-structured classifier utilizing the Classification and Regression Tree (CART) algorithm, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome. The main step while implementing a decision tree is to select the best attribute for the root and sub-nodes, and for that, the DT algorithm tries to maximize the information gain value, and iteratively the node having the highest information gain is chosen first. DT algorithm is further divided into three types -
   - ID3 DT (Iterative Dichotomiser 3)
   - C4.5 DT
   - J48 DT
2. Random Forest (RF) Algorithm - This type of supervised learning algorithm works on the ensemble learning concept, which includes combining multiple classifiers or *decision trees* to solve a complex problem and avoid the problem of overfitting. RF works in two parts; the first part revolves around creating the random forest by combining n number of DTs, whereas, in the second, the prediction for each tree is made.

RF and Bagging classifier (BC) come under the category of Bagging algorithms that work on the criteria of combining the results of multiple models.

3. Logistic Regression (LR) algorithm - LR predicts the output of a categorical dependent variable, but instead of giving the exact value as 0 and 1, it produces the probabilistic values between 0 and 1. In the case of LR, instead of fitting a regression line, a logistic function, also known as the sigmoid function, maps the predicted values to probabilities.
4. K-Nearest Neighbor(KNN) Algorithm - KNN works on the hypothesis that a similarity exists between the new and available cases. After analyzing the data, it puts the new case into the category that is most similar to the available ones, and this whole process takes place by calculating the Euclidean distance of K number of neighbors.
5. Support Vector Machine (SVM) - The major aim of the SVM algorithm is to build the best line or decision boundary or hyperplane capable of segregating the n-dimensional space into classes so that the new data point can be easily fitted in the correct category for future cases. SVM uses extreme points, also known as support vectors, to help create a hyperplane; hence the algorithm is called Support Vector Machine. Sequential Minimal Optimization (SMO) is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines (SVM). SVMs use the Kernel Trick and various kernel methods to transform linearly inseparable data into linearly separable data, thus finding an optimal boundary for possible outputs. Some of them are given below -
   - Linear kernel
   - Polynomial kernel (Poly)
   - Radial Basis Function (RBF)
   - Kernel change detection algorithm (KCD)
6. Naïve Bayes (NB) Algorithm - It is one of the fastest machine learning models based upon the Bayes theorem or Bayes' Rule, often used to determine the probability of a hypothesis with prior knowledge depending upon the conditional probability. The NB algorithm is mainly used in text classification problems with a high-dimensional training dataset. Three types of Naive Bayes model are commonly used in the literature -
   - Gaussian
   - Multinomial
   - Bernoulli
7. Bayesian networks (Bayesnet) - These are probabilistic graphical models that utilize Bayesian inference for probability computations. It is a directed acyclic graph in which each edge corresponds to a conditional dependency, and each node corresponds to a unique random variable.

Regression This type of learning is used when a relationship exists between the input and the output variable and the prediction is of continuous variables. Some commonly used regression algorithms are linear regression, regression trees, nonlinear regression, Bayesian linear regression, polynomial regression [214] etc.

1. Linear Regression - With the concept of performing predictive analysis and the assumption of variables having a linear relationship, linear regression finds how the value of the dependent variable changes according to the value of the independent variable. Mathematically, linear regression finds the best-fit line with the least error, which means that the Mean Squared Error (MSE) between the predicted and actual values should be minimized.

5.1.2 Unsupervised Machine Learning

In the absence of labeled data, the models are trained using the unlabelled data set and are allowed to act without supervision, and the correct corresponding technique is called unsupervised machine learning. Unlike supervised learning, unsupervised learning cannot be used directly for a regression or classification problem because the related output data doesn't exist. The main aim of unsupervised learning is to analyze and locate the underlying structure of data, categorize that data according to similarities, and present that data in a compact format. Unsupervised learning is more suited for real-world problems as we often do not have input data with the corresponding output data. Unsupervised learning can be further divided into two types -

1. Clustering
2. Association

Clustering In this type of unsupervised learning, the data points are grouped into clusters of similar data points, and this is done by finding the presence or absence of similar patterns in the unlabelled data set. Commonly used for statistical data analysis, this technique provides each cluster with a cluster ID which is further utilized by machine learning algorithms to simplify the processing of huge and complex problems. One of the most commonly used clustering algorithms is the K- Means algorithm, capable of efficiently dividing the samples into different clusters of equal variances with the linear complexity of **O(n)**. Other clustering methods used by researchers over the years include Agglomerative Hierarchical Clustering (AHC) [51] Farthest First (FF) clustering, Filtered clustering (FC), Density-Based clustering (DB) [202], Hidden Markov Model (HMM), Google Distance (GD) Clustering [208] etc.

Association Rule Learning Association rule learning aims to find relations or associations among data variables using the concept of *If and Else* statements. Association learning relies upon support, confidence, and lift to compute the associations between thousands of items. Some of the commonly used association rule learning algorithms are discussed in brief below -

1. Apriori algorithm - The algorithm relies upon breadth-first search and Hash Tree to efficiently compute and find the itemset associations after performing iterations on the large dataset.
2. F-P Growth algorithm - Frequent Pattern Growth (F-P) Algorithm revolves around finding frequent patterns without candidate generation. Unlike Apriori's functioning, which includes the generate and test strategy, it constructs an FP Tree to fragment the items' paths and generate frequent patterns.

5.1.3 Boosting in Machine learning

Combining various weak classifiers to build strong classifiers is known as the boosting technique in machine learning, and it is one of the most popular learning ensemble modeling techniques. Iterations to combine old ones and build new models are performed until training data capable of producing optimum predictions is prepared. Adaptive boosting (AdaBoost) was the first algorithm to combine various weak classifiers into a single strong classifier in the history of machine learning. Some of the other major boosting algorithms used in the literature are as follows -

- Gradient Boosting Machine (GBM)
- Extreme Gradient Boosting Machine (XGBM)
- Light GBM
- CatBoost

5.2 Deep Learning

Deep learning can be defined as a subset of machine learning as it has similar working but different capabilities and approaches. Deep learning, also known as Deep Neural Networks (DNN), is inspired by the functioning of the human brain cells/ neurons and leads to the concept of artificial neural networks capable of learning and discovering insights from data. Some popular deep learning models used by researchers over the years are discussed in brief below -

1. Convolutional Neural Network (CNN) - These networks take images as input, then assign importance, i.e., learnable weights and biases, to parts of the input to differentiate from one another. Through the help of relevant filters, a CNN successfully captures the Spatial and Temporal dependencies in an image.
2. Recurrent Neural Network (RNN) - Often used in NLP or language translation, these networks use sequential or time series data as inputs. Unlike other neural networks, the output of RNN depends on the prior elements within the sequence as it takes information from prior inputs too. Some popular variants of RNN often used in the literature are given below -
    - Long short-term memory (LSTM) - LSTM was proposed to solve the vanishing gradient and short-term memory problems. LSTM has "cells" type structures in the hidden layers of the neural network, which have three gates–an input gate, an output gate, and a forget gate capable of controlling the flow of information.
    - Gated recurrent units (GRUs) - Similar to LSTM, even GRUs are capable of dealing with the short-term memory problem of RNN models, but unlike LSTM, GRUs use hidden states and have two gates: a reset and an update gate.

3. Artificial Neural Network (ANN) - The working of ANN is similar to the working of nerve cells in the human brain. An artificial neural network has three or more layers that are interconnected. The first layer consists of input neurons. Those neurons send data to the deeper layers and then transfer the final output data to the last output layer. A fully connected multilayer neural network is called a Multi-layer Perceptron (MLP).
4. Deep Belief Neural (DBN) Networks - These are the classifiers that use layers of stochastic latent variables, which make up the network. The top couple of layers in the DBN have no direction, but the layers above them have direct links to lower layers. DBNs have an edge over traditional neural networks as they can be generative and discriminative models.
5. Deep Autoencoders (DAE) - The working of autoencoders is a bit different as they function as a special type of feedforward neural network in which the input is the same as the output. They tend to convert the information into a lower dimensional code, also known as latent-space representation, which is further used to reconstruct the result. The structure of an autoencoder comprises three parts: an encoding method, a decoding method, and a loss function.
6. Self-Organising Map (SOM) - As depicted by the name, the map organizes itself without any supervision from anyone or anything. The training of SOM is done through a fierce neural network, a single-layer feedforward network that resembles the brain's functioning. Apart from being used as a classifier, it can also be used as a non-linear dimensionality reduction technique.

Table 6 summarizes researchers' choice between ML, DL, or any other algorithm. In response to RQ4, we conclude by the information in Table 6 that most researchers and practitioners opted for traditional supervised machine learning algorithms like DT, SVM, RF, and NB due to their convenient and simple working. But nowadays, researchers have realized that machine learning has some limitations too. It relies too much on human intervention, so it might not work well with complex feature engineering problems like Natural Language Processing (NLP),

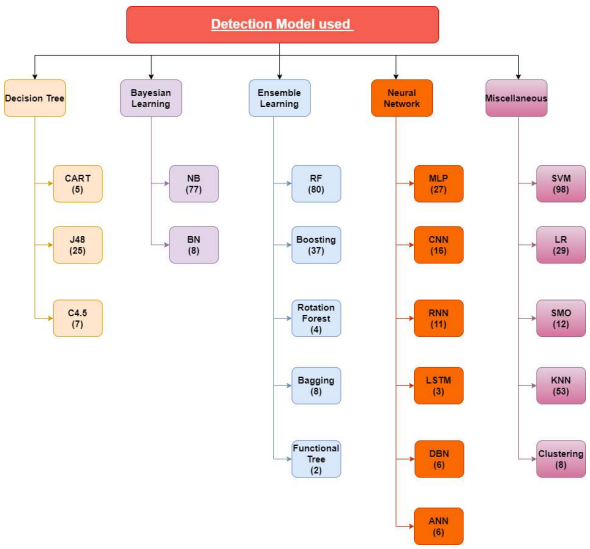


Fig. 4: Statistics depicting the usage of various detection models

Image recognition, etc. Hence, as seen in Table 6, researchers have explored a new area of Deep Neural networks and models like CNN, ANN, DAE, etc., for the past few years.

Figure 4 gives us a closer understanding of Table 6. We classify the majorly used detection models from Table 6 into five categories: Decision Trees, Bayesian Learning, Ensemble Learning, Neural Networks, and Miscellaneous to analyze the usage of different classifiers. For instance, Figure 4 highlights that five, 25, and seven papers out of the 200 related work reviewed in our work involved the usage of CART, J48, and C4.5 Decision trees, respectively. However, note that 46.5% of the total work utilized the DT classifier without mentioning the particular type. With the help of the information highlighted by the figure, we can conclude that the five most used detection models include SVM [64]-[66], DT [110]-[115], RF [124]-[131] from Ensemble learning, NB [165]-[167] from Bayesian Learning and KNN [73]-[75] with 98 (49%), 93 (46.5%), 80 (40%), 77 (38.5%) and 53 (26.5%) corresponding number (percentage) of studies that involved the classifiers mentioned respectively. It is worth noting that due to the growing popularity of deep learning, researchers have extensively started using neural network classifiers in their work. Neural networks [72]-[76] have been involved in more than 35% of the studies mentioned in our review. Apart from the ML/DL

Table 6: Model used

Related works	Machine Learning (ML)	Deep Learning (DL)	Others
Zhang et al. [19]			Permission behavioural pattern monitoring model
Li et al. [20]	SVM, Rotation Forest, PART, FT, Random Committee, RF.		
Sahin et al. [21]	NB, SMO, RF, C4.5 DT, LR, KNN	MLP	
Talha et al. [22]			Detection algorithm based upon permission score.
Varma et al. [23]	NB, J48 DT, RF	MLP, Multi-class classifier	
Mahindru et al. [24]	NB, J48 DT, RF, Simple Logistic (SL), kstar		
Dogru et al. [25]			Detection algorithm based upon permission score
Rathore et al. [26]			
Shang et al. [27]	Improved NB classification algorithm		
Tchakounte et al. [28]			Algorithm based upon permission patterns
Ju et al. [16]			Manual pattern recognition to existing malware permissions patterns
Ilham et al. [29]	RF, SVM, J48 DT		
Sahin et al. [30]	NB, KNN		
Angelo et al. [31]	J48 DT, NB	MLP	
Xiong et al. [32]			Developed a classifier name *ENCLAMALD* based upon permission patterns
Lu at el. [33]	RF, DT, Apriori algorithm		
Kavitha et al. [34]			
E. Amer [35]	RF, AdaBoost, SVM, DT	MLP	
Chakravarty et al. [36]	J48 DT, Random Committee, SMO, Randomizable filtered classifiers	MLP	
Pondugula et al. [37]		DNN	
Sahal et al. [38]	J48 DT, NB, SVM, KNN		
Tuan Mat et al. [39]	NB		
Wang et al. [40]	SVM, DT, RF	DCN	
Park et al. [41]	RF		
Liang et al. [42]			Developed their own rule-based algorithm
Enck et al. [14]			Analysis of the newly produced Android OS in 2009
Enck et al. [17]			Defined a new Application working upon its own Syntax Language
Wang et al. [43]			Algorithm based upon calculating the risk score of each app
Peng et al. [44]	NB and its modifications		
Pandita et al. [45]			NLP techniques to read application descriptions
Samra et al. [46]	k-means		
Yerima et al. [47]	Bayesian Classifier		
Aung et al. [48]	K-means clustering, J48 DT, RF, CART		
Yerima et al. [49]	Bayesian Classifier		
Sanz et al. [50]	SL, NB, Bayesnet, SMO, IBK, J48 DT, RandomTree		
Moonsamy et al. [51]	AHC		

Backes et al. [52]			Algorithm to assess and rewrite application policies
Wu et al. [53]	NB, KNN		
Kato et al. [54]	Stacking Ensemble Learning (SEL), RF		
Arora et al. [55]			Algorithm to form graphs based upon permission patterns and frequency
Alsoghyer et al. [56]	RF, J48 DT, SMO, NB		
Saleem et al. [57]			
Ghasempour et al. [58]	DT, SVM		
Shrivastava et al. [59]			Algorithm to calculate the risk score of an application using the feature's frequency
Upadhayay et al. [60]	RF, SVM , NB		
Lee et al. [61]	KNN, SVM, AdaBoost, Extra Tree, RF		
Surendran et al. [62]	Tree augmented NB		
A. T. Kabakus [63]		CNN	
Wang et al. [64]	SVM, GBDT, XGBM, LightGBM, CatBoost, LR		
Akbar et al. [65]	SVM, Rotation Forest, RF, NB		
Zhu et al. [66]	SVM, An ensemble taking DT as a base and involving PCA to form Rotation Forest		
Wang et al. [67]		Graph convolutional network	
N. McLauglin [68]		CNN	
Wang et al. [69]	SVM, RF, DT		
Grace et al. [70]			Risk level based detection of own algorithm
Liu et al. [71]	Multi-layered gradient boosting DT algorithm, SVM, DT, RF, XGBM		
Bayazit et al. [72]		RNN-based LSTM, BiLSTM, GRU	
Lee et al. [73]	J48 DT, RF, Decision Table, NB, SMO, SVM, LR, IBK, KNN, AdaBoost	MLP	
Zhu et al. [74]	SVM, KNN	Merged Sparse Auto-Encoder (MSAE), Stacked Denoising Auto-encoders (SDAE).	
Almahmoud et al. [75]	SVM, KNN, NB, RF, DT	RNN	
Feng et al. [76]		CNN	
Kandukuru et al. [77]	DT		
Arora et al. [78]	Apriori and FP-Growth algorithms		
Ding et al. [79]	KNN, multinomial NB, DT, RF, SVC, NuSVC, LinearSVC, LR, GBDT, XGBM	Combination of a residual network (ResNet) and LSTM	
Sahin et al. [80]	KNN, NB, SVM, DT, Bagging-DT algorithms		Linear regression-based and bagging method ensemble algorithms
Idrees et al. [81]			Algorithm based on the identification of distinct usage patterns of permissions and intents
Khariwal et al. [82]	RF, SVM, NB		
Idrees et al. [83]	NB, Kstar, Prism		
Zhu et al. [15]		CNN based Multi-Head Squeeze-and-Excitation Residual block	
Bai et al. [84]	CatBoost		
Taheri et al. [85]	Modified versions of KNN		
Alazab et al. [86]	RF, J48, RT, KNN and NB		

Mathur et al. [87]	KNN, SVM, LR, RF, ET, XGBM, AdaBoost, BG		
Imtiaz et al. [88]	DT, SMO, NB	MLP, ANN	
Liu et al. [89]	AdaBoost, Bagging, ExtraTrees, GBM, RF, Voting		Subspace based Restricted Boltzmann Machines (SBRBM)
Chen et al. [90]	LightGBM SVM, GBM, AdaBoost, ExtraTreesClassifier, RF, KNN, LR	Neural Networks	
Guan et al. [91]	RF, KNN, NB, SVM		Class Imbalanced Learning (CIL), Synthetic Minority Over-Sampling Technique (SMOTE), Random under-sampling (RUS), Balance Cascade
Mohamed et al. [92]	KNN, NB, SVM, DT		
Varma et al. [93]	RF, SVM, KNN, DT, Nearest Centroid (NC)		Bat Optimization Algorithm for Wrapper-based Feature Selection (BOAWFS)
Gyunka et al. [94]	NB, SL, RF, Partial DT, KNN, SVM		
Taha et al. [95]	LightGBM, SVM, KNN, LR, NB, DT.		
Peng et al. [96]		MLP, CNN	
Ashwini et al. [97]	Ridge Classifier, XGBM, RF, SVC		
Jiang et al. [98]	J48, KNN, NB, SVM		
Wang et al. [99]	LR, DT, SVM	DBN	
Rana et al. [100]	DT, RF, Extremely randomized trees (ERT), GB Tree, SVM, LR		
Lu et al. [101]		DBN, GRU	
Millar et al. [102]		CNN	
Barrera et al. [103]		SOM	
Shabtai et al. [104]	DT, NB, BN, BBN, BDT, PART, RF		
Felt et al. [105]			Own approach Stowaway to compare permissions required to invoke API methods and actually requested permissions
Erickson et al. [106]			
Sarma et al. [107]	SVM		
Frank et al. [108]			
Jhu et al. [109]	Naive Bayes with Multinomial Event Model		
Peiravian et al. [110]	SVM, DT, Bagging predictor		
Sanz et al. [111]	KNN, DT, Bayesnet, SVM		
Feldman et al. [112]	NB, SVM, KNN, C4.5 DT		
Pehlivan et al. [113]	Bayesian Classification, CART, DT, RF, SMO		
Rahman et al. [114]	NB, KNN, DT, RF, Decision Forest		
Rovelli et al. [115]	C4.5 DT, kstar Lazy Learning, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), NB, AdaBoost		
Arp et al. [116]	SVM		
Yerima et al. [117]	RF, NB, DT, random trees, SL		
Kang et al. [118]	NB		
Zhao et al. [119]	SVM, KNN, J48, NB		
Qiao et al. [120]	SVM, RF	ANN	
Chen et al. [121]	SVM, C4.5 DT, NB, KNN, Bagging predictor	MLP	
Demertzis et al. [122]		Evolving Spiking Neural Network (eSNN), MLP, Radial Basis functions, GMDH PNN, eSNN	Extreme learning machine (ELM)

Verma et al. [123]	SVM, Minimal complexity machine		
Wang et al. [124]	NB, LR, DT, Bayesnet, RF, SVM		
Tangil et al. [125]	Extra Trees, SVM, RF, XGBoost		
Wang et al. [126]	KNN, J48 DT, RF		
Li et al. [127]	NB, DT, LR, SVM, RF		
Bhattacharya et al. [128]	Bayesnet, DT, IBK, J48 DT, JRip, Kstar, LR, RF, SMO, NB, etc.	MLP	
Xie et al. [129]	DT, RF, SVM, KNN, NB		
Xie et al. [130]	SVM, LR, RF	CNN	
Ren et al. [131]	SVM, KNN, AdaBoost , RF, GBM		
Tao et al. [132]	RF, SVM, DT		
Namrud et al. [133]	SVM, K means, SOM		
Alswaina et al. [134]	Extremely Randomized Trees (ET), SVM, ID3 DT, RF, BC, KNN	NN	
Qiu et al. [135]	SVM, DT	DNN	
Zhu et al. [136]	SVM, KNN, DT, LR, RF, ET, Adaboost, GBDT	MLP	
Feng et al. [137]	LR, NB-SVM	GRU , CNN	
Aonzo et al. [138]			Proposed their own linear and non linear classifiers
Urooj et al. [139]	AdaBoost, SVM, RF, KNN, NB, RBF, DT, SVM		
Wang et al. [140]			Evaluated their own text semantic-based approach
Wang et al. [141]	SVM, KNN, RF		
Zhang et al. [142]			Own algorithm to detect bugs and Implicitly Malicious Behavior
Kesswani et al. [143]	NB		
Ibrahim et al. [144]	Stochastic Gradient Descent (SGD), DT, RF, SVC, KNN, XGBoost, GaussianNB	GRU	
Arshad et al. [145]	SVM, RF, NB, DT		
Yuan et al. [146]	NB, C4.5 DT, LR, SVM	DBN, CNN, MLP	
Zhou et al. [147]	RF, KNN, SVM	GRU	
Cilleruelo et al. [148]	SVM, Stochastic gradient descent(SGD), RF, XGBM		
Firdaus et al. [149]	NB, FT, J48 DT, RF	MLP	
Wang et al. [150]	KNN, SVM, NB, J48 DT		
Singh et al. [151]	SVM(Linear)		
Rafiq et al. [152]	SVM, LR, DT, XGBM, RF, KNN		
Mahdavifar et al. [153]	RF, SVM, KNN	Pseudo-Label Stacked Auto-Encoder (PLSAE), Pseudo-Label Deep Neural Network (PLDNN), Label Programming (LP)	
Seraj et al. [154]	RF, Gaussian NB, SVC, KNN	MLP	
Mahindru et al. [155]	DT, LR, NB, SVM with linear kernel, polynomial kernel, and RBF kernel	SOM, NN	
Sahin et al. [21]	KNN, NB, RF, C4.5 DT, LR	SMO, MLP	
Anupama et al. [156]	LR, CART, RF, SVM	DNN, CNN	
Chen et al. [157]	RF	CNN	
Mahindru et al. [158]	LSSVM with distinct kernel functions (polynomial, RBF and linear), LR, NB, DT	NN	
Tchakounté et al. [159]			Algorithm based on fuzzy hash processed static features and computed similarity score

Nissim et al. [160]	SVM		Active Learning methods
Peynirci et al. [161]	ID3, CART, J48 DT, RF, NB, RBF, SLogReg, Prism, MODLEM, OneR, NNGE, SVM	MLP	
Nauman et al. [162]	Bayesian Statistical Inference	Neural networks, DAE , DBN, CNN , LSTM	
Bhattacharya et al. [163]	Bayesnet, NB, Decision Table, Random Tree, RF, J48 DT	SMO, MLP	Quick Reduct algorithm
Bao et al.[164]	NB, Multinomial classification		Algorithms based upon predicting permissions from API and text classification
Medrano et al. [165]	-	-	
Mat et al. [165]	NB		
Shatnawi et al. [166]	LR, SVM, KNN, NB, GaussianNB		
Smmarwar et al. [167]	OEL, fine tree, NB, cubic SVM, Weighted KNN, Bagging, Adaboost, GentleBoost		
Arif et al. [168]			Information Gain
Manzanares et al. [169]			Extracted applications to form a comprehensive dataset
Bhat et al. [170]	DT, LR, SVM, NB, RF		
Elayan et al.[171]	SVM, KNN, NB, DT, RF	GRU	
Syrris et al. [172]	NB, RF, SVM	ANN	
Idrees et al. [173]	NB, DT, Decision Table, RF	SMO, MLP	
Rehman et al. [174]	SVM, KNN, DT, Linear Discriminant Analysis		
Martin et al. [175]	AdaBoost, Bagging (with RF estimators), ExtraTrees, GBM, RF, Voting classifier combining an RF, KNN, DT		
Navarro et al. [176]	RF, CART, Bagging		
Milosevic et al. [177]	C4.5 DT, NB, SVM with SMO, RF, JRIP, LR, AdaBoost, Farthest First, Simple K-means and Expectation maximization (EM) algorithms		
Alzaylaee et al. [178]	SVM Linear and RBF, NB, SL, DT, PART, J48 DT	MLP	
Cai et al. [179]	SVM, LR, RF, KNN, Gaussian NB, multinomial NB, LDA, QDA, RF	MLP	
Badhani et al. [180]	DT, ELM, LR, RIPPER, SVM-linear		
Hijawi et al. [181]	RF, J48 DT, NB, SVM, KNN	MLP	
Sheen et al. [182]	Adaboost, IBK and J48 DT, Decision Stump, Random tree, NB, SVM		
Nisha et al. [183]	KNN, SVM, RF, GBT	CNN	
Song et al. [184]			Dangerous permissions matching model to compute a threat degree threshold
Zhang et al. [185]			Algorithm to identify explicit and implicit permission use points to further create permission use graphs to analyze the permissions behavioral pattern
Yang et al. [186]			Algorithm to construct State Transition Graphs (STG) from permissions to analyze permission-related behavior only under the combinations of the relevant permissions
Thiyagarajan et al. [187]	DT		

Qaisar et al. [188]	SVM, DT		Case-based learning approach to store and learn permissions like features
Appice et al. [189]	RF, SVM, stacking, ensemble	Deep-Net	Algorithm involving clustering-based sampling with the specific multi-view learning approach
Zhu et al. [190]	RF, SVM		
A. Altaher [191]		Neuro-fuzzy inference system	
Su et al. [192]		DBN, SVM	
Mahindru et al. [193]	SVM, NB, RF, LR, K-Means, FF, FC, DB, Bayesnet, Adaboost, DT, KNN, J48 DT	MLP, DNN, SOM, Y-J48, Y-SMO, Y-MLP, best training ensemble approach (BTE), majority voting ensemble approach (MVE) and nonlinear ensemble decision tree forest approach (NDTF)	
Dehkordy et al. [194]	SVM, KNN, ID3 DT		
Nguyen et al. [195]	SVM	DNN	
Taheri et al. [196]	SVM	Robust NN, C4N CNN	
Mahesh et al. [197]		CNN, ANN , DBN, Method level behavioral semantic approach (MLBS) DE-CNN	CNN-Adaptive Red Fox Optimization (CNN-ARFO)
Firdaus et al. [198]	RF, Bayesnet, SVM	MLP, Voted perceptron (VP), Radial basis function network (RBFN) ANN	
Shrivastava et al. [199]			Frequency-based rules defined over permissions and intents to predict malicious risk level of an application
Varsha et al. [200]	RF, Rotation forest, SVM		
M. Deypir [201]			Entropy-based method to calculate information gain score for a permission and correspondingly risk score for an application
Mahindru et al. [202]	K-mean, SVM(linear, poly, RBF), DT, LR, NB	NN, SOM, FF, FC and DB	
Keyvanpour et al. [203]	KNN, SL, NB, J48 DT, RT, RF		
Razak et al. [204]	RF, J48 DT, KNN, AdaBoost	MLP	
Xie et al. [205]	SVM		
Mahindru et al. [206]	LOGR, SVM, NB, DT	Radial Basis Function Neural Network RBFM-KCM, RBFN-FCM, RBFN-RAN, ANN, with three different ensemble methods (i.e., BTE, MVE, and NDTF	
Alecakir et al. [18]		RNN GRU, MLP	
Ali et al. [207]	SVM (Linear, RBF, Poly), LR		
Sun et al. [208]	HMM, GD, SVM (Linear, RBF, Poly, KCD)		
AlJarrah et al. [209]	RF, SVM, LR, NB, KNN, DT		
Gharib et al. [210]	SVM, RF, NB, AdaBoost	DNN, DAE	
Sun et al. [211]	RF, DT, SVM		Positive and Unlabeled learning

classifiers mentioned in Figure 4, researchers have developed their algorithm, too, instead of using the traditional classifiers. For instance, in [55], Arora et al. proposed their algorithm to form graphs based on permission patterns and frequency as weights. **Hence, based on the results of the empirical evidence, we answer the fourth Research Question that the most commonly used ML models are Support Vector Machines, Decision Trees, and Random Forest due to their convenient and simple working but at the same time, due to the limitations of traditional supervised learning methods, one can assume that the popularity of neural network models will surely rise.**

6 Datasets used

Researchers and practitioners have adopted numerous malicious datasets in the past decade to carry out their research on Android malware. The datasets can be broadly divided into two types based on their data source: *in-the-lab* datasets and *in-the-wild* datasets. Traditional datasets such as Drebin[5] and Genome[6] are considered as baselines in the field of Android malware detection and fall under the category of *in-the-lab* datasets. On the other hand, datasets such as Virustotal[7], Virusshare[8], Contagio[9], AMD [215], McAfee[10] and Androzoo[11] are called *in-the-wild* datasets due to their continuously updating nature, and that is the reason why it's obvious that the performance of *in-the-wild* datasets is found to be more credible than that of *in-the-lab* datasets and in addition to the datasets mentioned above some researchers also utilized kaggle[12] and Anzhi[13] type datastores too. Apart from the datasets mentioned above, the datasets provided by the Canadian Institute of Cybersecurity[14] such as the CICInvesAndMal2019, CICAndMal2017, and CIC2020 have also been utilized by the researchers.

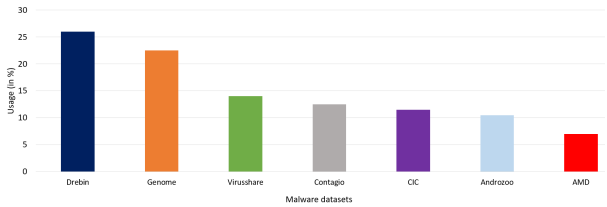


Fig. 5: Statistics depicting the most commonly used malware datasets

In response to RQ5, we present Table 7 as it summarizes the malware datasets used by researchers and practitioners to form their analysis/detection model. As seen from the table, most researchers have used traditional malware

[5]http://www.sec.cs.tu-bs.de/ danarp/Drebin/
[6]http://www.malgenomeproject.org/
[7]https://www.virustotal.com
[8]https://virusshare.com/
[9]http://contagiominidump.blogspot.hk/
[10]https://home.mcafee.com/Default.aspx?rfhs=1
[11]https://androzoo.uni.lu/
[12]https://www.kaggle.com
[13]http://www.anzhi.com/
[14]https://www.unb.ca/cic/datasets/index.html

Table 7: Dataset used

Related works	**Malware datasets used**
Zhang et al. [19]	Genome
Li et al. [20]	Anzhi Store
Sahin et al. [21]	[216]
Talha et al. [22]	Contagio, Drebin, Genome
Varma et al. [23]	[217]
Mahindru et al. [24]	[218], [219], Genome, and AndroMalShare[15]
Dogru et al. [25]	Drebin
Rathore et al. [26]	Drebin
Shang et al. [27]	-
Tchakounte et al. [28]	CIC-AndMal2017
Ju et al. [16]	-
Ilham et al. [29]	AMD
Sahin et al. [30]	Kaggle
Angelo et al. [31]	Unisa malware dataset (UMD)[16]
Xiong et al. [32]	[217]
Lu at el. [33]	Baidu application market[17], Genome, Virusshare
Kavitha et al. [34]	-
E. Amer [35]	Drebin, Genome
Chakravarty et al. [36]	AMD
Pondugula et al. [37]	For training data from IEEE DataPort, and sites like droidbench
Sahal et al. [38]	Contagio, Genome
Tuan Mat et al. [39]	Androzoo and Drebin
Wang et al. [40]	Virusshare, Genome
Park et al. [41]	AndroZoo
Liang et al. [42]	Genome
Enck et al. [14]	-
Enck et al. [17]	Android Market
Wang et al. [43]	Genome
Peng et al. [44]	Genome
Pandita et al. [45]	-
Samra et al. [46]	-
Yerima et al. [47]	Genome
Aung et al. [48]	-
Yerima et al. [49]	Genome
Sanz et al. [50]	VirusTotal
Moonsamy et al. [51]	Genome
Backes et al. [52]	-
Wu et al. [53]	Contagio
Kato et al. [54]	Drebin, Androzoo, VirusShare
Arora et al. [55]	Genome, Drebin, Koodous
Alsoghyer et al. [56]	HelDroid [220], RansomProber project[212], VirusTotal, Koodous
Saleem et al. [57]	Drebin
Ghasempour et al. [58]	Androzoo
Shrivastava et al. [59]	Drebin
Upadhayay et al. [60]	Genome, Drebin, Koodous
Lee et al. [61]	CICAndMal 2017
Surendran et al. [62]	Drebin, AMD, AndroZoo, external repositories
A. T. Kabakus [63]	Drebin ,KuafuDet[18] ,AndroZoo, VirusShare
Wang et al. [64]	Huawei application market[19], Xiaomi application market[20], 360 application market[21], Wandoujia application market[22]
Akbar et al. [65]	VirusShare
Zhu et al. [66]	VirusShare
Wang et al. [67]	Andro_dumpsys

Related works	Malware datasets used
N. McLauglin [68]	McAfee, vendor's internal dataset
Wang et al. [69]	Mal_com1, Mal_com2 and Mal_Zhou [220]
Grace et al. [70]	Github
Liu et al. [71]	VirusShare
Bayazit et al. [72]	CICInvesAndMal2019
Lee et al. [73]	Andro-AutoPsy Dataset [221]
Zhu et al. [74]	MUDFLOW [222], VirusShare
Almahmoud et al. [75]	CIC-AndMal2017, CIC-InvesAndMal2019, CIC-MalDroid2020
Feng et al. [76]	CICAndMal2017
Kandu et al.[77]	Genome
Arora et al. [78]	Genome
Ding et al. [79]	CICInvesAndMal2019
Sahin et al. [80]	M0Droid[223], AMD, Kaggle, [224]
Idrees et al. [81]	Contagio, Drebin, Genome, Virus Total, theZoo, MalShare, VirusShare
Khariwal et al.[82]	Genome, Drebin, Koodous
Idrees et al. [83]	Contagio, VirusTotal, appsapk, Android-mob
Zhu et al. [15]	VirusShare
Bai et al. [84]	Drebin
Taheri et al. [85]	Drebin, Contagio, Genome
Alazab et al. [86]	AndroZoo, Contagio, MalShare, VirusShare, VirusTotal
Mathur et al. [87]	Androzoo, AMD
Imtiaz et al. [88]	CICInves AndMal2019
Liu et al. [89]	OmniDroid, CIC2019, CIC2020
Chen et al. [90]	VirusShare
Guan et al. [91]	VirusShare
Mohamed et al. [92]	Genome, Maldroid
Varma et al. [93]	CICInvesAnd Mal2019
Gyunka et al.[94]	Genome, Contagio
Taha et al. [95]	Drebin
Peng et al. [96]	CICMalDroid 2020, CIC-InvesAndMal 2019, Drebin
Ashwini et al. [97]	Drebin
Jiang et al. [98]	Genome, Andro MalShare
Wang et al. [99]	Information Security Lab of Peking University
Rana et al. [100]	Drebin
Lu et al. [101]	VirusShare, Genome, Contagio
Millar et al. [102]	Genome, Drebin, vendor's internal repository, AMD
Barrera et al. [103]	No malware
Shabtai et al. [104]	No malware
Felt et al. [105]	No malware
Erickson et al. [106]	No malware
Sarma et al. [107]	Contagio
Frank et al. [108]	No malware
Jhu et al. [109]	Steamy Window[23]
Peiravian et al. [110]	[217]
Sanz et al. [111]	VirusTotal
Feldman et al. [112]	Contagio
Pehlivan et al. [113]	COMODO Security Solutions, Inc., a private security company[24]
Rahman et al. [114]	Kharon Malware Dataset[25], CICInvesAndMal2019, MalwareBazaar[26], Github, other platforms

Related works	Malware datasets used
Rovelli et al. [115]	Genome, Contagio
Arp et al. [116]	Genome,FakeInstaller, GoldDream[27], GingerMaster[28], DroidKungFu[29]
Yerima et al. [117]	McAfee
Kang et al. [118]	VirusShare, Contagio, Malware.lu
Zhao et al. [119]	Drebin
Qiao et al. [120]	Genome
Chen et al. [121]	360 APKs[30], MobiSec Lab Website[31], [217]
Demertzis et al. [122]	Magnum- Research[32]
Verma et al. [123]	Contagio, malware forums , security blogs, Genome
Wang et al. [124]	VirusTotal
Tang et al. [125]	Genome, Drebin
Wang et al. [126]	Drebin, Genome
Li et al. [127]	Drebin
Bhattacharya et al. [128]	Contagio
Xie et al. [129]	Genome, VirusShare, Drebin
Xie et al. [130]	Genome, VirusShare, Drebin, antivirus companies
Ren et al. [131]	Anzhi, AndroTotal, Drebin
Tao et al. [132]	VirusShare, Contagio
Namrud et al. [133]	AndroZoo
Alswaina et al. [134]	-
Qiu et al. [135]	-
Zhu et al. [136]	ViruShare
Feng et al. [137]	No Malware
Aonzo et al. [138]	AndroZoo
Urooj et al. [139]	MalDroid [225], DefenseDroid[33] and a small own generated dataset
Wang et al. [140]	No malware
Wang et al. [141]	FakeInst, Opfake, FakeInstaller, Droid-KungFu, GinMaster, Plankton
Zhang et al. [142]	No malware
Kesswani et al. [143]	No malware
Ibrahim et al. [144]	CICMalDroid 2020
Arshad et al. [145]	Drebin
Yuan et al. [146]	Genome, Contagio
Zhou et al. [147]	Genome
Cilleruelo et al. [148]	Malware selected on the basis of lifespan criteria from Google Play Store
Firdaus et al. [149]	Drebin
Wang et al. [150]	[217]
Singh et al. [151]	Drebin, polymorphic and metamorphic malware dataset
Rafiq et al. [152]	Drebin, AMD, Androzoo
Mahdavifar et al. [153]	VirusTotal, Contagio, AMD, CICMal-Droid202
Seraj et al. [154]	Applications classified as harmful via VirusTotal and downloaded from Google Play Store
Mahindru et al. [155]	Didn't mention properly
Sahin et al. [21]	VirusShare
Anupama et al. [156]	Drebin
Chen et al. [157]	AndroZoo
Mahindru et al. [158]	Genome, [218]
Tchakounté et al. [159]	AMD
Nissim et al. [160]	Contagio, Genome, third party application stores based in Russia, China, and Europe
Peynirci et al. [161]	Genome, Drebin, AndroZoo
Nauman et al. [162]	Drebin, VirusShare
Bhattacharya et al. [163]	Wang's repository dataset[34], Contagio, KEEL/UCI repository
Bao et al.[164]	No malware
Medrano et al. [165]	No malware
Mat et al. [165]	AndroZoo, Drebin
Shatnawi et al. [166]	CIC InvesAndMal2019

Related works	Malware datasets used
Smmarwar et al. [167]	CICInvesAndMal2019
Arif et al. [168]	AndroZoo, Drebin
Manzanares et al. [169]	Drebin, AMD, VirusTotal, VirusShare
Bhat et al. [170]	Virustotal, VirusShare, Drebin
Elayan et al.[171]	CICAndMal2017
Syrris et al. [172]	Drebin
Idrees et al. [173]	Contagiodump, Genome, Virus Total, theZoo, MalShare, VirusShare
Rehman et al. [174]	M0DROID
Martin et al. [175]	Koodous, AndroZoo
Navarro et al. [176]	AndroZoo, [217] , VirusShare, AndroMalShare
Milosevic et al. [177]	M0Droid
Alzaylaee et al. [178]	McAfee
Cai et al. [179]	AMD, Drebin
Badhani et al. [180]	Andro- DumpSys, AndroZoo, Contagio , AndroMalShare , AMD, VirusTotal
Hijawi et al. [181]	[218]
Sheen et al. [182]	Genome
Nisha et al. [183]	AndroZoo, VirusShare, Andro-MalShare, PRAGuard[226]
Song et al. [184]	Not mentioned
Zhang et al. [185]	Genome
Yang et al. [186]	No malware
Thiyagarajan et al. [187]	AndroZoo
Qaisar et al. [188]	Android PRAGuard, Drebin, Open-source apps, Kharon, Androzoo, CICAAGM
Appice et al. [189]	Alternative Chinese and Russian Markets, Android websites, malware forums, security blogs, Genome
Zhu et al. [190]	VirusShare
A. Altaher [191]	Genome
Su et al. [192]	Drebin, Genome, Contagio
Mahindru et al. [193]	Genome , AndroMalShare, [218]
Dehkordy et al. [194]	Drebin, AMD
Nguyen et al. [195]	AMD, Drebin
Taheri et al. [196]	Drebin, Contagio, Genome
Mahesh et al. [197]	Private companies
Firdaus et al. [198]	Drebin, Genome
Shrivastava et al. [199]	Third-party applications
Varsha et al. [200]	Drebin
M. Deypir [201]	[217], Security expert and World Wide Web
Mahindru et al. [202]	Genome, AndroMalShare, [218]
Keyvanpour et al. [203]	Drebin
Razak et al. [204]	Drebin
Xie et al. [205]	Anzhi
Mahindru et al. [206]	[218], [219], Sanddroid
Alecakir et al. [18]	No malware
Ali et al. [207]	No malware
Sun et al. [208]	Contagio
AlJarrah et al. [209]	CICMalDroid2020
Gharib et al. [210]	R-PackDroid[35], HellDroid[220], Contagio, Koodous
Sun et al. [211]	Drebin

datasets in their studies, such as Genome and Drebin, whereas some have shifted their focus to other alternatives from the *in-the-wild* dataset category, namely Contagio. Figure 5 gives us a better understanding of Table 7 by plotting the top seven malware datasets in terms of usage in %, i.e., the percentage of studies that have utilized the corresponding malware dataset from the total lot of research papers used in our review. The statistics depict the preference of researchers towards the traditional datasets as Drebin [95]-[97] stands at the top with 26% usage, followed by Genome [103]-[104] at 22.5%. *In-the-wild* datasets such as Virusshare [90]-[91], Contagio [94] and Androzoo [86]-[87] stand at third, fourth and sixth rank with 14%, 12.5% and 10.5% share respectively. As discussed above, the popularity of the datasets provided by the Canadian Institute of Cybersecurity [166]-[167] due to the concise pattern of information in datasets, and they take up almost 11.5% of the comprehensive studies taken up in this review. **Hence, based on the results presented above, we answer the fifth Research Question that the *In-the-lab* datasets such as Drebin and Genome occupy the largest proportion of the most preferred malware dataset by researchers.**

7 Challenges and Limitations

We answer our sixth Research Question in this and the subsequent section, wherein we describe a few limitations of proposed models mainly associated with permissions as a feature in the field of Android security followed by proposing some future research directions.

7.1 Drawbacks of using permissions as a feature

Static techniques are generally inexpensive in terms of complexity compared to dynamic approaches, as static features can be more easily extracted than dynamic ones. However, static methods have a few disadvantages, and consequently, models built over Android permissions are generally found incapable of recognizing the stealthier behavior of code obfuscation and dynamic code loading. As a result, some malicious apps may incorporate stealthy behavior and

evade permissions-based detection. Additionally, the inability of many detection models to track behaviors that do not cause permission checks is another limitation of using permissions as a feature [19], [185].

7.2 Faulty Research dataset

The training and validation dataset is the foundation for any Android malware analysis/detection model. Faulty research datasets, biased to one class, i.e., either benign or malware, outdated, or insufficient in size or distribution, can produce inaccurate results in terms of meaning and poor detection accuracy. It has been noticed that app developers often make careless spelling mistakes. Applications are found with no permission requests or faulty formatting of manifest files. These reasons surely cause hindrances to research. Results with insufficient data are less efficient and accurate because of the lack of training data. Even if the detector performs well in the experimental tests, it might still prove to be inefficient in real life. Hence, building a good and informative dataset of Android applications that request the actual and essentially needed permissions or features is an important pre-requisite step for any model. One of the main reasons for accumulating faulty applications in one's dataset is the open-source nature of the Android operating system itself. Users are given the freedom to download applications not only from the mainstream market but also from informal third-party markets and unknown malicious websites, which might result in the mixing of benign and malware application sets.

7.3 Feature Extraction and Processing

Gathering the right application samples is important; however, knowing how and which features to extract is itself a very extensive research topic. As described in the previous section, despite the numerous advantages of using static features, static analysis and consequently permission-based techniques are also found to have some disadvantages At the same time, there are also many challenges to feature extraction with dynamic analysis. For example, how to ensure that all malicious behaviors can be triggered during dynamic analysis and which emulator to use to produce real-time usage patterns of a user are issues that need to be solved.
The summary and statistics presented in section V indicate that most of the research is still based on traditional analysis features, such as permissions, because of their strong training effect, vast size, and prominent relationship with class labels and behavior.

7.4 Application of Machine Learning

Machine learning methods are used in the field of Android malware detection to improve the effectiveness and efficiency of detection. However, because of the phenomena such as concept drift, the performance of traditionally trained classifiers has started to decline. Concept drift, which is a problem that arises with the continuous development and evolution of Android malware, has caused researchers and practitioners to shift their focus toward neural network-based detection methods. For example, Bayazit et al. [72] chose static analysis to build their various RNN-based classifiers. The features that they used were permissions and intents. In the end, they compared the detection results of their various RNN-based algorithms to determine that the Bidirectional Long Short-Term Memory (BiLSTM) model outperforms all. Moreover, neural networks can be applied to bridge the gap between the semantic representation of APKs and Android malware detection. Hence, we conclude that neural networks have potential in the field of Android malware detection using static analysis. The preference of researchers and practitioners is starting to tilt more towards deep learning classifiers while building their analysis/detection models because of the added advantages compared with the traditional machine learning classifiers. However, deep learning algorithms can be computationally expensive and may require many resources to be successfully trained. This can be seen as a major barrier, especially for people who want access to high-performance computing services but lack the necessary resources.

7.5 Unreasonable reliability over users

The majority of the proposed models in the field of Android malware detection lie under the category of *Off-Device* models, i.e., they rely upon the user to provide the application data and then return the malware report of the corresponding applications after performing their proposed approach which is, in fact, a tedious and long method. Instead, *On-Device* models are the need of the hour, capable of performing malware analysis/detection using the user's smartphone to generate instant malware reports.

7.6 Application Collusion

Application collusion is an emerging threat to Android-based devices that seem almost immune to permissions-based detection systems. In app collusion, two or more Android apps somehow collude to perform a malicious action that they cannot accomplish independently [55]. In this way, they perform malicious tasks without showing malware behavior.

7.7 Zero day malware

Zero-day malware is the kind of threat for which no patch exists yet and therefore seems immune to traditional signature-based analysis that identifies the dangerous permissions usage patterns. In simple words, permission-based malware detection strategies are based on knowledge of the vulnerability, exploit in question, or analyzing behavioral patterns of application classes, which obviously is not available for zero-day threats. As a result, certain methods for mitigating these threats are ineffective.

8 Future Directions

In this section, we provide some directions for future work in the Android malware analysis/detection field using Android permissions based on our investigation and review.

1. The Android OS has more than 150 pre-defined permissions. Moreover, as applications become more complex with time, this number is bound to grow only. Permissions are one of the key components present within any Android application's manifest file and have been widely used in the literature for Android malware detection. However, the usage pattern of permissions between normal and malware applications is generally quite similar. For instance, we have collected 77,000 normal apps and an equal number of malware apps from Androzoo. Furthermore, we have extracted permissions from the manifest files of corresponding applications. Table 8 shows that 13 of the top 20 permissions are common in normal and malware datasets. If all these permissions are considered features while building an analysis or detection model, poor detection accuracy results are inevitable. Table 9 helps us to further prove our point by summarizing the detection results considering all features, i.e., the total lot of 129 permissions, for detection without performing any feature selection/ ranking/ other technique. As seen from the Table, the highest detection accuracy achieved by using the BC classifier is merely 78.64%, which is too low in Android security for any malware detection model. Hence, priority should be given to feature selection or ranking to filter out only the most relevant and influential set of features in the future.
2. Combining multiple static analysis techniques or features can always be beneficial when it comes to comparison with using just a single feature set. Permissions are known to be manually designed and generally are coarse-grained in nature. Intents possess similar characteristics like permissions, whereas opcode-based methods are found to be capable of capturing contextual semantics of applications but they both fail to produce structural semantics. Functional call graphs do the needful by extracting structural semantics. Hence, we can say that using two or more types of static features in combination might further improve the experimental results and thus a model's detection accuracy.
3. Since stealthier malware may evade permissions-based static detection, several works in the literature aim to use dynamic features such as network traffic and system calls to detect such stealthier malware. However, analyzing dynamic features is complex and time-consuming; therefore, hybrid detection models serve the purpose as they combine both

Table 8: Top 20 most frequently requested permissions from both normal and malware datasets with their corresponding frequency.

PERMISSIONS	Normal Frequency	PERMISSIONS	Malware Frequency
INTERNET	55063	INTERNET	55684
ACCESS_NETWORK_STATE	52391	ACCESS_NETWORK_STATE	55252
WRITE_EXTERNAL_STORAGE	38934	WRITE_EXTERNAL_STORAGE	54759
WAKE_LOCK	32527	ACCESS_WIFI_STATE	53886
ACCESS_WIFI_STATE	28554	READ_PHONE_STATE	53586
RECEIVE	23875	READ_EXTERNAL_STORAGE	46646
READ_EXTERNAL_STORAGE	22516	WAKE_LOCK	44003
VIBRATE	20472	GET_TASKS	43399
ACCESS_FINE_LOCATION	16968	CHANGE_WIFI_STATE	43165
ACCESS_COARSE_LOCATION	16650	ACCESS_COARSE_LOCATION	42425
RECEIVE_BOOT_COMPLETED	16519	VIBRATE	42325
CAMERA	14993	MOUNT_UNMOUNT_ FILESYSTEMS	41324
READ_PHONE_STATE	14176	ACCESS_FINE_LOCATION	40720
C2D_MESSAGE	12342	WRITE_SETTINGS	39497
BIND_GET_INSTALL_REFERRER_ SERVICE	10593	SYSTEM_ALERT_WINDOW	38594
BILLING	9905	CAMERA	36115
FOREGROUND_SERVICE	9587	CHANGE_NETWORK_STATE	30874
GET_ACCOUNTS	7806	RECEIVE_BOOT_COMPLETED	29441
WRITE_SETTINGS	7258	READ_LOGS	29112
BLUETOOTH	5820	RECORD_AUDIO	27010

Table 9: Detection results considering all features for detection without performing any feature selection/ ranking/ other technique

FEATURES used	Detection accuracy using various machine learning and deep learning classifiers (in %)					
	DT	RF	ANN	BC	NB	LR
All Permissions (129)	74.64	74.64	69.55	78.64	69.60	69.60

static and dynamic elements. Because of static features like permissions in hybrid models, some malware samples can be easily detected. And dynamic features can help detect stealthier samples. Hence, hybrid models may be better for effective Android malware detection.

4. Moreover, as discussed in [143], the malware authors tend to request more permissions than their application requires for its functioning to cover up or hide their malicious intent. Due to the absence of any robust permission check system that provides an in-depth comparison of the permissions requested and required, malware authors exploit it as much as possible. Hence, more research is required on this to fulfill the absence of such a permission check system by the OS shortly to reduce the malware attacks at the initial level only.
5. Currently, repositories like Androzoo and CIC provide standardized application dataset samples and libraries with comprehensive background information and essential filters that solve the faulty dataset issue up to some extent.

9 Conclusion

This paper summarizes the state-of-the-art studies and comprehensively reviews the research trend in Android malware analysis/detection using permissions as a key feature. Concretely, this Comprehensive Literature Review (CLR) is conducted with the close observation and study of 200 research papers ranging from the advent of Android in 2009 to the current research scenario in 2023, covering almost 14 years of research history.

Based on the six research questions put forward by us to understand the relevance of permissions as a means to analyze/detect Android malware, this CLR investigates the purpose of research (analysis or analysis combined with detection), the technique used to select or utilize or rank the extracted features, most commonly used features alongside with permissions, the malware datasets utilized by the researchers/ practitioners and lastly, the limitations, challenges in the field of Android malware analysis/detection in order to propose some future research directions.

According to this CLR, we answered the six research questions mentioned above and conclude that 1) Majority of researchers and practitioners have chosen the path of building Android malware detection models instead of merely analysing the malware; 2) The most popular techniques to handle the extracted features are the ML/DL classifiers as their hyper parameters can be easily tuned to reduce, rank, select even the unlabelled or unbalanced feature set, followed by statistical test such as Chi-square, Mann-Whitney test and many more; 3) Researchers have preferred using API calls and intents the most in combination with permissions while forming their Android malware analysis/detection models; 4) Based on the results of empirical evidence, we conclude that the most commonly used ML models are SVM, DT and RF due to their convenient and simple working but at the same time, due to the limitations of traditional supervised learning methods, one can assume that the popularity of neural network models will surely rise; 5) *In-the-lab* datasets such as Drebin and Genome occupy the largest proportion of the most preferred malware dataset by researchers.

Compliance with ethical standards

Conflict of interest - The authors declare that they have no conflict of interest.
Ethical approval - This article does not contain any studies with human participants or animals performed by any of the authors.
Research Data Policy and Data Availability - The authors declare that the data supporting the findings of this study are available within the paper.

References

[1] Lubuva, H., Huang, Q., Msonde, G.C.: A review of static malware detection for android apps permission based on deep learning. International Journal of Computer Networks and Applications **6**(5), 80–91 (2019)

[2] Alqahtani, E.J., Zagrouba, R., Almuhaideb, A.: A survey on android malware detection techniques using machine learning algorithms. In: 2019 Sixth International Conference on Software Defined Systems (SDS), pp. 110–117 (2019). IEEE

[3] Naway, A., Li, Y.: A review on the use of deep learning in android malware detection. arXiv preprint arXiv:1812.10360 (2018)

[4] Hahn, S., Protsenko, M., Müller, T.: Comparative evaluation of machine learning-based malware detection on android. Sicherheit 2016-Sicherheit, Schutz und Zuverlässigkeit (2016)

[5] Kouliaridis, V., Barmpatsalou, K., Kambourakis, G., Chen, S.: A survey on mobile malware detection techniques. IEICE Transactions on Information and Systems **103**(2), 204–211 (2020)

[6] Qamar, A., Karim, A., Chang, V.: Mobile malware attacks: Review, taxonomy & future directions. Future Generation Computer Systems **97**, 887–909 (2019)

[7] Hamed, Y.S.I., AbdulKader, S.N.A., Mostafa, M.-S.M.: Mobile malware detection: A survey. International Journal of Computer Science and Information Security (IJCSIS) **17**(1) (2019)

[8] Yan, P., Yan, Z.: A survey on dynamic mobile malware detection. Software Quality Journal **26**(3), 891–919 (2018)

[9] Odusami, M., Abayomi-Alli, O., Misra, S., Shobayo, O., Damasevicius, R., Maskeliunas, R.: Android malware detection: A survey. In: Applied Informatics: First International Conference, ICAI 2018, Bogotá,

Colombia, November 1-3, 2018, Proceedings 1, pp. 255–266 (2018). Springer

[10] BalaGanesh, D., Chakrabarti, A., Midhunchakkaravarthy, D.: Smart devices threats, vulnerabilities and malware detection approaches: a survey. European Journal of Engineering and Technology Research **3**(2), 7–12 (2018)

[11] Gyamfi, N.K., Owusu, E.: Survey of mobile malware analysis, detection techniques and tool. In: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1101–1107 (2018). IEEE

[12] Liu, K., Xu, S., Xu, G., Zhang, M., Sun, D., Liu, H.: A review of android malware detection approaches based on machine learning. IEEE Access **8**, 124579–124607 (2020)

[13] Pan, Y., Ge, X., Fang, C., Fan, Y.: A systematic literature review of android malware detection using static analysis. IEEE Access **8**, 116363–116379 (2020)

[14] Enck, W., Ongtang, M., McDaniel, P.: Understanding android security. IEEE security & privacy **7**(1), 50–57 (2009)

[15] Zhu, H.-j., Gu, W., Wang, L.-m., Xu, Z.-c., Sheng, V.S.: Android malware detection based on multi-head squeeze-and-excitation residual network. Expert Systems with Applications **212**, 118705 (2023)

[16] Ju, S.-h., Seo, H.-s., Kwak, J.: Research on android malware permission pattern using permission monitoring system. Multimedia Tools and Applications **75**, 14807–14817 (2016)

[17] Enck, W., Ongtang, M., McDaniel, P.: On lightweight mobile phone application certification. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, pp. 235–245 (2009)

[18] Alecakir, H., Can, B., Sen, S.: Attention: there is an inconsistency between android permissions and application metadata! International Journal of Information Security, 1–19 (2021)

[19] Zhang, Y., Yang, M., Yang, Z., Gu, G., Ning, P., Zang, B.: Permission use analysis for vetting undesirable behaviors in android apps. IEEE transactions on information forensics and security **9**(11), 1828–1842 (2014)

[20] Li, J., Sun, L., Yan, Q., Li, Z., Srisa-An, W., Ye, H.: Significant permission identification for machine-learning-based android malware detection. IEEE Transactions on Industrial Informatics **14**(7), 3216–3225 (2018)

[21] Şahin, D.Ö., Kural, O.E., Akleylek, S., Kılıç, E.: A novel permission-based android malware detection system using feature selection based on linear regression. Neural Computing and Applications, 1–16 (2021)

[22] Talha, K.A., Alper, D.I., Aydin, C.: Apk auditor: Permission-based android malware detection system. Digital Investigation **13**, 1–14 (2015)

[23] Varma, P.R.K., Raj, K.P., Raju, K.S.: Android mobile security by detecting and classification of malware based on permissions using machine learning algorithms. In: 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 294–299 (2017). IEEE

[24] Mahindru, A., Singh, P.: Dynamic permissions based android malware detection using machine learning techniques. In: Proceedings of the 10th Innovations in Software Engineering Conference, pp. 202–210 (2017)

[25] Doğru, İ.A., Önder, M.: Appperm analyzer: malware detection system based on android permissions and permission groups. International Journal of Software Engineering and Knowledge Engineering **30**(03), 427–450 (2020)

[26] Rathore, H., Sahay, S.K., Rajvanshi, R.,

Sewak, M.: Identification of significant permissions for efficient android malware detection. In: International Conference on Broadband Communications, Networks and Systems, pp. 33–52 (2020). Springer

[27] Shang, F., Li, Y., Deng, X., He, D.: Android malware detection method based on naive bayes and permission correlation algorithm. Cluster Computing **21**(1), 955–966 (2018)

[28] Tchakounté, F., Wandala, A.D., Tiguiane, Y.: Detection of android malware based on sequence alignment of permissions. Int. J. Comput.(IJC) **35**(1), 26–36 (2019)

[29] Ilham, S., Abderrahim, G., Abdelhakim, B.A.: Permission based malware detection in android devices. In: Proceedings of the 3rd International Conference on Smart City Applications, pp. 1–6 (2018)

[30] Şahın, D.Ö., Kural, O.E., Akleylek, S., Kiliç, E.: New results on permission based static analysis for android malware. In: 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1–4 (2018). IEEE

[31] D'Angelo, G., Palmieri, F., Robustelli, A.: A federated approach to android malware classification through perm-maps. Cluster Computing **25**(4), 2487–2500 (2022)

[32] Xiong, P., Wang, X., Niu, W., Zhu, T., Li, G.: Android malware detection with contrasting permission patterns. China Communications **11**(8), 1–14 (2014)

[33] Lu, T., Hou, S.: A two-layered malware detection model based on permission for android. In: 2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET), pp. 239–243 (2018). IEEE

[34] Kavitha, K., Salini, P., Ilamathy, V.: Exploring the malicious android applications and reducing risk using static analysis. In: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 1316–1319 (2016). IEEE

[35] Amer, E.: Permission-based approach for android malware analysis through ensemble-based voting model. In: 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), pp. 135–139 (2021). IEEE

[36] Chakravarty, S., *et al.*: Feature selection and evaluation of permission-based android malware detection. In: 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), pp. 795–799 (2020). IEEE

[37] Sirisha, P., Anuradha, T., *et al.*: Detection of permission driven malware in android using deep learning techniques. In: 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 941–945 (2019). IEEE

[38] Sahal, A.A., Alam, S., Soğukpinar, I.: Mining and detection of android malware based on permissions. In: 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 264–268 (2018). IEEE

[39] Mat, S.R.T., Ab Razak, M.F., Kahar, M.N.M., Arif, J.M., Zabidi, A.: Applying bayesian probability for android malware detection using permission features. In: 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), pp. 574–579 (2021). IEEE

[40] Wang, Z., Li, K., Hu, Y., Fukuda, A., Kong, W.: Multilevel permission extraction in android applications for malware detection. In: 2019 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–5 (2019). IEEE

[41] Park, J., Kang, M., Cho, S.-j., Han, H., Suh, K.: Analysis of permission selection techniques in machine learning-based malicious app detection. In: 2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE),

pp. 92–99 (2020). IEEE

[42] Liang, S., Du, X.: Permission-combination-based scheme for android mobile malware detection. In: 2014 IEEE International Conference on Communications (ICC), pp. 2301–2306 (2014). IEEE

[43] Wang, Y., Zheng, J., Sun, C., Mukkamala, S.: Quantitative security risk assessment of android permissions and applications. In: Data and Applications Security and Privacy XXVII: 27th Annual IFIP WG 11.3 Conference, DBSec 2013, Newark, NJ, USA, July 15-17, 2013. Proceedings 27, pp. 226–241 (2013). Springer

[44] Peng, H., Gates, C., Sarma, B., Li, N., Qi, Y., Potharaju, R., Nita-Rotaru, C., Molloy, I.: Using probabilistic generative models for ranking risks of android apps. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security, pp. 241–252 (2012)

[45] Pandita, R., Xiao, X., Yang, W., Enck, W., Xie, T.: {WHYPER}: Towards automating risk assessment of mobile applications. In: 22nd USENIX Security Symposium (USENIX Security 13), pp. 527–542 (2013)

[46] Samra, A.A.A., Yim, K., Ghanem, O.A.: Analysis of clustering technique in android malware detection. In: 2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 729–733 (2013). IEEE

[47] Yerima, S.Y., Sezer, S., McWilliams, G., Muttik, I.: A new android malware detection approach using bayesian classification. In: 2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA), pp. 121–128 (2013). IEEE

[48] Zarni Aung, W.Z.: Permission-based android malware detection. International Journal of Scientific & Technology Research **2**(3), 228–234 (2013)

[49] Yerima, S.Y., Sezer, S., McWilliams, G.: Analysis of bayesian classification-based approaches for android malware detection. IET Information Security **8**(1), 25–36 (2014)

[50] Sanz, B., Santos, I., Laorden, C., Ugarte-Pedrero, X., Bringas, P.G., Álvarez, G.: Puma: Permission usage to detect malware in android. In: International Joint Conference CISIS'12-ICEUTE 12-SOCO 12 Special Sessions, pp. 289–298 (2013). Springer

[51] Moonsamy, V., Rong, J., Liu, S.: Mining permission patterns for contrasting clean and malicious android applications. Future Generation Computer Systems **36**, 122–132 (2014)

[52] Backes, M., Gerling, S., Hammer, C., Maffei, M., Styp-Rekowsky, P.: Appguard–enforcing user requirements on android apps. In: Tools and Algorithms for the Construction and Analysis of Systems: 19th International Conference, TACAS 2013, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2013, Rome, Italy, March 16-24, 2013. Proceedings 19, pp. 543–548 (2013). Springer

[53] Wu, D.-J., Mao, C.-H., Wei, T.-E., Lee, H.-M., Wu, K.-P.: Droidmat: Android malware detection through manifest and api calls tracing. In: 2012 Seventh Asia Joint Conference on Information Security, pp. 62–69 (2012). IEEE

[54] Kato, H., Sasaki, T., Sasase, I.: Android malware detection based on composition ratio of permission pairs. IEEE Access **9**, 130006–130019 (2021)

[55] Arora, A., Peddoju, S.K., Conti, M.: Permpair: Android malware detection using permission pairs. IEEE Transactions on Information Forensics and Security **15**, 1968–1982 (2019)

[56] Alsoghyer, S., Almomani, I.: On the effectiveness of application permissions for android ransomware detection. In: 2020 6th Conference on Data Science and Machine Learning Applications (CDMA), pp. 94–99

(2020). IEEE

[57] Saleem, M.S., Mišić, J., Mišić, V.B.: Examining permission patterns in android apps using kernel density estimation. In: 2020 International Conference on Computing, Networking and Communications (ICNC), pp. 719–724 (2020). IEEE

[58] Ghasempour, A., Sani, N.F.M., Ovye, J.A.: Permission extraction framework for android malware detection. International Journal of Advanced Computer Science and Applications **11**(11) (2020)

[59] Shrivastava, G., Kumar, P.: Sensdroid: analysis for malicious activity risk of android application. Multimedia Tools and Applications **78**(24), 35713–35731 (2019)

[60] Upadhayay, M., Sharma, A., Garg, G., Arora, A.: Rpndroid: android malware detection using ranked permissions and network traffic. In: 2021 Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4), pp. 19–24 (2021). IEEE

[61] Lee, S.-A., Yoon, A.-R., Lee, J.-W., Lee, K.: An android malware detection system using a knowledge-based permission counting method. JOIV: International Journal on Informatics Visualization **6**(1), 138–144 (2022)

[62] Surendran, R., Thomas, T., Emmanuel, S.: A tan based hybrid model for android malware detection. Journal of Information Security and Applications **54**, 102483 (2020)

[63] Kabakus, A.T.: Droidmalwaredetector: A novel android malware detection framework based on convolutional neural network. Expert Systems with Applications **206**, 117833 (2022)

[64] Wang, X., Zhang, L., Zhao, K., Ding, X., Yu, M.: Mfdroid: A stacking ensemble learning framework for android malware detection. Sensors **22**(7), 2597 (2022)

[65] Akbar, F., Hussain, M., Mumtaz, R., Riaz, Q., Wahab, A.W.A., Jung, K.-H.: Permissions-based detection of android malware using machine learning. Symmetry **14**(4), 718 (2022)

[66] Zhu, H.-J., You, Z.-H., Zhu, Z.-X., Shi, W.-L., Chen, X., Cheng, L.: Droiddet: effective and robust detection of android malware using static analysis along with rotation forest model. Neurocomputing **272**, 638–646 (2018)

[67] Wang, W., Ren, C., Song, H., Zhang, S., Liu, P., et al.: Fgl_droid: an efficient android malware detection method based on hybrid analysis. Security and Communication Networks **2022** (2022)

[68] McLaughlin, N.: Malceiver: Perceiver with hierarchical and multi-modal features for android malware detection. arXiv preprint arXiv:2204.05994 (2022)

[69] Wang, W., Wang, X., Feng, D., Liu, J., Han, Z., Zhang, X.: Exploring permission-induced risk in android applications for malicious application detection. IEEE Transactions on Information Forensics and Security **9**(11), 1869–1882 (2014)

[70] Grace, M., Sughasiny, M.: Behaviour analysis of inter-app communication using a lightweight monitoring app for malware detection. Expert Systems with Applications **210**, 118404 (2022)

[71] Liu, K., Zhang, G., Chen, X., Liu, Q., Peng, L., Yurui, L.: Android malware detection based on sensitive patterns. Telecommunication Systems **82**(4), 435–449 (2023)

[72] Bayazit, E.C., Sahingoz, O.K., Dogan, B.: A deep learning based android malware detection system with static analysis. In: 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–6 (2022). IEEE

[73] Lee, J., Jang, H., Ha, S., Yoon, Y.: Android malware detection using machine learning with feature selection based on the genetic

algorithm. Mathematics **9**(21), 2813 (2021)

[74] Zhu, H.-J., Wang, L.-M., Zhong, S., Li, Y., Sheng, V.S.: A hybrid deep network framework for android malware detection. IEEE Transactions on Knowledge and Data Engineering **34**(12), 5558–5570 (2021)

[75] Almahmoud, M., Alzu'bi, D., Yaseen, Q.: Redroiddet: android malware detection based on recurrent neural network. Procedia Computer Science **184**, 841–846 (2021)

[76] Feng, J., Shen, L., Chen, Z., Wang, Y., Li, H.: A two-layer deep learning method for android malware detection using network traffic. IEEE Access **8**, 125786–125796 (2020)

[77] Kandukuru, S., Sharma, R.: Android malicious application detection using permission vector and network traffic analysis. In: 2017 2nd International Conference for Convergence in Technology (I2CT), pp. 1126–1132 (2017). IEEE

[78] Arora, A., Peddoju, S.K.: Ntpdroid: a hybrid android malware detector using network traffic and system permissions. In: 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), pp. 808–813 (2018). IEEE

[79] Ding, C., Luktarhan, N., Lu, B., Zhang, W.: A hybrid analysis-based approach to android malware family classification. Entropy **23**(8), 1009 (2021)

[80] Şahın, D.Ö., Akleylek, S., Kiliç, E.: Linregdroid: Detection of android malware using multiple linear regression models-based classifiers. IEEE Access **10**, 14246–14259 (2022)

[81] Idrees, F., Rajarajan, M., Chen, T.M., Rahulamathavan, Y., Naureen, A.: Andropin: Correlating android permissions and intents for malware detection. In: 2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 394–399 (2017). IEEE

[82] Khariwal, K., Singh, J., Arora, A.: Ipdroid: Android malware detection using intents and permissions. In: 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pp. 197–202 (2020). IEEE

[83] Idrees, F., Rajarajan, M.: Investigating the android intents and permissions for malware detection. In: 2014 IEEE 10th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 354–358 (2014). IEEE

[84] Bai, H., Xie, N., Di, X., Ye, Q.: Famd: A fast multifeature android malware detection framework, design, and implementation. IEEE Access **8**, 194729–194740 (2020)

[85] Taheri, R., Ghahramani, M., Javidan, R., Shojafar, M., Pooranian, Z., Conti, M.: Similarity-based android malware detection using hamming distance of static binary features. Future Generation Computer Systems **105**, 230–247 (2020)

[86] Alazab, M., Alazab, M., Shalaginov, A., Mesleh, A., Awajan, A.: Intelligent mobile malware detection using permission requests and api calls. Future Generation Computer Systems **107**, 509–521 (2020)

[87] Mathur, A., Podila, L.M., Kulkarni, K., Niyaz, Q., Javaid, A.Y.: Naticusdroid: A malware detection framework for android using native and custom permissions. Journal of Information Security and Applications **58**, 102696 (2021)

[88] Imtiaz, S.I., Rehman, S., Javed, A.R., Jalil, Z., Liu, X., Alnumay, W.S.: Deepamd: Detection and identification of android malware using high-efficient deep artificial neural network. Future Generation computer systems **115**, 844–856 (2021)

[89] Liu, Z., Wang, R., Japkowicz, N., Tang,

D., Zhang, W., Zhao, J.: Research on unsupervised feature learning for android malware detection based on restricted boltzmann machines. Future Generation Computer Systems **120**, 91–108 (2021)

[90] Chen, X., Shi, S., Xie, C., Yang, Z., Guo, Y., Fang, Y., Wen, W.: Suip: An android malware detection method based on data flow features. In: Journal of Physics: Conference Series, vol. 1812, p. 012010 (2021). IOP Publishing

[91] Guan, J., Jiang, X., Mao, B.: A method for class-imbalance learning in android malware detection. Electronics **10**(24), 3124 (2021)

[92] Mohamed, S.E., Ashaf, M., Ehab, A., Shereef, O., Metwaie, H., Amer, E.: Detecting malicious android applications based on api calls and permissions using machine learning algorithms. In: 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), pp. 1–6 (2021). IEEE

[93] Pushpa Latha, D.: Bat optimization algorithm for wrapper-based feature selection and performance improvement of android malware detection (2021)

[94] Gyunka, B.A., Abikoye, O.C., Adekunle, A.S.: Anomaly android malware detection: A comparative analysis of six classifiers. In: International Conference on Information and Communication Technology and Applications, pp. 145–157 (2020). Springer

[95] Taha, A.A., Malebary, S.J.: Hybrid classification of android malware based on fuzzy clustering and the gradient boosting machine. Neural Computing and Applications **33**, 6721–6732 (2021)

[96] Peng, T., Hu, B., Liu, J., Huang, J., Zhang, Z., He, R., Hu, X.: A lightweight multi-source fast android malware detection model. Applied Sciences **12**(11), 5394 (2022)

[97] Ashwini, S., Pai, M., Sangeetha, J.: Android malware classification based on static features of an application. In: Advances in Computing and Network Communications: Proceedings of CoCoNet 2020, Volume 2, pp. 567–581 (2021). Springer

[98] Jiang, X., Mao, B., Guan, J., Huang, X.: Android malware detection using fine-grained features. Scientific Programming **2020**, 1–13 (2020)

[99] Wang, J., Jing, Q., Gao, J., Qiu, X.: Sedroid: A robust android malware detector using selective ensemble learning. In: 2020 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–5 (2020). IEEE

[100] Rana, M.S., Sung, A.H.: Evaluation of advanced ensemble learning techniques for android malware detection. Vietnam Journal of Computer Science **7**(02), 145–159 (2020)

[101] Lu, T., Du, Y., Ouyang, L., Chen, Q., Wang, X.: Android malware detection based on a hybrid deep learning model. Security and Communication Networks **2020**, 1–11 (2020)

[102] Millar, S., McLaughlin, N., Rincon, J.M., Miller, P.: Multi-view deep learning for zero-day android malware detection. Journal of Information Security and Applications **58**, 102718 (2021)

[103] Barrera, D., Kayacik, H.G., Van Oorschot, P.C., Somayaji, A.: A methodology for empirical analysis of permission-based security models and its application to android. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, pp. 73–84 (2010)

[104] Shabtai, A., Fledel, Y., Elovici, Y.: Automated static code analysis for classifying android applications using machine learning. In: 2010 International Conference on Computational Intelligence and Security, pp. 329–333 (2010). IEEE

[105] Felt, A.P., Chin, E., Hanna, S., Song, D.,

Wagner, D.: Android permissions demystified. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, pp. 627–638 (2011)

[106] Erickson, J.L., Gibler, C.D., Chen, H.: Androidleaks: Detecting privacy leaks in android applications. Technical report, Sandia National Lab.(SNL-CA), Livermore, CA (United States) (2011)

[107] Sarma, B.P., Li, N., Gates, C., Potharaju, R., Nita-Rotaru, C., Molloy, I.: Android permissions: a perspective combining risks and benefits. In: Proceedings of the 17th ACM Symposium on Access Control Models and Technologies, pp. 13–22 (2012)

[108] Frank, M., Dong, B., Felt, A.P., Song, D.: Mining permission request patterns from android and facebook applications. In: 2012 IEEE 12th International Conference on Data Mining, pp. 870–875 (2012). IEEE

[109] Zhu, J., Guan, Z., Yang, Y., Yu, L., Sun, H., Chen, Z.: Permission-based abnormal application detection for android. In: Information and Communications Security: 14th International Conference, ICICS 2012, Hong Kong, China, October 29-31, 2012. Proceedings 14, pp. 228–239 (2012). Springer

[110] Peiravian, N., Zhu, X.: Machine learning for android malware detection using permission and api calls. In: 2013 IEEE 25th International Conference on Tools with Artificial Intelligence, pp. 300–305 (2013). IEEE

[111] Sanz, B., Santos, I., Laorden, C., Ugarte-Pedrero, X., Nieves, J., Bringas, P.G., Álvarez Marañón, G.: Mama: manifest analysis for malware detection in android. Cybernetics and Systems **44**(6-7), 469–488 (2013)

[112] Feldman, S., Stadther, D., Wang, B.: Manilyzer: automated android malware detection through manifest analysis. In: 2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems, pp. 767–772 (2014). IEEE

[113] Pehlivan, U., Baltaci, N., Acartürk, C., Baykal, N.: The analysis of feature selection methods and classification algorithms in permission based android malware detection. In: 2014 IEEE Symposium on Computational Intelligence in Cyber Security (CICS), pp. 1–8 (2014). IEEE

[114] Rahman, R., Islam, M.R., Ahmed, A., Hasan, M.K., Mahmud, H.: A study of permission-based malware detection using machine learning. In: 2022 15th International Conference on Security of Information and Networks (SIN), pp. 01–06 (2022). IEEE

[115] Rovelli, P., Vigfússon, Ý.: Pmds: permission-based malware detection system. In: Information Systems Security: 10th International Conference, ICISS 2014, Hyderabad, India, December 16-20, 2014, Proceedings 10, pp. 338–357 (2014). Springer

[116] Arp, D., Spreitzenbarth, M., Hubner, M., Gascon, H., Rieck, K., Siemens, C.: Drebin: Effective and explainable detection of android malware in your pocket. In: Ndss, vol. 14, pp. 23–26 (2014)

[117] Yerima, S.Y., Sezer, S., Muttik, I.: High accuracy android malware detection using ensemble learning. IET Information Security **9**(6), 313–320 (2015)

[118] Kang, H., Jang, J.-w., Mohaisen, A., Kim, H.K.: Detecting and classifying android malware using static analysis along with creator information. International Journal of Distributed Sensor Networks **11**(6), 479174 (2015)

[119] Zhao, K., Zhang, D., Su, X., Li, W.: Fest: A feature extraction and selection tool for android malware detection. In: 2015 IEEE Symposium on Computers and Communication (ISCC), pp. 714–720 (2015). IEEE

[120] Qiao, M., Sung, A.H., Liu, Q.: Merging permission and api features for android malware detection. In: 2016 5th IIAI International Congress on Advanced Applied

Informatics (IIAI-AAI), pp. 566–571 (2016). IEEE

[121] Chen, S., Xue, M., Tang, Z., Xu, L., Zhu, H.: Stormdroid: A streaminglized machine learning-based system for detecting android malware. In: Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security, pp. 377–388 (2016)

[122] Demertzis, K., Iliadis, L.: Bio-inspired hybrid intelligent method for detecting android malware. In: Knowledge, Information and Creativity Support Systems: Selected Papers from KICSS'2014-9th International Conference, Held in Limassol, Cyprus, on November 6-8, 2014, pp. 289–304 (2016). Springer

[123] Verma, S., Muttoo, S.K., Pal, S.: Mdroid: android based malware detection using mcm classifier. International Journal of Engineering Applied Sciences and Technology **1**(8), 206–215 (2016)

[124] Wang, K., Song, T., Liang, A.: Mmda: Metadata based malware detection on android. In: 2016 12th International Conference on Computational Intelligence and Security (CIS), pp. 598–602 (2016). IEEE

[125] Suarez-Tangil, G., Dash, S.K., Ahmadi, M., Kinder, J., Giacinto, G., Cavallaro, L.: Droidsieve: Fast and accurate classification of obfuscated android malware. In: Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy, pp. 309–320 (2017)

[126] Wang, X., Zhang, D., Su, X., Li, W., et al.: Mlifdect: android malware detection based on parallel machine learning and information fusion. Security and Communication Networks **2017** (2017)

[127] Li, D., Wang, Z., Li, L., Wang, Z., Wang, Y., Xue, Y.: Fgdetector: fine-grained android malware detection. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), pp. 311–318 (2017). IEEE

[128] Bhattacharya, A., Goswami, R.T.: Dmdam: data mining based detection of android malware. In: Proceedings of the First International Conference on Intelligent Computing and Communication, pp. 187–194 (2017). Springer

[129] Xie, N., Zeng, F., Qin, X., Zhang, Y., Zhou, M., Lv, C.: Repassdroid: Automatic detection of android malware based on essential permissions and semantic features of sensitive apis. In: 2018 International Symposium on Theoretical Aspects of Software Engineering (TASE), pp. 52–59 (2018). IEEE

[130] Xie, N., Di, X., Wang, X., Zhao, J.: Andro_md: android malware detection based on convolutional neural networks. International Journal of Performability Engineering **14**(3), 547 (2018)

[131] Ren, B., Liu, C., Cheng, B., Guo, J., Chen, J.: Mobisentry: Towards easy and effective detection of android malware on smartphones. Mobile Information Systems **2018**, 1–14 (2018)

[132] Tao, G., Zheng, Z., Guo, Z., Lyu, M.R.: Malpat: Mining patterns of malicious and benign android apps via permission-related apis. IEEE Transactions on Reliability **67**(1), 355–369 (2017)

[133] Namrud, Z., Kpodjedo, S., Bali, A., Talhi, C.: Deep-layer clustering to identify permission usage patterns of android app categories. IEEE Access **10**, 24240–24254 (2022)

[134] Alswaina, F., Elleithy, K.: Android malware permission-based multi-class classification using extremely randomized trees. IEEE Access **6**, 76217–76227 (2018)

[135] Qiu, J., Zhang, J., Luo, W., Pan, L., Nepal, S., Wang, Y., Xiang, Y.: A3cm: automatic capability annotation for android malware. IEEE Access **7**, 147156–147168 (2019)

[136] Zhu, H., Li, Y., Li, R., Li, J., You, Z.,

Song, H.: Sedmdroid: An enhanced stacking ensemble framework for android malware detection. IEEE Transactions on Network Science and Engineering **8**(2), 984–994 (2020)

[137] Feng, Y., Chen, L., Zheng, A., Gao, C., Zheng, Z.: Ac-net: Assessing the consistency of description and permission in android apps. IEEE Access **7**, 57829–57842 (2019)

[138] Aonzo, S., Merlo, A., Migliardi, M., Oneto, L., Palmieri, F.: Low-resource footprint, data-driven malware detection on android. IEEE Transactions on Sustainable Computing **5**(2), 213–222 (2017)

[139] Urooj, B., Shah, M.A., Maple, C., Abbasi, M.K., Riasat, S.: Malware detection: a framework for reverse engineered android applications through machine learning algorithms. IEEE Access **10**, 89031–89050 (2022)

[140] Wang, R., Wang, Z., Tang, B., Zhao, L., Wang, L.: Smartpi: Understanding permission implications of android apps from user reviews. IEEE Transactions on Mobile Computing **19**(12), 2933–2945 (2019)

[141] Wang, W., Gao, Z., Zhao, M., Li, Y., Liu, J., Zhang, X.: Droidensemble: Detecting android malicious applications with ensemble of string and structural static features. IEEE Access **6**, 31798–31807 (2018)

[142] Zhang, J., Tian, C., Duan, Z., Zhao, L.: Rtpdroid: Detecting implicitly malicious behaviors under runtime permission model. IEEE Transactions on Reliability **70**(3), 1295–1308 (2021)

[143] Kesswani, N., Lyu, H., Zhang, Z.: Analyzing android app privacy with gp-pp model. Ieee Access **6**, 39541–39546 (2018)

[144] İbrahim, M., Issa, B., Jasser, M.B.: A method for automatic android malware detection based on static analysis and deep learning. IEEE Access **10**, 117334–117352 (2022)

[145] Arshad, S., Shah, M.A., Wahid, A., Mehmood, A., Song, H., Yu, H.: Samadroid: a novel 3-level hybrid malware detection model for android operating system. IEEE Access **6**, 4321–4339 (2018)

[146] Yuan, Z., Lu, Y., Xue, Y.: Droiddetector: android malware characterization and detection using deep learning. Tsinghua Science and Technology **21**(1), 114–123 (2016)

[147] Zhou, H., Yang, X., Pan, H., Guo, W.: An android malware detection approach based on simgru. IEEE Access **8**, 148404–148410 (2020)

[148] Cilleruelo, C., De-Marcos, L., Martinez-Herráiz, J.-J., *et al.*: Malware detection inside app stores based on lifespan measurements. IEEE Access **9**, 119967–119976 (2021)

[149] Firdaus, A., Anuar, N.B., Karim, A., Razak, M.F.A.: Discovering optimal features using static analysis and a genetic search based method for android malware detection. Frontiers of Information Technology & Electronic Engineering **19**(6), 712–736 (2018)

[150] Wang, C., Xu, Q., Lin, X., Liu, S.: Research on data mining of permissions mode for android malware detection. Cluster Computing **22**, 13337–13350 (2019)

[151] Singh, A.K., Jaidhar, C., Kumara, M.A.: Experimental analysis of android malware detection based on combinations of permissions and api-calls. Journal of Computer Virology and Hacking Techniques **15**, 209–218 (2019)

[152] Rafiq, H., Aslam, N., Aleem, M., Issac, B., Randhawa, R.H.: Andromalpack: enhancing the ml-based malware classification by detection and removal of repacked apps for android systems. Scientific Reports **12**(1), 19534 (2022)

[153] Mahdavifar, S., Alhadidi, D., Ghorbani, A.A.: Effective and efficient hybrid android malware classification using pseudo-label stacked auto-encoder. Journal of network

and systems management **30**, 1–34 (2022)

[154] Seraj, S., Khodambashi, S., Pavlidis, M., Polatidis, N.: Hamdroid: permission-based harmful android anti-malware detection using neural networks. Neural Computing and Applications **34**(18), 15165–15174 (2022)

[155] Mahindru, A., Sangal, A.: Somdroid: Android malware detection by artificial neural network trained using unsupervised learning. Evolutionary Intelligence **15**(1), 407–437 (2022)

[156] Anupama, M., Vinod, P., Visaggio, C.A., Arya, M., Philomina, J., Raphael, R., Pinhero, A., Ajith, K., Mathiyalagan, P.: Detection and robustness evaluation of android malware classifiers. Journal of Computer Virology and Hacking Techniques **18**(3), 147–170 (2022)

[157] Chen, X., Yu, H., Yu, D., Chen, J., Sun, X.: Predicting android malware combining permissions and api call sequences. Software Quality Journal **31**(3), 655–685 (2023)

[158] Mahindru, A., Sangal, A.: Fsdroid:-a feature selection technique to detect malware from android using machine learning techniques: Fsdroid. Multimedia Tools and Applications **80**, 13271–13323 (2021)

[159] Tchakounté, F., Ngassi, R.C.N., Kamla, V.C., Udagepola, K.P.: Limondroid: a system coupling three signature-based schemes for profiling android malware. Iran Journal of Computer Science **4**, 95–114 (2021)

[160] Nissim, N., Moskovitch, R., BarAd, O., Rokach, L., Elovici, Y.: Aldroid: efficient update of android anti-virus software using designated active learning methods. Knowledge and Information Systems **49**, 795–833 (2016)

[161] Peynirci, G., Eminağaoğlu, M., Karabulut, K.: Feature selection for malware detection on the android platform based on differences of idf values. Journal of Computer Science and Technology **35**, 946–962 (2020)

[162] Nauman, M., Tanveer, T.A., Khan, S., Syed, T.A.: Deep neural architectures for large scale android malware analysis. Cluster Computing **21**, 569–588 (2018)

[163] Bhattacharya, A., Goswami, R.T., Mukherjee, K.: A feature selection technique based on rough set and improvised pso algorithm (psors-fs) for permission based detection of android malwares. International journal of machine learning and cybernetics **10**, 1893–1907 (2019)

[164] Bao, L., Lo, D., Xia, X., Li, S.: Automated android application permission recommendation. Science China Information Sciences **60**, 1–17 (2017)

[165] Rubio-Medrano, C.E., Soundrapandian, P.K.D., Hill, M., Claramunt, L., Baek, J., Ahn, G.-J.: Dypoldroid: Protecting against permission-abuse attacks in android. Information Systems Frontiers **25**(2), 529–548 (2023)

[166] Shatnawi, A.S., Yassen, Q., Yateem, A.: An android malware detection approach based on static feature analysis using machine learning algorithms. Procedia Computer Science **201**, 653–658 (2022)

[167] Smmarwar, S.K., Gupta, G.P., Kumar, S., Kumar, P.: An optimized and efficient android malware detection framework for future sustainable computing. Sustainable Energy Technologies and Assessments **54**, 102852 (2022)

[168] Arif, J.M., Ab Razak, M.F., Mat, S.R.T., Awang, S., Ismail, N.S.N., Firdaus, A.: Android mobile malware detection using fuzzy ahp. Journal of Information Security and Applications **61**, 102929 (2021)

[169] Guerra-Manzanares, A., Bahsi, H., Nõmm, S.: Kronodroid: Time-based hybrid-featured dataset for effective android malware detection and characterization. Computers & Security **110**, 102399 (2021)

[170] Bhat, P., Dutta, K.: A multi-tiered feature selection model for android malware

detection based on feature discrimination and information gain. Journal of King Saud University-Computer and Information Sciences **34**(10), 9464–9477 (2022)

[171] Elayan, O.N., Mustafa, A.M.: Android malware detection using deep learning. Procedia Computer Science **184**, 847–852 (2021)

[172] Syrris, V., Geneiatakis, D.: On machine learning effectiveness for malware detection in android os using static analysis data. Journal of information security and applications **59**, 102794 (2021)

[173] Idrees, F., Rajarajan, M., Conti, M., Chen, T.M., Rahulamathavan, Y.: Pindroid: A novel android malware detection system using ensemble learning methods. Computers & Security **68**, 36–46 (2017)

[174] Rehman, Z.-U., Khan, S.N., Muhammad, K., Lee, J.W., Lv, Z., Baik, S.W., Shah, P.A., Awan, K., Mehmood, I.: Machine learning-assisted signature and heuristic-based detection of malwares in android devices. Computers & Electrical Engineering **69**, 828–841 (2018)

[175] Martín, A., Lara-Cabrera, R., Camacho, D.: Android malware detection through hybrid features fusion and ensemble classifiers: The andropytool framework and the omnidroid dataset. Information Fusion **52**, 128–142 (2019)

[176] Navarro, L.C., Navarro, A.K., Grégio, A., Rocha, A., Dahab, R.: Leveraging ontologies and machine-learning techniques for malware analysis into android permissions ecosystems. Computers & Security **78**, 429–453 (2018)

[177] Milosevic, N., Dehghantanha, A., Choo, K.-K.R.: Machine learning aided android malware classification. Computers & Electrical Engineering **61**, 266–274 (2017)

[178] Alzaylaee, M.K., Yerima, S.Y., Sezer, S.: Dl-droid: Deep learning based android malware detection using real devices. Computers & Security **89**, 101663 (2020)

[179] Cai, L., Li, Y., Xiong, Z.: Jowmdroid: Android malware detection based on feature weighting with joint optimization of weight-mapping and classifier parameters. Computers & Security **100**, 102086 (2021)

[180] Badhani, S., Muttoo, S.K.: Cendroid—a cluster-ensemble classifier for detecting malicious android applications. Computers & Security **85**, 25–40 (2019)

[181] Alqatawna, J., Ala'M, A.-Z., Hassonah, M.A., Faris, H., *et al.*: Android botnet detection using machine learning models based on a comprehensive static analysis approach. Journal of Information Security and Applications **58**, 102735 (2021)

[182] Sheen, S., Anitha, R., Natarajan, V.: Android based malware detection using a multifeature collaborative decision fusion approach. Neurocomputing **151**, 905–912 (2015)

[183] OS, J.N., *et al.*: Detection of malicious android applications using ontology-based intelligent model in mobile cloud environment. Journal of Information Security and Applications **58**, 102751 (2021)

[184] Song, J., Han, C., Wang, K., Zhao, J., Ranjan, R., Wang, L.: An integrated static detection and analysis framework for android. Pervasive and Mobile Computing **32**, 15–25 (2016)

[185] Zhang, Y., Yang, M., Xu, B., Yang, Z., Gu, G., Ning, P., Wang, X.S., Zang, B.: Vetting undesirable behaviors in android apps with permission use analysis. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, pp. 611–622 (2013)

[186] Yang, S., Zeng, Z., Song, W.: Permdroid: automatically testing permission-related behaviour of android applications. In: Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 593–604 (2022)

[187] Thiyagarajan, J., Akash, A., Murugan, B.:

Improved real-time permission based malware detection and clustering approach using model independent pruning. IET Information Security **14**(5), 531–541 (2020)

[188] Qaisar, Z.H., Li, R.: Multimodal information fusion for android malware detection using lazy learning. Multimedia Tools and Applications, 1–15 (2022)

[189] Appice, A., Andresini, G., Malerba, D.: Clustering-aided multi-view classification: a case study on android malware detection. Journal of intelligent information systems **55**, 1–26 (2020)

[190] Zhu, H.-J., Jiang, T.-H., Ma, B., You, Z.-H., Shi, W.-L., Cheng, L.: Hemd: a highly efficient random forest-based malware detection framework for android. Neural Computing and Applications **30**, 3353–3361 (2018)

[191] Altaher, A.: An improved android malware detection scheme based on an evolving hybrid neuro-fuzzy classifier (ehnfc) and permission-based features. Neural Computing and Applications **28**, 4147–4157 (2017)

[192] Su, X., Shi, W., Qu, X., Zheng, Y., Liu, X.: Droiddeep: using deep belief network to characterize and detect android malware. Soft Computing **24**(8), 6017–6030 (2020)

[193] Mahindru, A., Sangal, A.: Mldroid—framework for android malware detection using machine learning techniques. Neural Computing and Applications **33**(10), 5183–5240 (2021)

[194] Dehkordy, D.T., Rasoolzadegan, A.: A new machine learning-based method for android malware detection on imbalanced dataset. Multimedia Tools and Applications **80**, 24533–24554 (2021)

[195] Nguyen, D.V., Nguyen, G.L., Nguyen, T.T., Ngo, A.H., Pham, G.T.: Minad: Multi-inputs neural network based on application structure for android malware detection. Peer-to-Peer Networking and Applications, 1–15 (2022)

[196] Taheri, R., Javidan, R., Pooranian, Z.: Adversarial android malware detection for mobile multimedia applications in iot environments. Multimedia Tools and Applications **80**, 16713–16729 (2021)

[197] Mahesh, P.S., Hemalatha, S.: An efficient android malware detection using adaptive red fox optimization based cnn. Wireless Personal Communications **126**(1), 679–700 (2022)

[198] Firdaus, A., Anuar, N.B., Razak, M.F.A., Sangaiah, A.K.: Bio-inspired computational paradigm for feature investigation and malware detection: interactive analytics. Multimedia Tools and Applications **77**, 17519–17555 (2018)

[199] Shrivastava, G., Kumar, P.: Intent and permission modeling for privacy leakage detection in android. Energy Systems **13**(3), 567–580 (2022)

[200] Varsha, M., Vinod, P., Dhanya, K.: Identification of malicious android app using manifest and opcode features. Journal of Computer Virology and Hacking Techniques **13**, 125–138 (2017)

[201] Deypir, M.: Entropy-based security risk measurement for android mobile applications. Soft Computing **23**(16), 7303–7319 (2019)

[202] Mahindru, A., Sangal, A.: Semidroid: a behavioral malware detector based on unsupervised machine learning techniques using feature selection approaches. International Journal of Machine Learning and Cybernetics **12**, 1369–1411 (2021)

[203] Keyvanpour, M.R., Barani Shirzad, M., Heydarian, F.: Android malware detection applying feature selection techniques and machine learning. Multimedia Tools and Applications **82**(6), 9517–9531 (2023)

[204] Razak, M.F.A., Anuar, N.B., Othman, F., Firdaus, A., Afifi, F., Salleh, R.: Bio-inspired for features optimization and malware detection. Arabian Journal for Science

and Engineering **43**, 6963–6979 (2018)

[205] Xie, N., Wang, X., Wang, W., Liu, J.: Fingerprinting android malware families. Frontiers of Computer Science **13**, 637–646 (2019)

[206] Mahindru, A., Sangal, A.: Hybridroid: an empirical analysis on effective malware detection model developed using ensemble methods. The Journal of Supercomputing **77**, 8209–8251 (2021)

[207] Ali, T., Khan, Y., Ali, T., Faizullah, S., Alghamdi, T., Anwar, S.: An automated permission selection framework for android platform. Journal of Grid Computing **18**, 547–561 (2020)

[208] Sun, J., Yan, K., Liu, X., Yang, C., Fu, Y.: Malware detection on android smartphones using keywords vector and svm. In: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), pp. 833–838 (2017). IEEE

[209] AlJarrah, M.N., Yaseen, Q.M., Mustafa, A.M.: A context-aware android malware detection approach using machine learning. Information **13**(12), 563 (2022)

[210] Gharib, A., Ghorbani, A.: Dna-droid: A real-time android ransomware detection framework. In: Network and System Security: 11th International Conference, NSS 2017, Helsinki, Finland, August 21–23, 2017, Proceedings 11, pp. 184–198 (2017). Springer

[211] Sun, L., Wei, X., Zhang, J., He, L., Philip, S.Y., Srisa-an, W.: Contaminant removal for android malware detection systems. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 1053–1062 (2017). IEEE

[212] Chen, J., Wang, C., Zhao, Z., Chen, K., Du, R., Ahn, G.-J.: Uncovering the face of android ransomware: Characterization and real-time detection. IEEE Transactions on Information Forensics and Security **13**(5), 1286–1300 (2017)

[213] Senliol, B., Gulgezen, G., Yu, L., Cataltepe, Z.: Fast correlation based filter (fcbf) with a different search strategy. In: 2008 23rd International Symposium on Computer and Information Sciences, pp. 1–4 (2008). IEEE

[214] Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with java implementations. Acm Sigmod Record **31**(1), 76–77 (2002)

[215] Fan, M., Liu, J., Luo, X., Chen, K., Tian, Z., Zheng, Q., Liu, T.: Android malware familial classification and representative sample selection via frequent subgraph analysis. IEEE Transactions on Information Forensics and Security **13**(8), 1890–1905 (2018)

[216] Wei, F., Li, Y., Roy, S., Ou, X., Zhou, W.: Deep ground truth analysis of current android malware. In: Detection of Intrusions and Malware, and Vulnerability Assessment: 14th International Conference, DIMVA 2017, Bonn, Germany, July 6-7, 2017, Proceedings 14, pp. 252–276 (2017). Springer

[217] Zhou, Y., Jiang, X.: Dissecting android malware: Characterization and evolution. In: 2012 IEEE Symposium on Security and Privacy, pp. 95–109 (2012). IEEE

[218] Abdul Kadir, A.F., Stakhanova, N., Ghorbani, A.A.: Android botnets: What urls are telling us. In: Network and System Security: 9th International Conference, NSS 2015, New York, NY, USA, November 3-5, 2015, Proceedings 9, pp. 78–91 (2015). Springer

[219] Lashkari, A.H., Kadir, A.F.A., Taheri, L., Ghorbani, A.A.: Toward developing a systematic approach to generate benchmark android malware datasets and classification. In: 2018 International Carnahan Conference on Security Technology (ICCST), pp. 1–7 (2018). IEEE

[220] Andronio, N., Zanero, S., Maggi, F.: Heldroid: Dissecting and detecting mobile ransomware. In: Research in Attacks, Intrusions, and Defenses: 18th International Symposium, RAID 2015, Kyoto, Japan, November 2-4, 2015. Proceedings 18, pp. 382–404 (2015). Springer

[221] Jang, J.-w., Kang, H., Woo, J., Mohaisen, A., Kim, H.K.: Andro-autopsy: Anti-malware system based on similarity matching of malware and malware creator-centric information. Digital Investigation **14**, 17–35 (2015)

[222] Avdiienko, V., Kuznetsov, K., Gorla, A., Zeller, A., Arzt, S., Rasthofer, S., Bodden, E.: Mining apps for abnormal usage of sensitive data. In: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, vol. 1, pp. 426–436 (2015). IEEE

[223] Damshenas, M., Dehghantanha, A., Choo, K.-K.R., Mahmud, R.: M0droid: An android behavioral-based malware detection model. Journal of Information Privacy and Security **11**(3), 141–157 (2015)

[224] Arslan, R.S.: Androanalyzer: android malicious software detection based on deep learning. PeerJ Computer Science **7**, 533 (2021)

[225] Mahdavifar, S., Kadir, A.F.A., Fatemi, R., Alhadidi, D., Ghorbani, A.A.: Dynamic android malware category classification using semi-supervised deep learning. In: 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), pp. 515–522 (2020). IEEE

[226] Maiorca, D., Ariu, D., Corona, I., Aresu, M., Giacinto, G.: Stealth attacks: An extended insight into the obfuscation effects on android malware. Computers & Security **51**, 16–31 (2015)

SynOpen S. Kumar et al. THIEME OPEN ACCESS Review

A Decade of Exploration of Transition-Metal-Catalyzed Cross-Coupling Reactions: An Overview

Saurav Kumar[a]
Jyoti[a]
Deepak Gupta[a]
Gajendra Singh[b]
Anil Kumar*[a]

[a] Department of Applied Chemistry, Delhi Technological University, Delhi-110042, India
anil_kumar@dce.ac.in
[b] Department of Chemistry, Deshbandhu College, University of Delhi, Delhi-110019, India

X = H, Halogen, BR_2, SiR_3, SO_2R, NR_2, COOH etc.
Y = H, NH, OH, SH, PH, CH, RM, BR_2 etc.
R^1 = Alkyl, Aryl, Heteroaryl
R^2 = Alkyl, Aryl, Heteroaryl

Received: 01.08.2023
Accepted after revision: 21.09.2023
Published online: 13.11.2023 (Version of Record)
DOI: 10.1055/s-0040-1720096; Art ID: SO-2023-08-0055-RV

License terms:
© **2023**. The Author(s). This is an open access article published by Thieme under the terms of the Creative Commons Attribution License, permitting unrestricted use, distribution and reproduction, so long as the original work is properly cited. (https://creativecommons.org/licenses/by/4.0/)

Abstract During the previous couple of decades, transition-metal (Fe, Co, Cu, Ni, Ru, Rh, Pd, Ag, Au) catalyzed inter- and intramolecular coupling reactions have attracted huge attention for the construction of C–C and C–heteroatom (like C–N, C–P, C–O, C–S, etc.) bonds to synthesize a diverse range of polymers, fine chemicals, and agrochemicals (mainly fungicides, herbicides, and insecticides), as well as biologically and pharmaceutically important organic molecules. Furthermore, the employment of lower cost and easily available metals such as first-row transition-metal salts or metal complexes of Fe, Co, Cu, Ni as catalysts compared to the precious metals such as Pd, Ag, Au in cross-coupling reactions have led to major advances in applications within the fields of synthesis. A number of cross-coupling reactions catalyzed by transition metals have been explored, including Suzuki, Heck, Sonogashira, Stille, Kumada, Kochi, Murahashi, Corriu, and Negishi reactions, as well as carbonylative, decarboxylative, reactions and α-arylations. In this review, we offer a comprehensive summary of the cross-coupling reaction catalyzed by different transition metals from the year 2009 to date.

1 Introduction
2 Pd-Catalyzed Reactions
2.1 C–C Cross-Coupling Reactions
2.2 C–N Cross-Coupling Reactions
2.3 C–P Cross-Coupling Reactions
3 Ni-Catalyzed Cross-Coupling Reactions
3.1 C–C Cross-Coupling Reactions
4 Cu-Catalyzed Cross-Coupling Reactions
4.1 C–C Cross-Coupling Reactions
4.2 C–O Cross-Coupling Reactions
4.2 C–N Cross-Coupling Reactions
4.4 C–P Cross-Coupling Reactions
4.5 C–Se Cross-Coupling Reactions
4.6 C–S Cross-Coupling Reactions
5 Fe-Catalyzed Reactions
5.1 C–C Cross-Coupling Reactions
5.2 C–S Cross-Coupling Reactions
6 Co-Catalyzed Reactions
7 Transition-Metal Nanoparticle-Promoted Reactions
7.1 Pd Nanoparticles
7.2 Cu Nanoparticles
8 Miscellaneous Reactions
9 Perspectives and Future Directions

Keywords cross coupling, transition metals, catalysis

1 Introduction

Over the past years, the construction of carbon–carbon and carbon–heteroatom bonds via cross-coupling reactions catalyzed by transition metals, such as Suzuki–Miyaura,[1] Heck,[2] Sonogashira,[3] Stille,[4] Negishi,[5] Kumada,[6] and Hiyama[7] reactions, have remained the most widely employed synthesis protocols in the chemical industry. These reactions represent the fundamental criteria for a number of basic technologies in modern synthetic organic chemistry and have been widely applied in a variety of academic and industrial process,[8,9] including the synthesis of natural products,[10,11] biologically active small molecule, materials science, medicinal, supramolecular catalysis, and coordination chemistry. In addition, several of these reactions have been commercially employed in the fields of pharmaceuticals, agrochemical conjugated polymers,[12,13] and crystalline liquids,[14–16] in the active components of organic light-emitting diodes (OLEDs),[17,18] and as industrial chemicals[12] etc.

The first breakthrough in the direction of cross-coupling was the copper-catalyzed synthesis of biaryl compounds from aryl halides published by F. Ullmann in 1901.[19] This discovery was not limited to a mere presentation of new synthesis methodology, but rather brought the realization that carbon–carbon bonds can be made in a laboratory synthetically. After a long gap of nearly seven decades, the discovery by Ullmann gained the recognition and a variety

Biographical Sketches

Jyoti received her Bachelor's degree in Chemistry from the University of Delhi and her Master's degree from Maharishi Dayanand University. After receiving the prestigious CSIR-NET Junior Research Fellowship, she started her Ph.D. journey in July 2018 under the supervision of Prof. Sudhir G. Warkar and Prof. Anil Kumar at the Department of Applied Chemistry, DTU. Her research focuses on the synthesis and applications of Cobalt corroles. She is the recipient of Commendable Research Excellence award in 2021 at DTU. She wishes to explore the wider perspectives of macrocyclic chemistry in future.

Saurav Kumar hails from Uttar Pradesh, India. He completed his B.Sc. (H) Chemistry in 2016 at the University of Delhi. Following that, he pursued his M.Sc. in Organic Chemistry at CCS University in 2018. In the same year, he was awarded with the CSIR-NET JRF. In 2018, Saurav embarked on his Ph.D. journey at Delhi Technological University under the guidance of Prof. Anil Kumar, focusing on organic synthesis.

Deepak Gupta is a guest faculty member at the Department of Applied Chemistry, Delhi Technological University. His research is focused on electrochemical conversions and developing sustainable materials for energy storage. He has published more than 20 research articles with total impact factor of greater than 100. He is a recipient of research funding from various prestigious agencies such as Council of Scientific and Industrial Research (India), DAAD (Germany), European Research Council (Belgium) and Science and Engineering Research Board (India). He is a member of Royal Society of Chemistry and reviewer with various high-impact journals.

Gajendra Singh is an Associate Professor at the Department of Chemistry, Deshbandhu College (University of Delhi). He completed his education at the University of Delhi and CCS University. Dr. Singh has published several research papers in national and international journals. He has also presented papers at approximately 30 conferences and seminars. In addition to his research and academic achievements, he also serves as a member of the editorial board of the Journal of Heterocyclic Letters and Universe Journal of Education & Humanities. He is a lifetime member of the Indian Society of Analytical Scientists (ISAS).

Anil Kumar is a full professor at the Department of Applied Chemistry, Delhi Technological University (formerly, Delhi College of Engineering), Delhi, India. He received his master's and doctorate degrees in chemistry from University of Roorkee, Roorkee (Now IIT Roorkee) and Indian Institute of Technology, Kanpur, India, respectively. He has worked with Professor Gross at Technion, Israel Institute of Technology, Haifa, Israel, as both a post-doctoral fellow and a visiting associate professor. Kumar gives credit to all his mentors 'Gurujis' (Prof. S. Sarkar, Prof. C. H. Hung and Prof. Zeev Gross) for enlightening him through the path of knowledge. His current research interest is corrole and benziporphodimethene–based coordination chemistry and its applications. Kumar lives in Delhi with his wife (Hemlata) and two children (Sarthak Pal and Tanish Pal).

of modifications and new directions subsequently emerged. In particular, the discovery of Kumada coupling in 1972, in which reactive chemical halides and alkenyl/aryl halides were combined using Ni or Pd catalysts, paved the way for the discovery of modern transition-metal (TM) catalyzed cross-coupling methods.[20,21] These developments were followed by Heck in 1972, where unsaturated halides and olefins were combined with Pd catalysts,[22–24] and Sonogashira in 1975, where terminal alkynes and aryl or vinyl halides were combined with Pd and Cu catalysts.[25,26] In an attempt to extend the outreach of these methodologies, Negishi (1977) used Pd or Ni catalysts to combine organozinc compounds with organic halides or triflates.[27,28] Using a similar strategy, in 1978, Stille coupled organotin compounds with a variety of organic electrophiles using Pd,[29,30] Suzuki, in 1979, presented the coupling of boric acid and organohalogen compounds using Pd,[31–33] Hiyama, in 1988, used organosilanes and organic halides with Pd,[34–36] Buchwald–Hartwig, 1994, coupled amines with aryl halides using Pd,[37] and other cross-couplings have been described.[38–40]

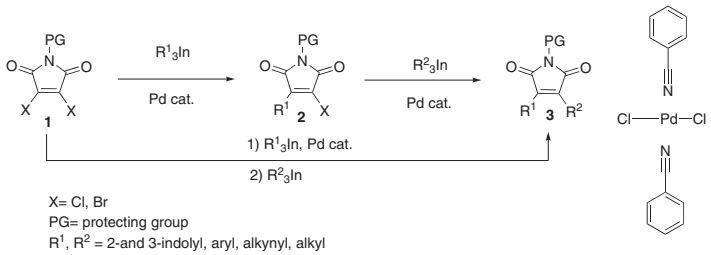
The importance of these palladium-catalyzed cross-couplings was finally recognized when the Nobel Prize in Chemistry in 2010 was jointly awarded to A. Suzuki, R. Heck, and E. Negishi.[41–43] Since the early developments in the dynamic area of cross-coupling reactions nearly fifty years ago, the diversity, scope, reactivity, value effectiveness, toxicity, required synthetic skill, and number of workable applications and limitations of TMs such as palladium,[44–48] iron,[38,49–52] cobalt,[53–56] nickel,[57–61] copper,[62–64] rhodium,[65–69] ruthenium,[70,71] and iridium,[72] has led to thousands of publications in this field, and many reviews and books have cataloged the advancements. With all this, bottlenecks in cross-coupling reactions have encouraged scientists and researchers to formulate novel catalysts primarily based on naturally abundant and environmentally benign elements. Therefore, compared to the widespread applications of late and noble transition-metals in TM-catalyzed cross-couplings, much attention is paid to the first-native transition metals such as Fe, Co, Ni, and Cu due to their obvious advantages, such as high earth abundance, low cost, reduced toxicity, higher nucleophilicity, unique catalytic properties, and environmental friendliness.[73–75]

In particular, Pd-catalyzed coupling appears to be one of the most commonly use reactions for producing good quality chemical compounds on a good scale and it represents one of the most powerful and diverse techniques available to synthetic organic chemists.[76,77] In fact, extensive research has confirmed that nickel-based catalysts are more potent and flexible catalysts for C–C,[78] C–O,[78] C–P,[79] and C–N[80] bond construction.

Moreover, the past 20 years have witnessed an impressive expansion in interest in the advancement of iron-based cross-coupling responses.[38,81,82] Developed by Kharasch and Field in 1941, iron-catalyzed cross-coupling reactions were first detailed in which Grignard reagents were combined with aryl halides under the influence of $FeCl_3$.[83] It is intriguing to note here that it required an additional 30 years for a subsequent report to be published that sped up progress in this field; Kochi clarified the cross-coupling of Grignard reagents with alkenyl halides catalyzed by $FeCl_3$.

Following Kochi's report, simultaneous work[84] led by the groups of Julia,[85] Molander,[86] Cahiez,[87] and Fürstner,[88] in addition to considerable contributions from the groups of Hayashi,[89] Nakamura,[90] and Itami,[91] Bedford,[92] Knochel,[93] and Shi[94] during the 1990s and early 2000s, dynamically marked the resurgence of Fe-catalyzed reactions. From that point forward, various new iron-catalyzed cross-coupling responses have been discovered that are highly appealing to research in organic synthesis.[81] Nevertheless, the developments towards a viable and economical protocol for iron-mediated cross-coupling catalysis is sluggish in comparison to analogous palladium and nickel-primarily based catalysis.

Given the growing recognition of the importance of cross-coupling chemistry, this review offers a complete overview of the successful applications of numerous transition-metals in cross-coupling strategies that have been carried out as key steps, including Suzuki, Heck, Sonogashira, Stille, Kumada, Kochi, Murahashi, Corriu, and Negishi reactions, in addition to carbonylative, decarboxylative, C–N cross-coupling reactions and α-arylative reactions, to synthesize heterocycles, organic materials, natural products, and medicinally relevant compounds. The practical challenges and perspectives for these TM-catalyzed coupling protocols for the discovery of polymers, natural products,



Scheme 1 Pd-catalyzed direct and sequential cross-coupling reaction of triorganoindium reagents and 3,4-dihalomaleimides[95]

agrochemicals, and biologically active compounds are briefly discussed in the final section. This work includes references published from the year 2011 to date that cover the most important developments in this rapidly progressing field. The content of this review is categorized into various transition-metal-catalyzed reactions such as Pd-mediated reactions.

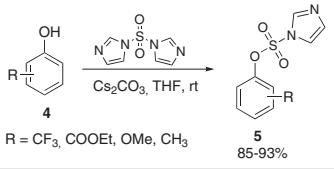
2 Pd-Catalyzed Reactions

The discussion in this section is bounded to the use of palladium catalysis in various cross-coupling reactions and is generally presented in chronological order.

2.1 C–C Cross-Coupling Reaction

In 2009, Bouissane and co-workers[95] discovered a palladium-mediated sequential or stepwise one-pot cross-coupling reaction with various triorganoindium reagents (40–50 mol%) with 3,4-dihalomaleimides **1** to afford a variety of aryl, heteroaryl, alkyl, alkynyl, 2- and 3-indolyl 3,4-disubstituted maleimides with satisfactory yields with high selectivity and atom economy (Scheme 1).

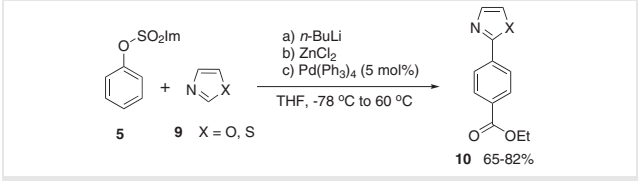
At the same time, Raju and co-workers first synthesized aryl imidazolylsulfonates **5** as a cost-effective alternative to triflates, which was shown to participate as a fully competent electrophilic coupling partner in palladium-catalyzed cross-coupling Negishi and Suzuki–Miyaura reactions in excellent yields (Scheme 2, Scheme 3, and Scheme 4).[96]



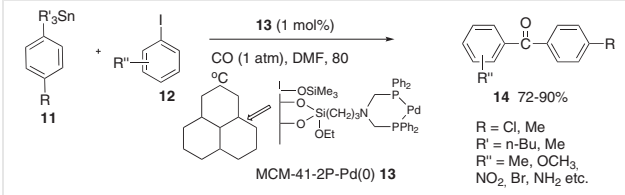
Scheme 2 Synthesis of imidazolylsulfonates[96]

The first successful heterogeneous carbonylative Stille cross-coupling reaction of organostannanes **11** with aryl iodides **12** was demonstrated in 2009 by Cai et al.[97] in the presence of a catalytic amount of an MCM-41-supported bidentate phosphine palladium(0) complex [MCM-41-2P-Pd(0)] (**13**) (Scheme 5). The reaction was carried out at 80 °C under carbon monoxide atmosphere in the presence of DMF, producing a variety of unsymmetrical ketones **14** in high reaction yields.

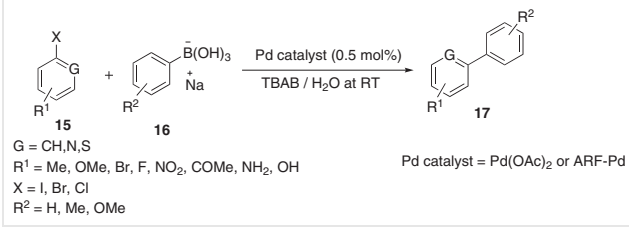
In 2010, Basu and co-workers[98] deciphered the ligand-free, on-water, Pd-catalyzed Suzuki–Miyaura (SM) coupling of the easily accessible sodium salt of aryl trihydroxyborate (**16**) with a variety of aryl halides (**15**) under aerobic conditions. The protocol was also applicable to very challenging substrates like aryl chlorides bearing electron-withdrawing groups, in good to excellent yields. Further, the authors demonstrated that this protocol was effective with heterogeneous palladium-catalysts and also exhibited the synthesis of pharmaceutically important benzotriazole **20** and benzimidazole-based biphenyl scaffolds **19** (Scheme 6 and Scheme 7).



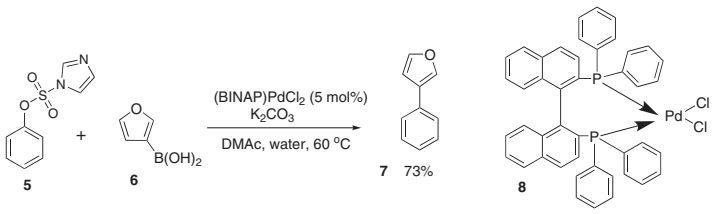
Scheme 4 Aryl imidazolylsulfonates as coupling partner in Pd-catalyzed Negishi cross-coupling reaction[96]



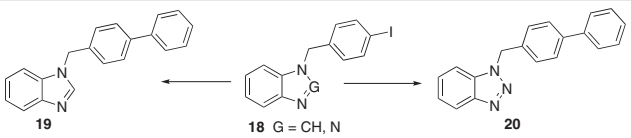
Scheme 5 Heterogeneous carbonylative Stille cross-coupling reaction of organostannanes[97]



Scheme 6 Suzuki–Miyaura cross-coupling using palladium acetate[98]



Scheme 3 Aryl imidazolylsulfonates as coupling partner in Pd-catalyzed Suzuki–Miyaura cross-coupling reaction[96]



Scheme 7 Synthesis of benzotriazole and benzimidazole-based biphenyl scaffolds[98]

Lee and co-workers[99] described the synthesis of heterogeneous silica gel-supported β-ketoiminatophosphane-Pd complex (Pd@SiO_2) (**21**) and examined its catalytic activity for Sonogashira, Suzuki, and Stille coupling reactions of a broad range of heteroaryl chlorides with different nucleophilic partners such as aryl boronic acids, organostannanes, and alkynes, providing yields up to 96, 94 and 96%, respectively (Scheme 8). The reaction was carried out an aqueous medium with 0.5 mol% catalyst loading, as mild conditions.

Milton and his research group[100] presented a microwave-accelerated synthesis of the [$PdCl_2$(L)] pre-catalysts synthesized from Na_2PdCl_4 and studied the reactivity of the Grignard cross-coupling by screening various ferrocene ligands such as dppf, dippf, dtbpf, and dtbdppf in a new solvent (Scheme 9). The solvent Me-THF has been gaining attention as a greener substitute to THF with no added reaction solvents. The authors performed the cross-coupling of Grignard reagents at 5 molar concentration in Me-THF with the correct matching of catalyst to substrate, and achieved good conversions in short times. This method significantly reduced the amount of solvent in both the Grignard synthesis and Grignard cross-coupling reactions as compared to other typical procedures based on THF. The reaction was found to be strongly dependent on the ligand structure, where 1,1-bis-diphenylphosphino-ferrocene and 1-di-*tert*-butyl-1-diphenylphosphino-ferrocene seemed to be the suitable ligands.

Peng et al.[101] disclosed stilbazo (stilbene-4,4-bis[(1-azo)-3,4-dihydroxybenzene]-2,2-disulfonic acid diammonium salt) (**24**) promoted, ligand-free Suzuki–Miyaura reaction in the presence of palladium catalyst in water at room temperature, which tolerated functional groups well and proceeded with high efficiency (Figure 1).

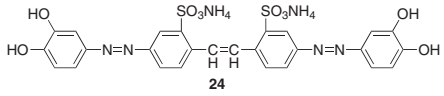


Figure 1 Stilbazo promoted Suzuki–Miyaura reaction[101]

Marziale and co-workers[102] tested various new palladacyclic catalysts (**25**–**27**) in aqueous Suzuki–Miyaura coupling conditions and concluded that catalyst **25** displayed high activities at room temperature for a broad range of products and afforded high yields (Figure 2). The isolated products were of high purity and could be separated by simple filtration.

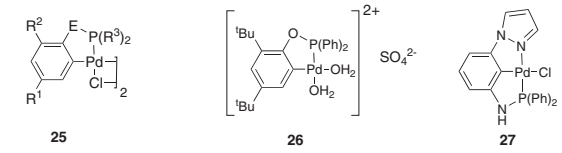
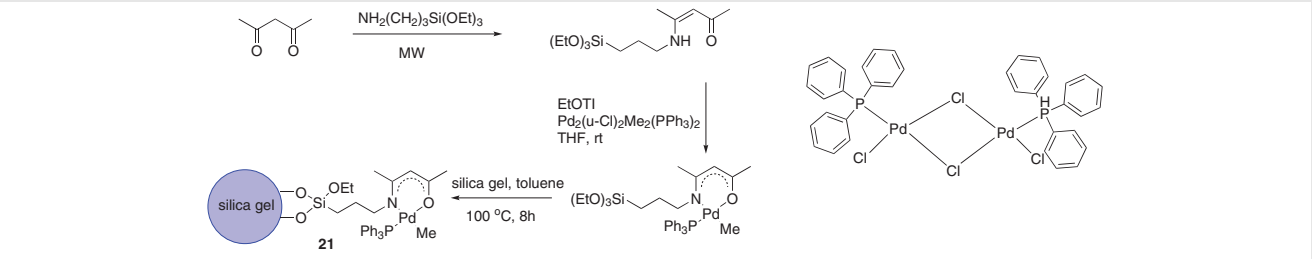
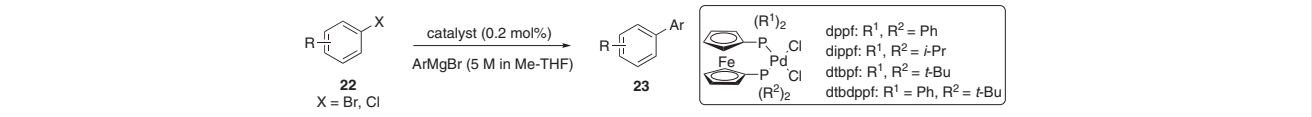


Figure 2 Palladacyclic catalysts tested in Suzuki–Miyaura coupling[102]

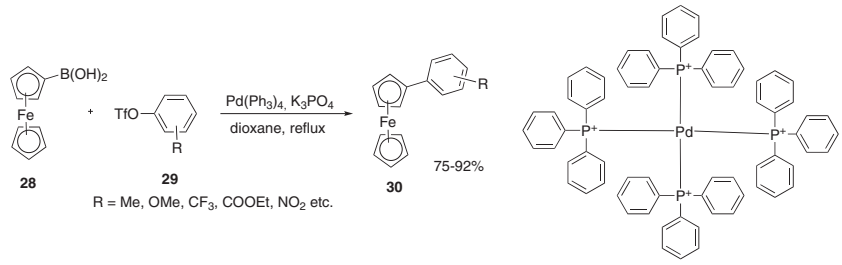
Long and co-workers[103] synthesized monosubstituted ferrocene derivatives **30** by using Suzuki cross-coupling reaction of ferroceneboronic acid (**28**) with a variety of aryl and vinyl triflates **29**. The reaction was carry out in the presence of $Pd(PPh_3)_4$ (0.025 equiv) and K_3PO_4 (2 equiv) in refluxing dioxane in excellent yields (Scheme 10). The electronic and steric effects were also observed for *ortho*-, *meta*-, and *para*-substituents of aryl triflates.



Scheme 8 Synthesis of silica gel-supported β-ketoiminatophosphane-Pd complex **21**[99]



Scheme 9 Cross-coupling of Grignard reagent with alkyl halides in Me-THF[100]



Scheme 10 Pd-catalyzed Suzuki cross-coupling reaction of ferroceneboronic acid[103]

For the first time, Chen and co-workers[104] used oxadisilole (**31**) as a coupling partner in the cross-coupling reaction with aryl halides catalyzed by palladium, affording 2-aryl naphthalenes **33** (Scheme 11). The reaction was carried out in the presence of tetrabutylammonium fluoride and this report offered a new path for the synthesis of functionalized acenes and related structures.

Molander and co-workers[105] devised a new avenue to introduce the amidomethyl functional group into substrates (Scheme 12). The amidomethyltrifluoroborates **35** were first synthesized and applied in the cross-coupling as coupling partners with a number of aryl and heteroaryl chlorides in a one-pot fashion.

In 2011, Liu and co-workers[106] achieved the ligand-free SM reaction of arylboronic acids with aryl bromides or nitrogen-based heteroaryl halides in aqueous DMF with K_2CO_3 and a catalytic amount of $PdCl_2$ at room temperature in moderate to excellent yields. This mild and simple method tolerated various functional groups and showed that the water/DMF ratio and presence of base are crucial in the reaction.

Islam et al.[107] disclosed a synthetic route involving novel polystyrene-assisted palladium(II) complex (PS-$[(C_6H_5CH{=}N)Pd(OAc)]_2$) (**37**) for phosphine-free and copper-free Sonogashira reactions using Et_3N as a base under aerobic conditions in aqueous (DMF/H_2O) medium (Figure 3).

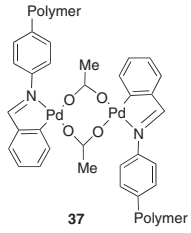
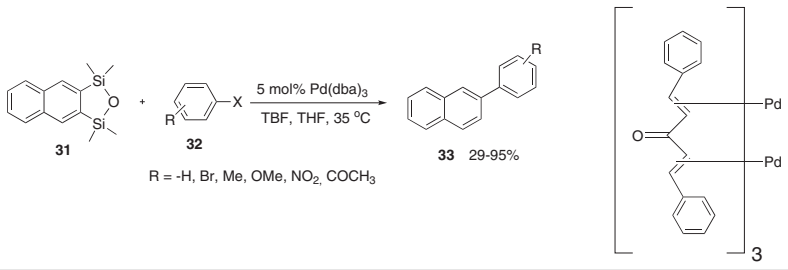
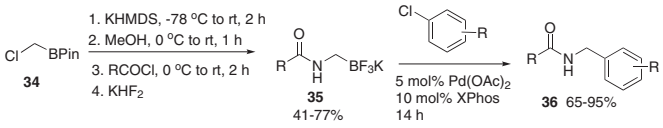


Figure 3 Polystyrene-supported palladium(II) complex catalyzed copper-free Sonogashira reactions[107]

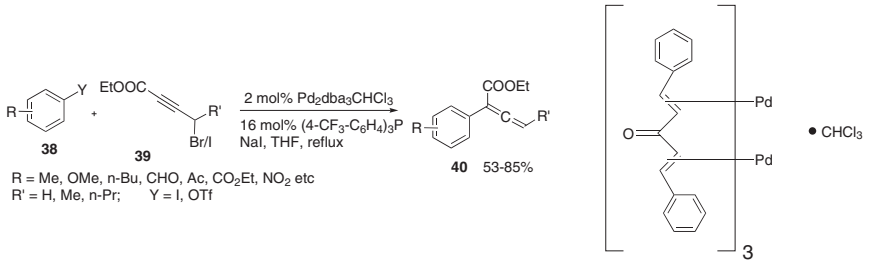
Lee and co-workers[108] published a synthetically valuable process to an introduce aryl group to the C2-position of 2,3-alkadienoates **40** via Pd-catalyzed selective allenyl cross-coupling reactions of an electron-withdrawing or electron-donating group containing aromatic iodides with organoindium reagents **39**; i.e, 2-aryl-2,3-alkadienoates and ethyl 4-bromo-2-alkynoates, generated under in situ conditions, with good yield (Scheme 13).



Scheme 11 Pd-catalyzed cross-coupling of oxadisilole with aryl halides[104]



Scheme 12 Synthesis of amidomethyltrifluoroborates and cross-coupling with aryl chlorides[105]



Scheme 13 Pd-catalyzed allenyl cross-coupling reaction[108]

Ali and co-workers[109] used palladium-catalyzed cross-coupling reactions (Buchwald–Hartwig, Sonogashira, and Suzuki–Miyaura) for the synthesis of a chain of peptides that were mono-functionalized with phthalocyanines (Pc) (**41**) at the N/C-terminal with moderate yields (Figure 4). The authors conjugated Pc with peptide moieties to help establish the selectivity for potential imaging probes for positron emission tomography and fluorescence for applications in the medical field.

A phosphine-free system developed by Modak et al.[110] demonstrated that a new functionalized mesoporous polymer (MPTAT-1) (**42**) developed by radical polymerization of 2,4,6-triallyloxy-1,3,5-triazine (TAT) in an aqueous medium

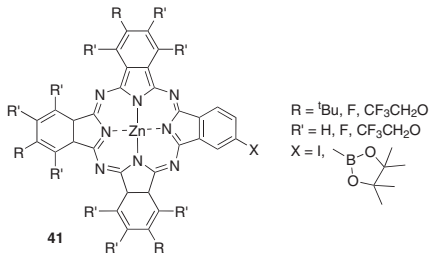


Figure 4 Structure of phthalocyanines (Pc) as reaction partner for various cross-coupling reactions[109]

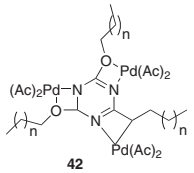
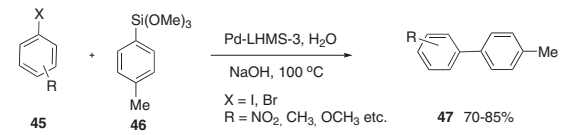


Figure 5 MPTAT-1 for Suzuki–Miyaura, Mizoroki–Heck, and Sonogashira cross-coupling reactions[110]

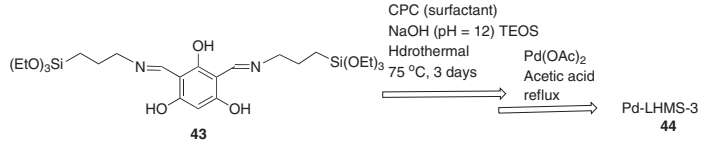
in the presence of an anionic surfactant (sodium dodecyl sulfate) as template, was an effective catalyst for several cross-coupling reactions such as Suzuki–Miyaura, Mizoroki–Heck, and Sonogashira (Figure 5). The template-free MPTAT-1 provides assistance in immobilizing Pd(II) and appears to be a very good catalytic scheme for eco-friendly conditions such as the use of water as reaction medium.

Later, an additional report was published by the same group[111] in which they prepared a Pd-grafted periodic mesoporous organosilica material (Pd-LHMS-3) (**44**) containing a phloroglucinol-diimine moiety within the pores (Scheme 14). This heterogeneous catalyst was investigated for its catalytic activity in Hiyama and Sonogashira couplings, and in cyanation reactions. The Hiyama cross-couplings executed by this protocol were fluoride-free and performed in water at alkaline pH conditions (Scheme 15). Similarly, copper-free Sonogashira cross-coupling reaction proceeded in water with a base such as hexamine. The catalyst promoted the cyanide-free cyanation of aryl halides with $K_4[Fe(CN)_6]$ as the cyanide source rather than using toxic KCN, NaCN or $Zn(CN)_2$. Excellent yields were presented for the synthesis of unsymmetrical biphenyls, di-substituted alkynes, and substituted benzonitriles under eco-friendly reaction conditions (Scheme 16).

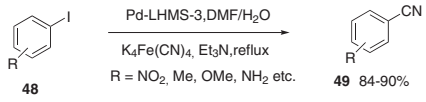
Susanto et al.[112] developed a thermally stable, fluorous oxime-based palladacycle, which promoted various microwave-assisted cross-couplings such as Suzuki–Miyaura, So-



Scheme 15 Hiyama cross-coupling reaction using Pd-LHMS-3[111]



Scheme 14 Synthetic pathway for the synthesis of Pd-LHMS-3[111]



Scheme 16 KCN-free cyanation using Pd-LHMS-3[111]

nogashira and Stille reactions in aqueous media (Figure 6). According to this report, Pd leaching was extremely low and **50** was reused five times with no significant loss of activity.

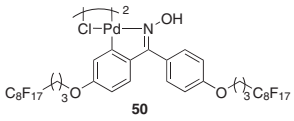


Figure 6 Fluorous oxime-based palladacycle for Suzuki, Sonogashira and Stille cross-couplings[112]

Mondal et al.[113] produced a catalytic system based on $PdCl_2$ and sodium sulfate generated in situ for ligand-free cross-coupling reaction in water at room temperature. They also produced a similar catalytic system based on $PdCl_2$ and sodium chloride or sodium acetate that was found to be equally effective in Suzuki–Miyaura cross-couplings. The Wang group[114] published a report on one-pot tandem Pd(II)-catalyzed Diels–Alder/cross-coupling reactions of 2-boron substituted dienes (Scheme 17). They prepared and characterized several new 2-boron substituted dienes and examined their reactivity in Diels–Alder reactions. The boron substituted cycloadducts thus formed were used in Suzuki cross-couplings.

The Liu research group[115] examined $Pd(OAc)_2/(i\text{-}Pr)_2NH/H_2O$ as a catalytic system for ligand-free and aerobic Suzuki reaction in water in the absence of any additive. It was demonstrated that the protocol tolerated a broad scope of aryl halides with either hydrophobic or hydrophilic groups. Further, the base was found to play a crucial role in this reaction. Keller et al.[116] synthesized a series of novel dendritic thiazolyl phosphine ligands and deployed them in palladium-catalyzed Suzuki couplings using $Pd(OAc)_2$ (Figure 7). The efficiency of the catalysts were compared with those of the corresponding triphenylphosphines; for example, in contrast to their triphenylphosphine counterparts, the thiazolyl phosphine-based catalytic arrangements (**54**). The reactions were also conducted either in water/THF mixtures or in pure water in the case of a single substrate.

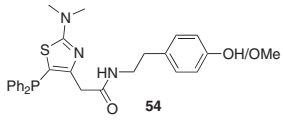


Figure 7 Dendritic thiazolyl phosphine ligands for Pd-catalyzed Suzuki cross-couplings[116]

Saha and co-workers[117] extended a new protocol for the development of eight-membered benzoxocinoquinoline (**55**) from quinolines by using a basic alumina-supported microwave-assisted intramolecular Heck reaction (Figure 8).

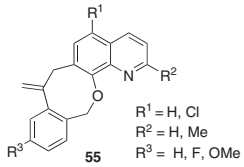
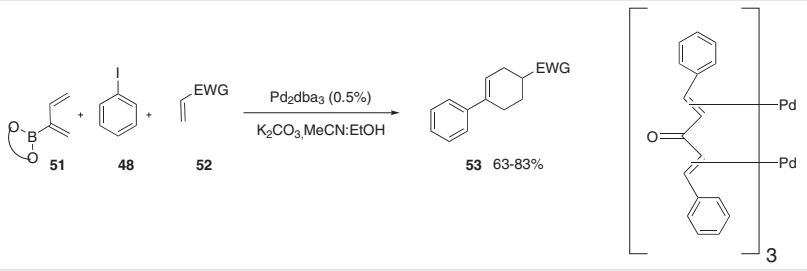


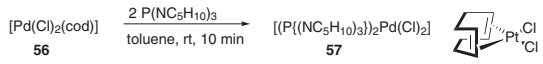
Figure 8 Benzoxocinoquinolines from Pd-catalyzed MW intramolecular Heck reaction[117]

In 2013, the Lu group[118] utilized a nonionic designer amphiphile, (TPGS-750-M) for $Pd(P(t\text{-}Bu)_3)_2$/DABCO catalyzed Stille couplings between a broad range of substrates (aryl and alkenyl halides) and organostannanes in water at room temperature. Oberholzera and Frech[119] designed a series of highly active, cheap, easily accessible, and air-stable dichloro-bis(aminophosphine) complexes of palladium of general formula $[(P\{(NC_5H_{10})_3\text{-}n\text{-}(C_6H_{11})_n\})_2Pd(Cl)_2]$ (**57**) (n = 0–3) for Heck cross-coupling reactions with a catalyst loading of 0.05 mol% in DMF at 100 °C using tetrabutylammonium bromide (Scheme 18). The active form of this catalyst was utilized (namely, nanoparticles) in the Heck reaction, which displayed an outstanding functional group tolerance towards aryl bromides containing fluoro, chloro, nitro, nitriles, aldehydes, ketones, esters, ethers, trifluoromethane



Scheme 17 Tandem Diels–Alder/Suzuki cross-coupling reaction[114]

groups, anilines, amides, phenols, and methylsulfanyl groups, and heterocyclic aryl bromides, such as pyridines and thiophene groups.



Scheme 18 Synthesis of a dichloro-bis(aminophosphine) complex of palladium[119]

The Yang group[120] developed a novel heterogeneous palladium catalyst by anchoring Pd(II) onto poly(undecylenic acid-co-*N*-isopropylacrylamide-co-potassium 4-acryloxyoylpyridine-2,6-dicarboxylate)-coated Fe_3O_4 (Fe_3O_4@PUNP) magnetic microgel (**58**) (Scheme 19).

Ding and co-workers[121] fabricated the Pd/bentonite catalyst by a simple impregnation method for Suzuki–Miyaura reaction. Clay is abundant, nontoxic, cheap, and a good support for the preparation of green catalysts. This methodology tolerated aryl bromides and iodides using several EDG and EWG such as -$COCH_3$, -OCH_3, -CH_3, -F, -NO_2,-CN, and -Cl in the coupling reaction with loading of catalyst Pd (0.06 mol%) in very low amount under ambient temperature. Yan and co-workers developed an effective palladium-catalyzed synthesis of arylethylene (mono-substituted alkenes) and diarylethylene (disubstituted alkenes) by coupling various aryl halides with olefins in the presence of easily available ligands (Scheme 20).[122]

Based on functionalized β-cyclodextrin, Zhang and co-workers[123] designed and prepared a novel water-soluble complex PdLn@β-CD (**62**) for Suzuki–Miyaura coupling reactions in aqueous medium (Figure 9). This catalyst was based on click-triazole-functionalized β-cyclodextrin and gave high turnover frequencies and turnover numbers of up to 4.9 × 108 h^{-1} and 9.9 × 10^8, respectively.

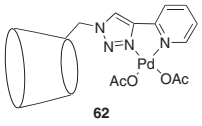
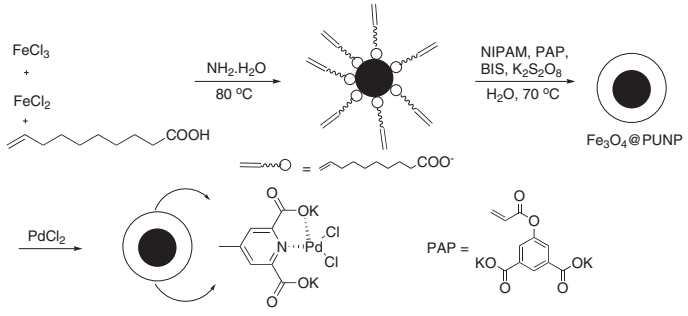
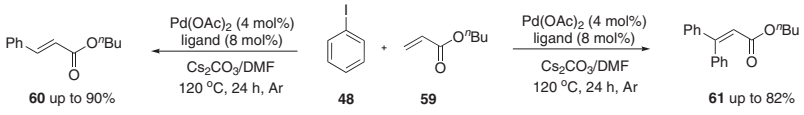


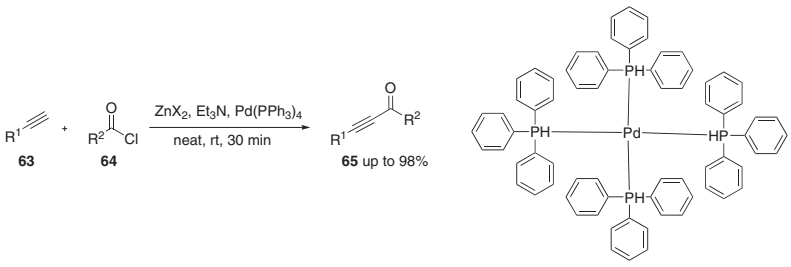
Figure 9 PdLn@β-CD for Suzuki–Miyaura coupling reaction[123]



Scheme 19 Synthesis of Fe_3O_4@PUNP-Pd catalyst[120]



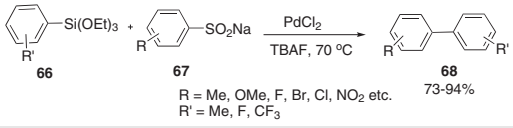
Scheme 20 Pd-catalyzed synthesis of arylethylene and diarylethylene[122]



Scheme 21 Pd-catalyzed synthesis of ynones[124]

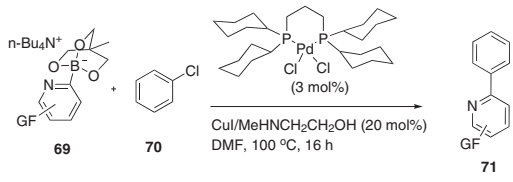
Yuan and co-workers[124] devised a mild and solvent-free procedure for the development of ynones by $Pd(PPh_3)_4$ catalyzed cross-coupling of in-situ-generated alkynylzinc derivatives with acyl chlorides in high yields (Scheme 21).

Cheng and co-workers[125] reported Pd-catalyzed Hiyama-type cross-couplings of various organosilanes (**66**) with arenesulfinates (**67**) in good to excellent yields at 70 °C under aerobic conditions (Scheme 22).



Scheme 22 Pd-catalyzed Hiyama-type cross-coupling of organosilanes with arenesulfinates[125]

Pd-catalyzed Suzuki–Miyura cross-coupling reactions were demonstrated by the Sakashita research group,[126] in which newly synthesized tetrabutylammonium 2-pyridyltriolborate salts **69** were allowed to couple with various challenging aryl/heteroaryl chlorides **70** to produce 2-arylpyridine derivatives **71** in the presence of $PdCl_2$dcpp (3 mol%) and CuI/$MeNHCH_2CH_2OH$ (20 mol%) without using bases, with excellent yields (Scheme 23). These tetrabutylammonium salts showed better reactivity than the correspondent lithium salts.



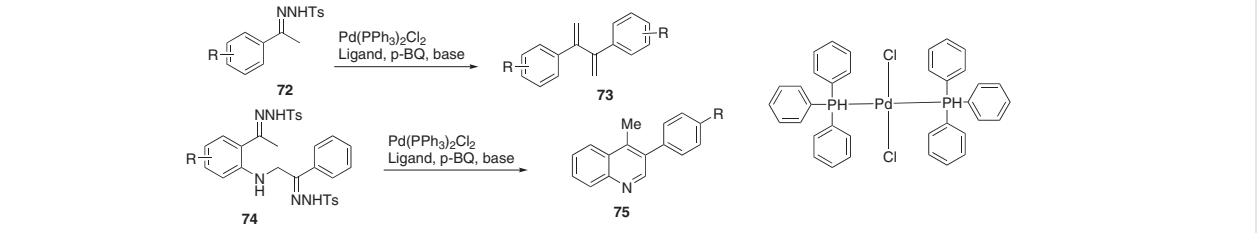
Scheme 23 Pd-catalyzed SM cross-coupling reactions of tetrabutylammonium 2-pyridyltriolborate salts as coupling partner[126]

Ojha et al.[127] achieved the challenging task of regioselective formation of highly branched dienes by a novel palladium-catalyzed selective coupling reaction of hydrazones **72** with *t*-BuOLi and *p*-benzoquinone (Scheme 24). They utilized carbene transfer reactions in the Pd(II) catalyzed coupling of hydrazones under oxidative conditions, which led to the formation of a Pd-bis-carbene complex with α-hydrogens finally affording branched dienes **73**. The reaction was versatile and compatible with a range of functional groups to open new avenues for synthesizing useful heterocyclic molecules **75**.

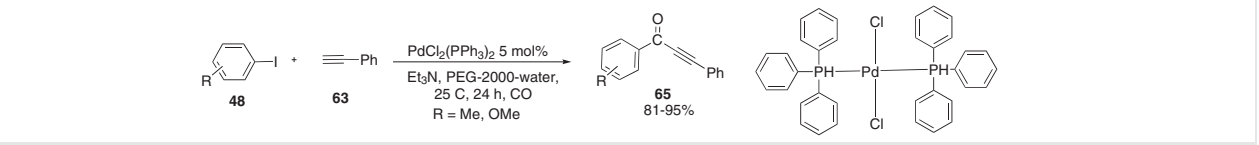
In 2014, the Liu research team[128] envisioned an efficient, recyclable, green and ligand-free method for the Suzuki coupling of aryl or heteroaryl halides in the presence of potassium aryltrifluoroborates with water, in air, using a $Pd(OAc)_2$-H_2O-PEG system to give the desired products in high reaction yields. The catalytic system was recycled eight times without appreciable loss in activity. Similarly, the water and PEG-2000 solvent mixture was utilized by the Zhao research group[129] in which they described the carbonylative Sonogashira coupling reaction of terminal alkynes with aryl iodides in the presence of $PdCl_2(PPh_3)_2$ and Et_3N as a base under an atmospheric pressure of CO at 25 °C, giving a scope of alkynyl ketones with satisfactory yields (Scheme 25). This protocol could be effortlessly extended to the synthesis of 2-substituted flavones from *o*-iodophenol and terminal alkynes.

The Nehra group[130] prepared an efficient ionic-liquid-tagged Schiff base palladium complex (**76**) that was stable in air and was water-soluble (Figure 10). This complex showed catalytic activity and was investigated for Heck and Suzuki cross-coupling reactions in aqueous media. The protocol was found to be very effective for more challenging substrates such as chlorides.

Li et al.[131] reported stereospecific cross-coupling between aryl chlorides and unactivated secondary alkylboron



Scheme 24 Pd-catalyzed coupling of hydrazones[127]



Scheme 25 Carbonylative Sonogashira coupling reaction[129]

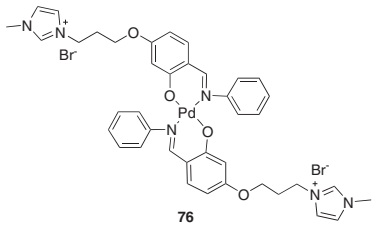
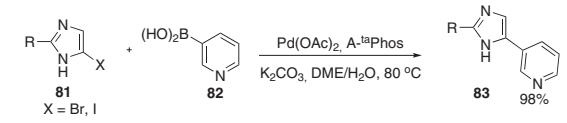


Figure 10 Ionic-liquid-tagged Schiff base palladium complex[130]

nucleophiles under Pd catalysis. Secondary alkyltrifluoroborates and secondary alkylboronic acids were tolerated in this protocol without noteworthy isomerization of the alkyl nucleophile. In this cross-coupling process, optically active secondary alkyltrifluoroborate reagents underwent stereospecific inversion of configuration. This protocol could be utilized in the construction of optically active drugs from optically active alkylboron compounds. Nadaf et al.[132] described the palladium-catalyzed Suzuki–Miyaura cross-coupling reactions between newly synthesized potassium *N*-methyltrifluoroborate isoindolin-1-one (**77**) and aryl and heteroaryl chlorides to prepare libraries of substituted *N*-benzyl isoindolin-1-ones **78** (Scheme 26).

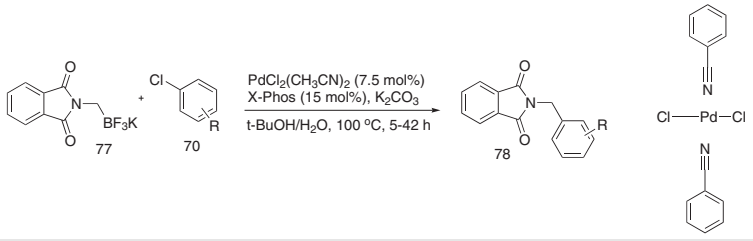
In cross-coupling reactions, one of the limitations is that azobenzenes act as electrophiles, whereby metalation by halogen-metal exchange causes reduction of the azo group yielding hydrazine derivatives in place of the desired metallated azobenzenes. Strueben[133] provided a solution to this problem by developing a mild method to prepare mono- and distannylated azobenzenes (**79**), which were used as nucleophilic partners in Pd-catalyzed Stille cross-coupling reactions with electron-deficient and electron-rich aryl bromides, resulting in the formation of the cross-coupled products **80** in yields as high as 70 to 93% (Scheme 27).

Sun et al.[134] reported a three-step synthesis of 4,8-azaboranaphthalene (ABN) on a gram scale and showed that the reaction tolerated a variety of functional groups and cross-coupling partners in various Sonogashira, palladium-catalyzed Suzuki, and Heck cross-coupling reactions. One of the advantages of the coupled product bearing an ABN motif was that it showed a fluorescence response toward Cd(II) and Zn(II) ions. Thus, this protocol is very significant in designing various fluorescent chemosensors. The Tan group[135] published a Suzuki–Miyaura cross-coupling reaction catalyzed by palladium of unprotected haloimidazoles (**81**) with various aryl- and heteroarylboronic acids providing a wide array of functionalized imidazoles (**83**) (Scheme 28). The total synthesis of nortopsentin D was also demonstrated by the group utilizing this method.

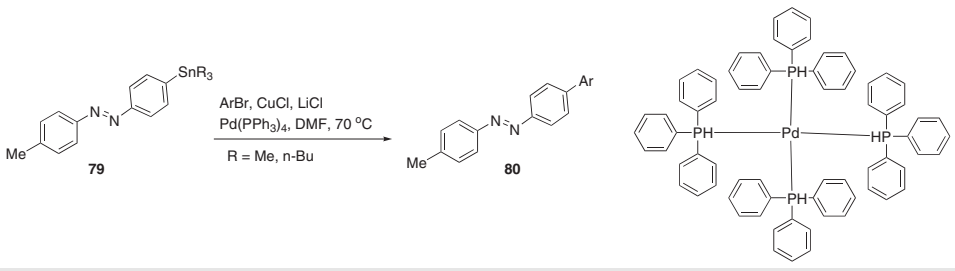


Scheme 28 Pd-catalyzed SM cross-coupling reaction of unprotected haloimidazoles[135]

In 2015, Shen et al.[136] synthesized D-glucosamine-derived triazole@palladium catalyst (**84**) via a suitable route in high reaction yields, and its catalytic activity was studied in Heck cross-coupling reactions between olefins (Figure 11). Additionally, the easy synthesis of marketed antitumor drug Axitinib (**85**) was also demonstrated by the group utilizing this protocol.



Scheme 26 Synthesis of substituted *N*-benzyl isoindolin-1-ones[132]



Scheme 27 Pd-catalyzed Stille cross-coupling reactions of monostannylated azobenzenes[133]

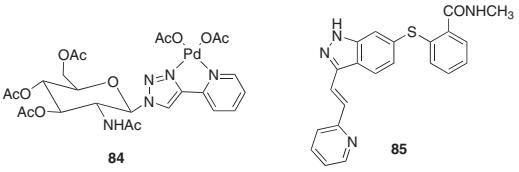
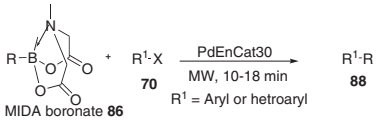


Figure 11 D-Glucosamine-derived triazole@palladium catalyst **84** and axitinib **85**[136]

An improved and highly efficient procedure was reported by da Silva and co-workers,[137] who described ligand-free microwave-enhanced Suzuki cross-coupling reaction of (het)aryl halides and (hetaryl, allyl)arylboronic acid N-methyl-iminodiacetic acid (MIDA) ester (**86**) for the synthesis of biaryls, bipyridyls, thienylpyridine, and allylphenols using polyurea microencapsulated Pd catalyst (Pd EnCat 30) in excellent yields in just 10–18 min (Scheme 29). This method did not utilize any phosphine-ligand and was performed in an aqueous medium.



Scheme 29 MW-assisted SM cross-coupling reaction of (het)aryl halides and MIDA ester[137]

Boruah and co-workers[138] demonstrated a highly efficient, economical alternative and eco-friendly procedure for Suzuki–Miyaura cross-coupling reactions catalyzed by palladium acetate. The reactions were carried out in neat 'water extract of banana' at room temperature in the air within 5–90 min in the absence of any ligand, external base, organic medium, and external promoters like additives. Strappaveccia and co-workers[139] investigated the synthesis of stilbines, cinnamate esters and acids by using of GVL in the Pd-catalyzed Heck reaction. Moreover, poly(phenylenevinylene) (PPV) semiconductors were also prepared using this protocol in high yields with very low amount of Pd-content. In a later investigation by this group,[140] they presented γ-valerolactone (GVL) as a non-toxic, biodegradable, biomass-derived dipolar aprotic solvents like DMF or NMP for the Sonogashira cross-coupling reaction using DABCO as a base, affording the desired products in 62–96% yields. Because of the biomass-derived reactants, this method offers an eco-friendly approach leading to higher sustainability as well as high chemical efficiency.

In 2016, Khana et al.[141] prepared an ionic Pd(II) complex stabilized by a water-soluble pyridinium-modified β-cyclodextrin, affording the N-octyl-pyridine-2-amine backbone (Pd(II)@Pyr:β-CD) (**89**) and explored the activity of catalyst in Heck cross-coupling and Suzuki–Miyaura coupling reactions in eco-friendly aqueous medium, which provided high yields of the coupled product (Figure 12). Aryl chlorides were also efficiently coupled with phenylboronic acid/styrene using this procedure.

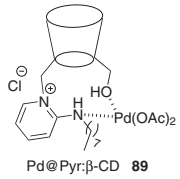
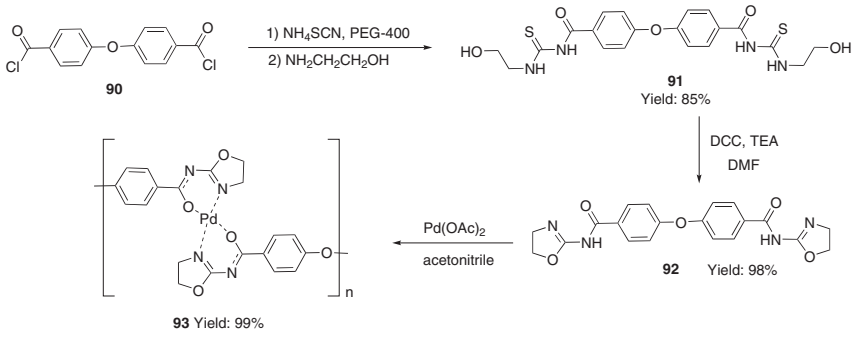


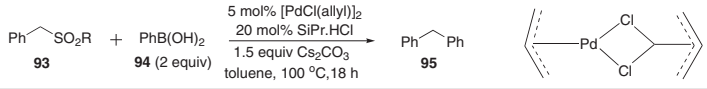
Figure 12 Pyridinium-modified β-cyclodextrin bearing the N-octyl-pyridine-2-amine[141]

The Jadhav research group[142] proposed a ligand-free Suzuki–Miyaura reaction and base-free Heck reactions to synthesize a variety of biaryls, acrylates, and prochiral ketones under mild reaction conditions using Pd supported on activated carbon (Pd/C) in an aqueous hydrotropic solution. A hydrotrope was used as a precious, green reaction medium for the first time. Wang and co-workers[143] developed a novel oxazoline-based palladium microsphere complex (**93**) by the self-assembly of the bisoxazoline (**92**) and $Pd(OAc)_2$ (Scheme 30). This solid microsphere catalyst was explored for phosphine-free Suzuki–Miyaura cross-coupling reactions in aqueous media.

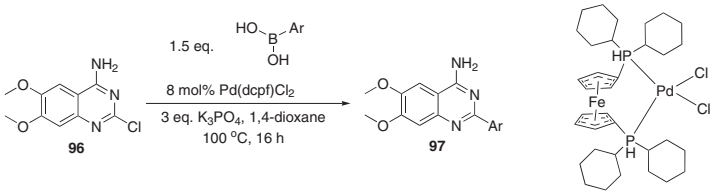
In 2017, Nambo and co-workers[144] described the Pd-catalyzed Suzuki–Miyaura cross-coupling reactions of fluorinated sulfone derivatives as effective electrophiles (Scheme 31). C–SO_2 bonds were activated by introducing an



Scheme 30 Synthesis of oxazoline-based palladium microsphere complex **93**[143]



Scheme 31 Pd-catalyzed SM cross-coupling reactions of fluorinated sulfone[144]

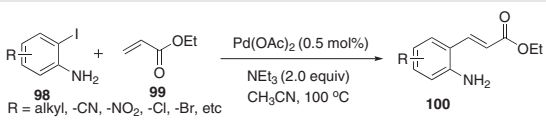


Scheme 32 Synthesis of 4-aminoquinazoline biaryl compounds[145]

EWG on the aryl ring of the sulfones under Pd-catalysis, leading to a diversity of multiple arylated products in satisfactory yields.

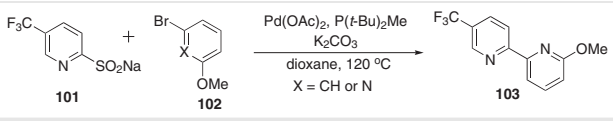
Pulipati and co-workers[145] proposed a vigorous approach for the synthesis of 4-aminoquinazoline biaryl compounds from arylboronic acids and quinazoline containing an unprotected NH_2 group (**96**) via Suzuki–Miyaura coupling reaction using Pd(dcpf)Cl_2 (Scheme 32). The synthesized compounds were also assessed for antimicrobial and antifungal biological activity.

Chen and co-workers[146] developed Pd-catalyzed ligand-free Heck reaction between 2-iodoanilines (**98**) and acrylate (**99**) in CH_3CN using Pd$(OAc)_2$ (5.0 mol%) as catalyst and NEt_3 as a base to afford 2-alkenylanilines (**100**) in high yields of up to 93% (Scheme 33).

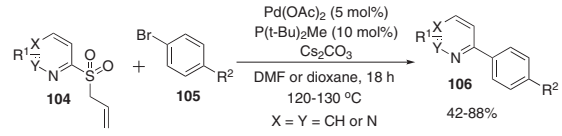


Scheme 33 Pd-catalyzed ligand-free Heck reaction[146]

Jadhav et al.[147] utilized a Pd$(PPh_3)_4$/Et_3N/H_2O/98 °C catalyst system for the Mizoroki–Heck coupling reaction carried out in the absence of any additives under aerobic conditions with TOF of 12 to 14 h^{-1} in a very short reaction time. This procedure was applicable for a broad range of electron-donating and electron-withdrawing aryl chlorides and bromides. The Clavé research group[148] developed a bio-based plant-derived EcoPd from the roots of *Eichhornia crassipes* for the Suzuki cross-coupling of heteroaryl compounds for the synthesis of a broad range of heterocyclic–heterocyclic biaryl and heterocyclic biaryl compounds with a small amount of catalyst. The reaction promoted the Suzuki cross-coupling without ligands or additives. In 2018, Markovic et al.[149] described a Pd-catalyzed coupling reaction between heterocyclic sulfinates (**101**) and aryl or heteroaryl halides (**102**), affording high yields of the corresponding biaryls (**103**) (Scheme 34). Furthermore, the heterocyclic allylsulfones (**104**) can function as ideal sulfinate reagents and, when reacted with aryl halide, deallylation could be restricted, leading to efficient desulfinylative cross-coupling under palladium(0) catalysis (**106**) (Scheme 35). Additionally, the authors also prepared pharmaceutical agents etoricoxib and crizotinib using allyl heteroaryl sulfone coupling partners.



Scheme 34 Pd-catalyzed cross-coupling reaction of heterocyclic sulfinates[149]



Scheme 35 Pd-catalyzed desulfinylative cross-coupling reaction[149]

Sharma and co-workers[150] synthesized palladium(II) complex of hemilabile N–O ligand (picolinate) (**107**) for Mizoroki–Heck couplings with high TOF up to >10,000 h^{-1} in just 15 min with high selectivity of >99% to the desired products (Figure 13).

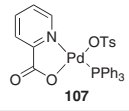
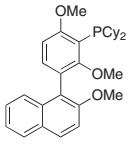


Figure 13 Palladium(II) complex of hemilabile N–O ligand (picolinate)[150]

Another exceptional contribution was reported by Landstrom et al.[151] who synthesized a new biaryl phosphine-containing ligand (EvanPhos) (**108**) in only two steps (Figure 14). The complexation was done with either Pd$(OAc)_2$ or $(CH_3CN)PdCl_2$ converting into a highly active Pd(0) species, which served as a precatalyst for Suzuki–

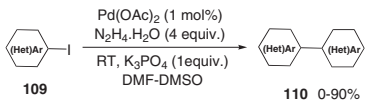
Miyaura cross-couplings of functionalized reaction partners in either water containing nano micelles or uncommon solvent EtOAc. This catalytic system was very effective even at low amounts of 0.05–0.5 mol%. The high rate of reaction was further improved when the reaction was performed in aqueous micellar media instead EtOAc.



108

Figure 14 Biaryl phosphine-containing ligand (EvanPhos)[151]

In 2019, Gong and co-workers[152] developed an Ullmann biaryl synthesis using $Pd(OAc)_2$ and $N_2H_4{\cdot}H_2O$ as the reducing reagent for the coupling of both electron-deficient as well as electron-rich aryl or heteroaryl iodides **109** leading to a variety of biaryls **110** at room temperature in good to excellent yields. The *in-situ* generated palladium nanoparticles were found to be active catalysts. The advantages of this protocol were cheap reducing agent, cost-effectiveness, and no need for metal reductants (Scheme 36).



109 110 0-90%

Scheme 36 Pd-catalyzed Ullmann biaryl synthesis[152]

Chehrouri and co-workers[153] demonstrated the synthesis of palladium complexes derived from 3-pentadecyl-1,3,4-oxadiazole-2(3*H*)-thiones **111**, **112** or 4-amino-3-pentadecyl-3*H*-1,2,4-triazole-3-thiones **113**, **114** (Figure 15). The complexes thus obtained were explored for their catalytic activity in Mizoroki–Heck and Tsuji–Trost reactions, providing very high chemical yield.

The Balfour research group[154] presented the synthesis of a library of 2,6-disubstituted-azaindoles **116**, based on a tandem Sonogashira coupling/5-*endo*-*dig*/Sonogashira coupling sequence (Scheme 37) This protocol tolerated alkynes containing alcohols, aliphatic chains, and aromatic substituents.

Inspired by advantages of heterogeneous catalysts such as high stability, easy separation from reaction mixture, and good recyclability over homogeneous catalysts Liu et

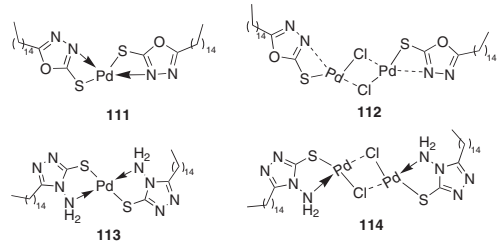
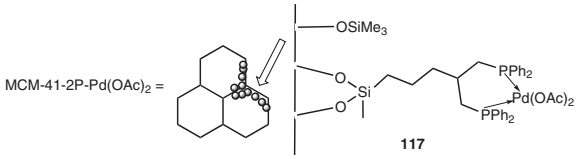


Figure 15 Palladium complexes developed by Chehrouri et al.[153]

al.[155] published the synthesis of three pyridine-functionalized *N*-heterocyclic carbene-Pd complexes (HCP-Pd) using a simple external cross-linking reaction. In each complex, Pd was immobilized on the hypercrosslinked polymer (HCP) via the formation of a six-membered ring by pyridine, bidentate ligands of NHC, and Pd^{2+}. The newly synthesized catalysts were very effective for the coupling reaction in an aqueous medium under mild conditions. The microporous structure of the support ensured the high dispersion of palladium active sites. Of the three complexes, the recovery and reusability were easier for the HCP-Pd-I catalyst compared to the HCP-Pd-II and III catalysts.

In 2020, Zhou et al.[156] developed a heterogeneous Pd-catalyzed carbonylative Sonogashira coupling by using an MCM-41-2P-$Pd(OAc)_2$ (**117**) as a catalyst (Scheme 38). In 2021, Bangar and co-workers[157] used mono- and bidentate chelating oximes as ligands for the Pd catalysis of Suzuki and Heck coupling.

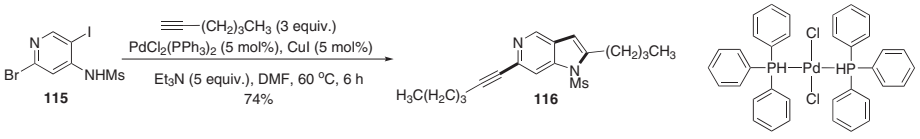


117

Scheme 38 Design of MCM-41-2P-$Pd(OAc)_2$[156]

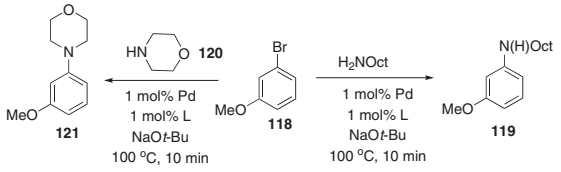
2.2 C–N Cross-Coupling Reaction

In 2010, Fors and co-workers[158] disclosed an alternative approach to catalyst advancement, in which they prepared a multiligand-based Pd catalyst (Scheme 39). The designed catalyst was then allowed to catalyze C–N cross-coupling reactions. This catalytic system exhibited the same catalyst activity and substrate scope.

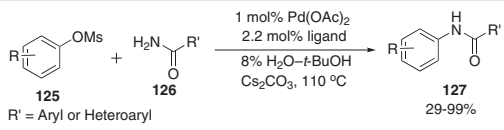


115 116

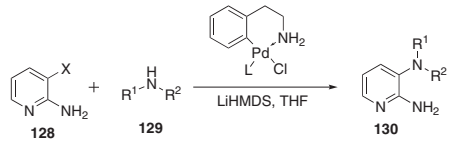
Scheme 37 Pd-catalyzed tandem Sonogashira coupling/5-*endo*-*dig*/Sonogashira coupling sequence[154]



Scheme 39 C–N cross-coupling reactions using multiligand-based Pd catalyst[158]



Scheme 40 Pd-catalyzed coupling of amides and aryl mesylates[159]



Scheme 41 Pd-catalyzed C–N cross-coupling of unprotected 3-halo-2-aminopyridines[160]

Similar to the Fors report, another group led by Dooleweerdt et al.[159] also described a palladium catalyst based on biaryl phosphine ligands (**120–124**) (Figure 16) that allowed the coupling of amides and an array of aryl/heteroaryl mesylates (**125**) (electron-rich, -neutral, and -deficient) to afford the corresponding N-aryl amides **127** in high yields (Scheme 40). Benzamides and aliphatic and heterocyclic amides were also investigated as excellent coupling partners in this protocol.

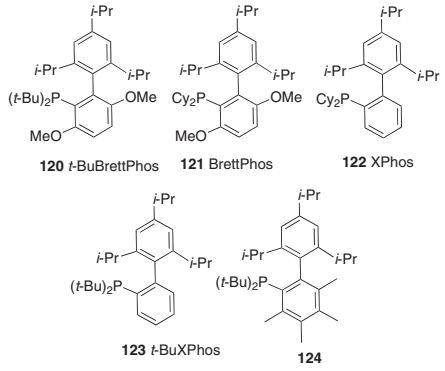
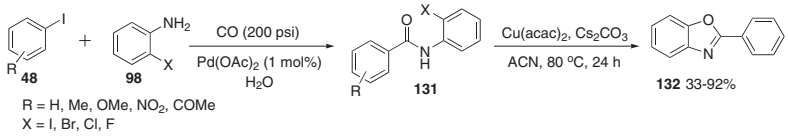


Figure 16 Biarylphosphine ligands used by Dooleweerdt et al.[159]

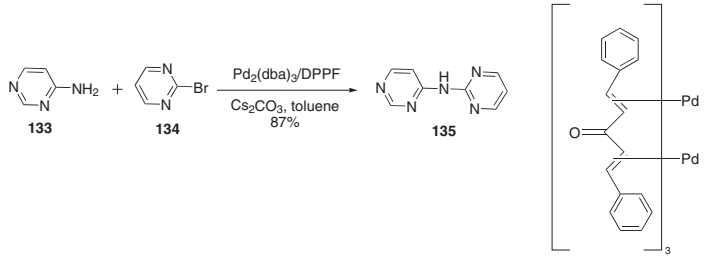
In 2011, Perez and co-workers[160] presented an unprecedented approach to Pd-catalyzed C–N cross-coupling of unprotected 3-halo-2-aminopyridines (**128**) with an array of primary and secondary amines yielding N3-substituted-2,3-diaminopyridines **130** (Scheme 41). The reaction was performed with BrettPhos- and RuPhos-precatalysts in combination with LiHMDS for this C–N cross-coupling reaction.

Tambade et al.[161] disclosed a phosphine-free $Pd(OAc)_2$ catalyzed procedure for aminocarbonylation or carbonylative cross-coupling that enabled the coupling of a wide range of substituted aryl iodide with *ortho*-haloaniline to form *ortho*-haloanilide (**131**) in water, affording good yields (Scheme 42). Further, *ortho*-haloanilides **131** underwent cyclization for the synthesis of benzoxazoles **132** using $Cu(acac)_2$ catalyst.

In 2013, Zhang and co-workers[162] devised a Pd-catalyzed method for the cross-coupling of heteroaryl halides and electron-deficient heteroaromatic amines in the presence of $Pd_2(dba)_3$ as a catalyst, 1,10-bis(diphenylphosphi-



Scheme 42 Pd-catalyzed aminocarbonylation cross-coupling[161]



Scheme 43 Pd-catalyzed C–N cross-coupling[162]

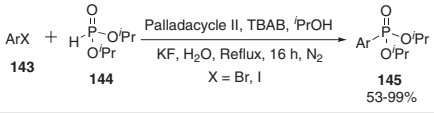
no)ferrocene (DPPF) as ligand, and Cs_2CO_3 as a base (Scheme 43). This methodology allowed the coupling of several rarely reported electron-deficient heteroaromatic amines in good yields.

In 2014, Wagner et al.[163] described a versatile green catalytic system ([(cinnamyl)PdCl]$_2$/*t*-BuXPhos) (**123**) for coupling of arylbromides or chlorides with a wide range of amines, carbamates, ureas, and amides under Buchwald–Hartwig cross-coupling reaction conditions in an aqueous micellar medium. The procedure was functional-group tolerant; for example, for esters and halides the reactions were carried out at 30–50 °C providing the target compounds in good to excellent yields (Scheme 44). Compared to the previously reported Takasago's catalyst system (cBRIDP ligand in combination with [(allyl)PdCl]$_2$), this catalytic system was found to be much more efficient for Buchwald–Hartwig reactions with benzamide derivatives or aliphatic primary amines. No racemization was experienced in this method when a substrate with a chiral center was used.
In 2020, Fan et al.[164] described the development of a Pd-catalyzed decarbonylative C–N coupling under a nitrogen atmosphere (Scheme 45).

In 2021, Feng and co-workers[165] achieved the direct cross-coupling of NH-sulfoximines through N-benzylation via visible-light photocatalysis. Patel et al.[166] described Pd-mediated simultaneous CH–CX and CH–NH bond activation followed by intramolecular cyclization reaction to form quinolin-fused benzo[*d*]azeto[1,2-*a*]benzimidazole analogs.

2.3 C–P Cross-Coupling Reaction

In 2013, Xu and co-workers[167] developed palladacycle-catalyzed phosphonation of aryl halides with diisopropyl H-phosphonate (**144**) using cyclopalladated ferrocenylimines **141**–**142** (Figure 17) with bulky phosphine ligands of X-Phos in water affording the phosphonated products **145** in excellent reaction yields (Scheme 46). The inactive electron-rich and electron-neutral aryl chlorides reacted well in this process. The weak base KF was enough for the activation of C–Cl bond instead of strong bases such as NaO^{t}Bu or KO^{t}Bu.



Scheme 46 Palladacycle-catalyzed phosphonation reaction[167]

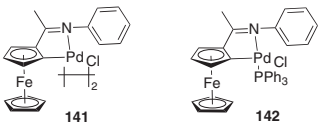
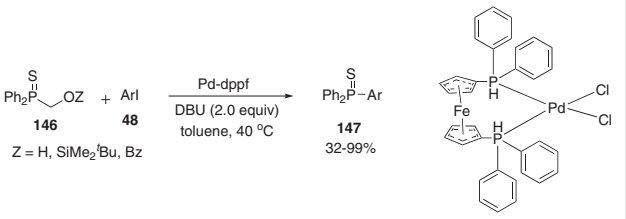
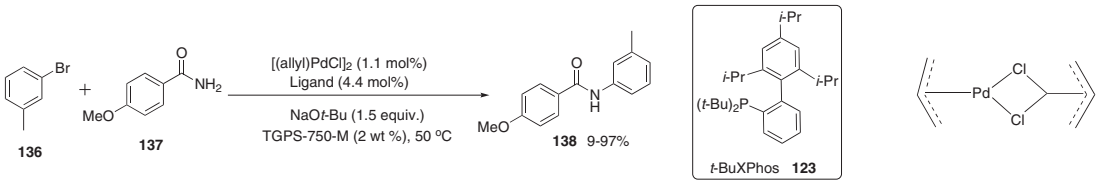


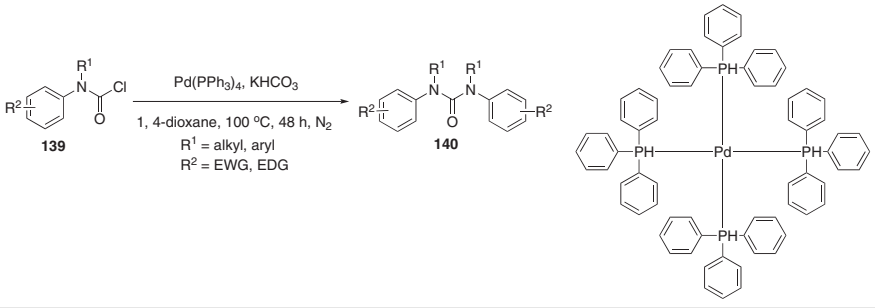
Figure 17 Cyclopalladated ferrocenylimine ligands[167]



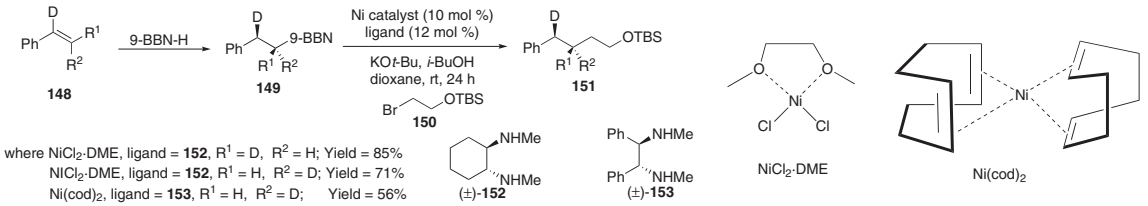
Scheme 47 Pd-catalyzed deformylative C–P cross-couplings[168]



Scheme 44 Pd-catalyzed Buchwald–Hartwig cross-coupling reaction[163]

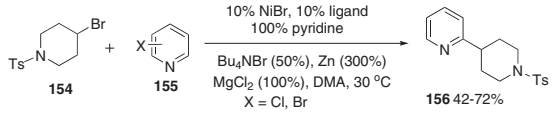


Scheme 45 Pd-catalyzed decarbonylative C–N coupling[164]



Scheme 48 Ni-catalyzed SM cross-coupling reaction[169]

The Hayashi group[168] achieved the syntheses of tertiary phosphine derivatives by Pd-catalyzed deformylative C–P cross-couplings of hydroxymethylphosphine derivatives **146** (Scheme 47). Triarylphosphine synthesis was also achieved by sequential triple couplings using this protocol.



Scheme 49 Ni-catalyzed alkylation of halo-pyridines[170]

3 Ni-Catalyzed Cross-Coupling Reactions

We will discuss in this section the use of nickel catalysis in various cross-coupling reactions and is generally presented in chronological order.

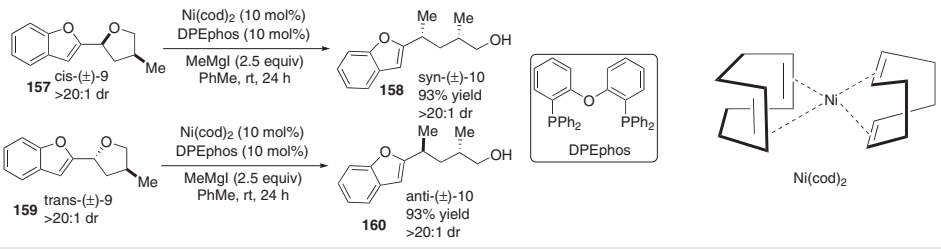
3.1 C–C Cross-Coupling Reaction

In 2011, the Taylor group[169] introduced deuterium-labeled alkylborane reagents **149**, which were allowed to undergo nickel-catalyzed Suzuki cross-coupling reactions in the presence of diamine ligands **152** and **153** (Scheme 48), resulting in transmetalation from boron to nickel with retention of configuration.

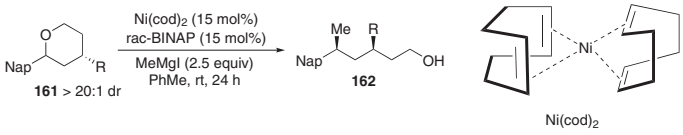
In 2014, Liu et al.[170] disclosed a method for Ni-catalyzed cross-electrophile coupling of secondary alkyl bromides **154** with halogenated pyridines **155** using zinc as a reductant, yielding different alkyl-substituted pyridines **156** in moderate to excellent yields (Scheme 49). This report provided a solution to the unreported instances in the previous literature on alkylation of halo-pyridines.

Tollefson's research group[171] reported the stereospecific Ni-catalyzed ring-opening cross-couplings of *O*-heterocycles such as aryl-substituted tetrahydropyrans, tetrahydrofurans, and lactones to give acyclic alcohols and carboxylic acids (Scheme 50). This method paved the way for the stereochemical synthesis of acyclic polyketide analogs. The authors showed that Ni-catalyzed Kumada-type coupling of aryl-substituted tetrahydropyrans and tetrahydrofurans proceeded with a variety of Grignard reagents to provide acyclic alcohols with excellent diastereoselectivity (Scheme 51).

One year later, in 2015, the Tollefson research group[172] presented further research findings on the Ni-catalyzed Kumada, Negishi, and Suzuki cross-coupling reactions of benzylic ethers such as methyl ethers, tetrahydrofurans, tetrahydropyrans, esters, and lactones as one of the reaction



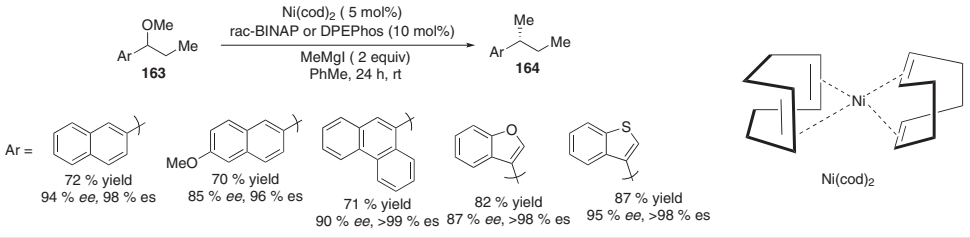
Scheme 50 Ni-catalyzed stereospecific ring-opening cross-couplings of O-heterocycles[171]



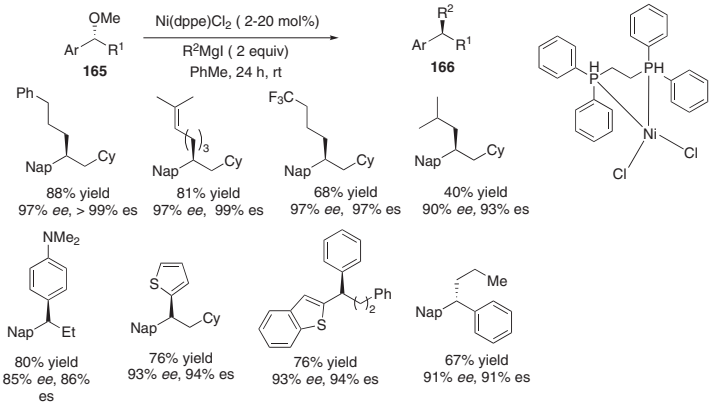
Scheme 51 Ni-catalyzed stereospecific ring-opening cross-couplings of tetrahydropyrans[171]

partners (Scheme 52). Several Grignard reagents such as aryl, methyl, and *n*-alkyl Grignard reagents were engaged in Kumada coupling reactions. Specifically, with methylmagnesium iodide as coupling partner, the ligands DPEphos or rac-BINAP afforded the highest reaction yield and stereospecificity (Scheme 53). The functional group tolerance was described in Negishi cross-coupling reactions using dimethylzinc. Similarly, Suzuki reactions using arylboronic esters were also reported, with different stereochemical outcomes employing different achiral ligands giving opposite enantiomers of the product. Using *N*-heterocyclic carbene ligand (SIMes) in Suzuki reaction caused inversion in the product, and use of the electron-rich phosphine PCy_3 gave retention with stereospecificity (Scheme 54). Various pharmacophores units such as 1,1-diarylalkane and 2-arylalkane have been synthesized by using these cross-coupling reactions.

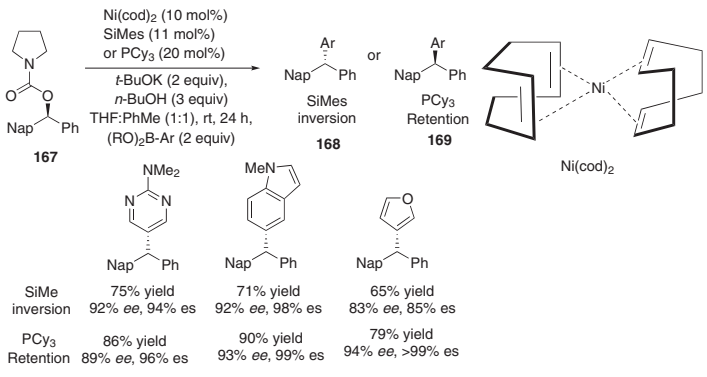
Dawson and Jarvo[173] published a similar kind of approach as previously described by the Tollefson group (in which they also have been the members), whereby they emphasized the development of stereospecific reactions for use in the field of pharmaceutical chemistry (Scheme 55). They reported a highly stereospecific gram-scale Kumada cross-coupling reaction with inversion at the benzylic position using a sustainable and inexpensive nickel catalyst.



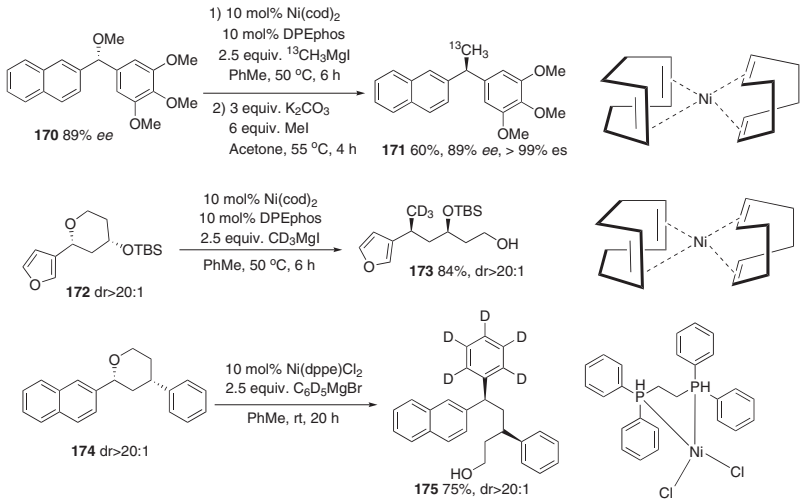
Scheme 52 Ni-catalyzed Kumada cross-coupling reaction[172]



Scheme 53 Ni-catalyzed Kumada cross-coupling reaction[172]

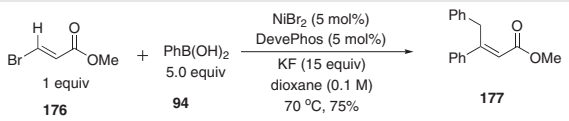


Scheme 54 Ni-catalyzed Suzuki cross-coupling reaction[172]



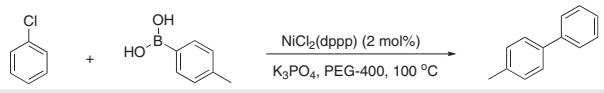
Scheme 55 Kumada cross-couplings of isotopically labeled Grignard reagents[173]

An unprecedented report was published by the Funicello research group[174] in which they carried out the Ni-catalyzed C–Br/C–H double phenylation of methyl 4-bromocrotonate (**176**) affording a useful bis-arylated synthon through a cross-coupling reaction (Scheme 56).



Scheme 56 Ni-catalyzed C–Br/C–H double phenylation[174]

In 2019, Liao and co-workers[175] developed a catalytic system consisting of 2 mol% $NiCl_2$(dppp) in PEG-400 for Suzuki–Miyaura coupling reaction at 100 °C using a base (i.e K_3PO_4), providing a range of biaryls with high reaction yields (Scheme 57). The $NiCl_2$(dppp)/PEG-400 catalytic system could be simply recycled and re-applied up to five times without significant loss of activity. The main advantages of this protocol lie in the fact that it avoids the use of toxic and easily volatile toluene or dioxane as solvent, and solves the critical problem of nickel catalyst reuse.



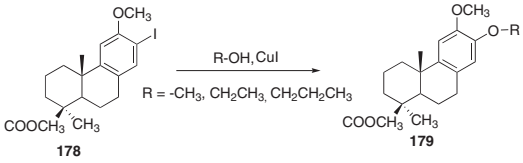
Scheme 57 Ni-catalyzed Suzuki–Miyaura coupling reaction[175]

4 Cu-Catalyzed Cross-Coupling Reactions

We will discuss in this section the use of copper catalysis in various cross-coupling reactions and is generally presented in chronological order.

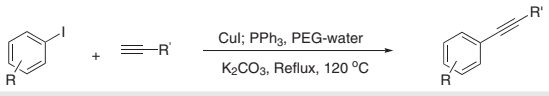
4.1 C–C Cross-Coupling Reactions

In 2010, Yalavarty and co-workers[176] found a new copper-catalyzed method of synthesizing podocarpic acid ether derivatives through the one-step cross-coupling reaction of methyl 13-iodo-*O*-methylpodocarpate (**178**) with alcohols in excellent yields (Scheme 58). Copper iodide was utilized as an inexpensive catalyst to achieve this transformation.



Scheme 58 Cu-catalyzed synthesis of podocarpic acid ether derivatives[176]

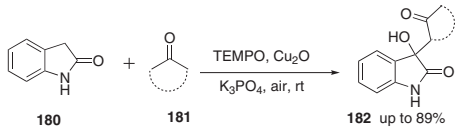
In 2011, the Chen research group[177] established an efficient CuI/PPh_3/PEG-H_2O catalytic system for Sonogashira coupling of electron-deficient or electron-rich aryl iodides with terminal acetylenes in water-polyethylene glycol under microwave irradiation or reflux to provide good to excellent yields (Scheme 59).



Scheme 59 Cu-catalyzed Sonogashira coupling reaction[177]

In 2015, Wang et al.[178] developed easy and efficient protocol that allowed the synthesis of a variety of 3-(2-oxoalkyl)-3-hydroxyoxindoles **182** through tandem oxidative cross-couplings of oxindoles (**180**) with ketones by using

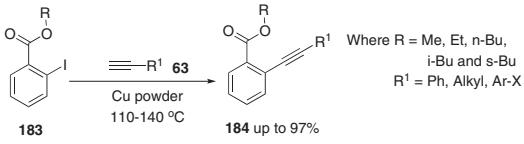
Cu_2O as a catalyst and 2,2,6,6-tetramethylpiperidine *N*-oxyl (TEMPO) in the air with high reaction yields (Scheme 60). This methodology offers a possible approach through the generation of all-carbon quaternary centers at the C3 position of oxindoles with outstanding regioselectivity under mild conditions.



Scheme 60 Cu-catalyzed tandem oxidative cross-couplings of oxindoles[178]

In 2016, Sagadevan and co-workers[179] devised a novel visible-light-initiated Cu-catalyzed process for the cross-coupling reaction of terminal alkynes to furnish bio-active 1,3-unsymmetrical conjugated diynes at room temperature. This method did not require pre-functionalized substrates, ligands, bases, additives, or costly palladium/gold catalysts.

In 2017, Ali and co-workers[180] presented a Cu-catalyzed Sonogashira reaction of alkyl-2-iodobenzoates **183** with alkynes under solvent-, co-catalyst-, and base-free conditions providing coupling product yields up to 97% (Scheme 61). According to the authors, the reported compounds may act as anti-fobic and anti-cizmatic agents and also have the potential to control diseases such as Alzheimer's and schizophrenia.

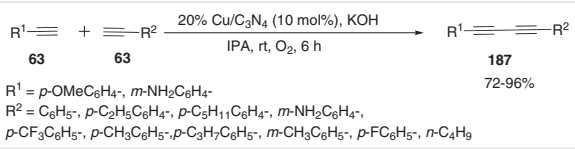


Scheme 61 Cu-catalyzed Sonogashira reaction of alkyl-2-iodobenzoates[180]

In another report in 2018 by Charpe et al.,[181] in which Sagadevan was co-worker, described the first report on visible-light initiated Cu-catalyzed denitrogenative oxidative coupling of 2-hydrazinopyridines (**185**) with terminal alkynes to provide 2-(alkyl/arylethynyl) pyridines **186** at room temperature with N_2 and water as the only byproducts (Scheme 62). The reaction proceeded by formation of copper(II) superoxo/peroxo complex *in situ*. This method offered the green synthesis of 2-methyl-6-(phenylethynyl)pyridine (MPEP), mGluR5 receptor antagonists, and 2-((3-methoxyphenyl)ethynyl)-6-methylpyridine (M-MPEP).

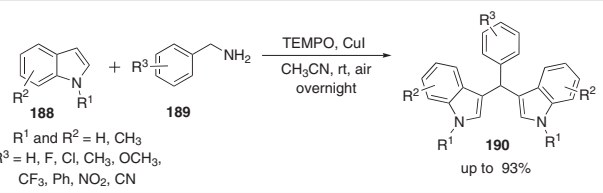
Scheme 62 Visible-light-initiated Cu-catalyzed denitrogenative oxidative coupling[181]

Xu and co-workers[182] prepared an environmentally friendly Cu/C_3N_4 composite and examined it as a highly effective catalyst for the homo- and cross-coupling reaction of terminal alkynes affording symmetrical and unsymmetrical 1,3-diynes **187** in good yields (Scheme 63). The reaction was performed with oxygen as an oxidant in an isopropanol solution with excellent functional group tolerance under ambient conditions.



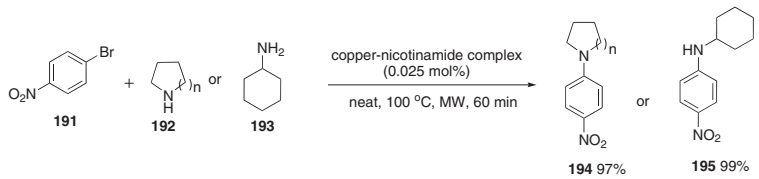
Scheme 63 Cu/C_3N_4 composite-catalyzed coupling of terminal alkynes[182]

Liao et al.[183] found TEMPO/CuI to be an effective catalyst for the cross-coupling of benzylic amines **189** with indoles **188**, generating the corresponding bis(indolyl)phenylmethanes **190** under air at room temperature in high yields (Scheme 64).



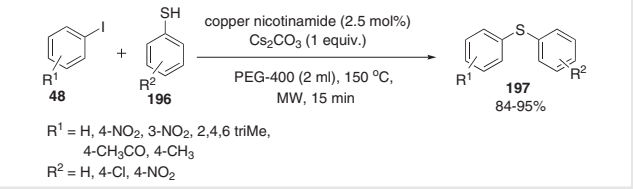
Scheme 64 TEMPO/CuI-catalyzed cross-coupling of benzylic amines with indoles[183]

A mixed example of a Cu-catalyzed coupling reaction was described by Baig and co-workers,[184] who synthesized a versatile crystalline copper(II)-nicotinamide complex that

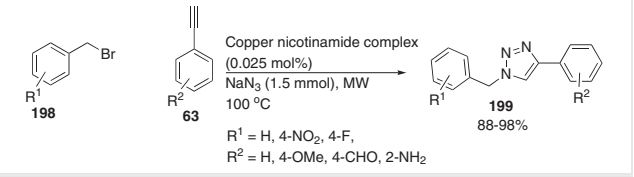


Scheme 65 Copper(II)-nicotinamide complex-catalyzed MW-enhanced C–N coupling reaction[184]

efficiently catalyzed the MW-accelerated C–N, C–S bond-forming and cycloaddition reactions (Scheme 65, Scheme 66, and Scheme 67).



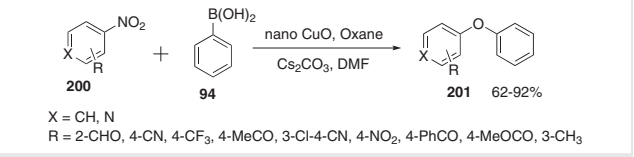
Scheme 66 Copper(II)-nicotinamide complex-catalyzed MW-enhanced C–S coupling reaction[184]



Scheme 67 Copper(II)-nicotinamide complex-catalyzed MW-enhanced cycloaddition reaction[184]

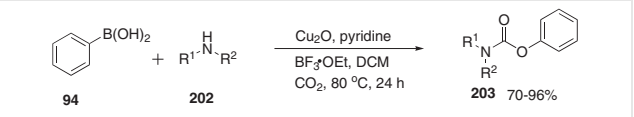
4.2 C–O Cross-Coupling Reaction

In 2012, Zhang and co-workers[185] developed the first example of a Cu-catalyzed coupling of nitroarenes with arylboronic acid, providing diaryl ethers **201** in moderate to excellent yields (Scheme 68). The reaction did not involve any ligand, and deuterium labeling in mechanistic studies showed that water was essential for this transformation.



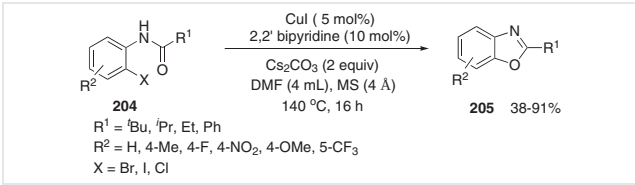
Scheme 68 Cu-catalyzed coupling of nitroarenes with arylboronic acid[185]

In 2017, Xiong and co-workers[186] described the first report on the Cu-catalyzed oxidative coupling reaction of carbon dioxide, amines, and arylboronic acids to synthesize various *O*-aryl carbamates **203** using $BF_3{\cdot}OEt_2$ (Scheme 69). A wide functional group tolerance could be seen in this transformation.



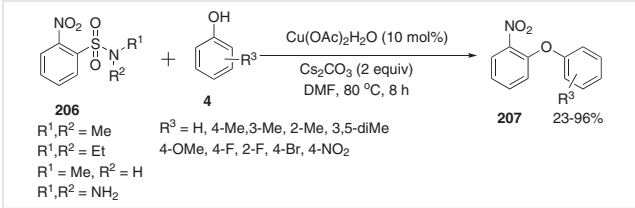
Scheme 69 Cu-catalyzed oxidative coupling reaction[186]

In 2019, a new method for the synthesis of bioactive 2-substituted benzoxazoles **205** was developed by the Saranya research group[187] via Cu-catalyzed intramolecular C–O cross-coupling of 2-haloanilides **204** in moderate to good yields (Scheme 70). This transformation occurred by employing CuI (5 mol%)/2,2′-bipyridine (10 mol%) as a catalytic system, Cs_2CO_3 (2 equiv) as base, and DMF solvent with 4 Å molecular sieves at 140 °C. The reaction was observed to be influenced by the amide and aromatic substituents of 2-haloanilides.



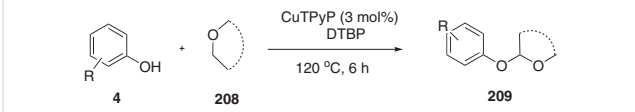
Scheme 70 Cu-catalyzed intramolecular C–O cross-coupling reaction[187]

Chen et al.[188] observed an unprecedented ligand-free Cu-catalyzed *O*-arylation of arenesulfonamides **206** with phenols generating a range of unsymmetric biaryl ethers **207** in excellent yields (Scheme 71). The reaction involved cleavage of C–S bond with excellent regioselectivity and good functional groups tolerance on phenols.

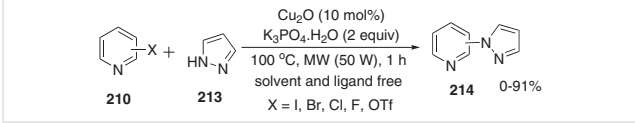


Scheme 71 Cu-catalyzed *O*-arylation of arenesulfonamides[188]

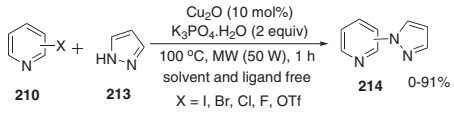
Recently, the Wang group[189] investigated the cross-dehydrogenative coupling (CDC) reaction between the C(sp^3)–H bond and the hydroxyl group of phenol substrates (Scheme 72).



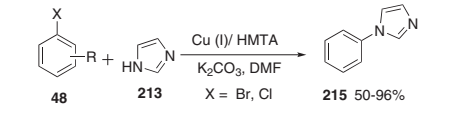
Scheme 72 Cu-catalyzed cross-dehydrogenative coupling (CDC) reaction[189]



Scheme 73 Cu-catalyzed N-arylation of nitrogen-containing heterocycles[190]



Scheme 74 MW-promoted Cu-catalyzed amination of halopyridines[191]



Scheme 75 Cu(I)/HMTA-catalyzed C–N cross-coupling of imidazole and aryl halides[192]

4.3 C–N Cross-Coupling Reaction

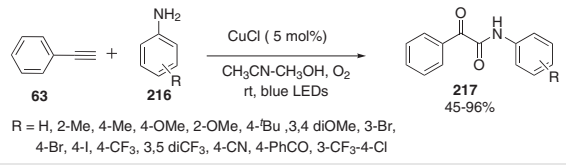
In 2010, the Li group[190] developed a simple Cu-catalyzed method for N-arylations of nitrogen-containing heterocycles and aliphatic amines in water as a solvent and (1*E*,2*E*)-oxalaldehyde dioxime (**211**) as a ligand at 100 °C (Scheme 73).

In 2011, Liu and co-workers[191] described a microwave-promoted solvent- and ligand-free Cu-catalyzed amination of several halopyridines **210** with various nitrogen nucleophiles **213**, giving corresponding *N*-heteroarylated products **214** in good yields (Scheme 74).

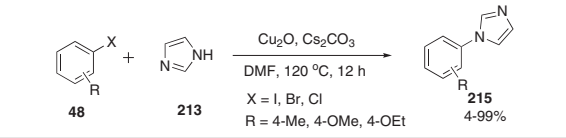
In 2012, Cao and co-workers[192] reported an efficient C–N cross-coupling reaction that allowed the coupling of imidazole (**213**) with aryl chlorides or bromides by employing an inexpensive catalytic system Cu(I)/HMTA, providing products in moderate to good yields. Moreover, the presence of electron-withdrawing or electron-donating groups in the aryl halides had no adverse effect on the outcome of the reaction (Scheme 75).

In 2015, Sagadevan and co-workers[193] reported a copper(I) chloride catalyzed green process for direct oxidative C_{sp}-N coupling reactions of anilines and alkynes affording biologically important α-ketoamides **217** under visible-light irradiation at room temperature without the need for a base, ligands, or an external oxidant (Scheme 76).

In 2017, Wang et al.[194] established a new Cu-catalyzed ligand-free method for Ullmann-type N-arylation of N-containing heterocycles **213** with aryl **48** or heteroaryl bromides or iodides without the protection of an inert gas, affording the desired products with high reaction yields (Scheme 77). In 2021, Bai et al.[195] reported a simple strategy for the C–N cross-coupling of indazole with a modification of substituted aryl bromides under ligand-free conditions.



Scheme 76 Cu-catalyzed direct oxidative C–N coupling reaction[193]



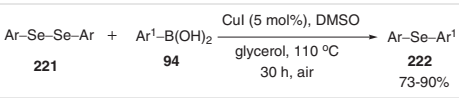
Scheme 77 Cu-catalyzed Ullmann-type N-arylation[195]

4.4 C–P Cross-Coupling Reaction

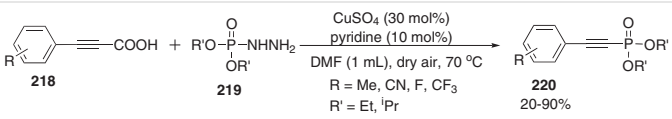
In 2016, the first attractive synthetic tool was provided by the Chen research group[196] for the synthesis of valuable alkynylphosphonates **220**, which involved Cu-catalyzed decarboxylative coupling of various arylpropiolic acids **218** with readily available dialkyl hydrazinylphosphonates **219** giving up to 90% yield (Scheme 78).

4.5 C–Se Cross-Coupling Reaction

In 2012, Ricordi and co-workers[197] described the Cu-catalyzed cross-coupling reaction of diaryl diselenides **221** including arylboronic acids with CuI and DMSO as additive and glycerol as a recyclable solvent, affording the corresponding diaryl selenides **222** in high reaction yields (Scheme 79). The reaction was performed under an open atmosphere at 110 °C.



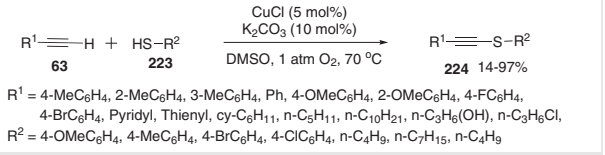
Scheme 79 Cu-catalyzed cross-coupling reaction of diaryl diselenides[197]



Scheme 78 Cu-catalyzed decarboxylative coupling reaction[196]

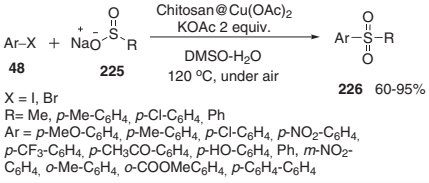
4.6 C–S Cross-Coupling Reaction

In 2013, Yang et al.[198] demonstrated a Cu-catalyzed aerobic cross-dehydrogenative coupling reaction for the synthesis of alkynyl sulfides **224** from terminal alkynes with thiols using K_2CO_3 and molecular O_2 as the oxidant under mild reaction conditions (Scheme 80).



Scheme 80 Cu-catalyzed aerobic cross-dehydrogenative coupling[198]

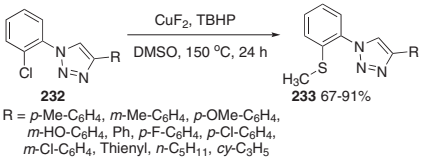
In 2014, Shen and co-workers[199] utilized chitosan@copper as a recoverable catalyst for the synthesis of aryl sulfones **226** from aryl halide and sodium sulfinates through cross-coupling reactions with high reaction yields (Scheme 81). Interestingly, the antiulcer drug zolimidine (**231**) could easily be synthesized by employing this protocol (Scheme 82).



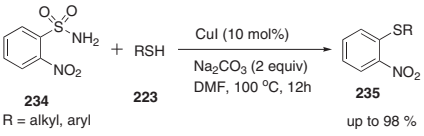
Scheme 81 Chitosan@copper-catalyzed synthesis of aryl sulfones[199]

In 2018, Yu and co-workers[200] demonstrated the regioselective C–S coupling of 1,4-disubstituted 1,2,3-triazole halides (**232**) mediated by CuF_2 using DMSO as a methylthiolation source, in which the *ortho*-C–X bond in the N(1) aryl group was selectively cross-coupled without affecting other C–X bonds (Scheme 83).

Chen et al.[201] presented a novel Cu(I)-catalyzed method for the cross-coupling of 2-nitro benzenesulfonamides (**234**) with thiols generating unsymmetrical sulfides (**235**) in high to excellent yields by using CuI in DMF as solvent at 100 °C (Scheme 84). This method offered **234** as a new coupling partner for the first time and occurred through cleavage of the Ar–SO_2NH_2 bond without cleavage of the C–NO_2 bond.

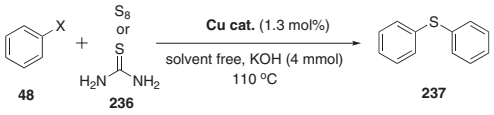


Scheme 83 Cu-catalyzed regioselective C–S coupling[200]



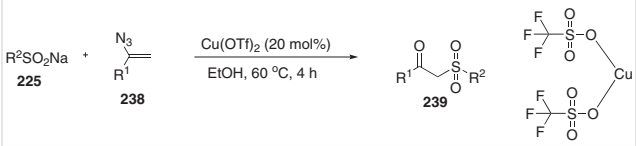
Scheme 84 Cu-catalyzed cross-coupling of 2-nitrobenzenesulfonamides with thiols[201]

In 2019, Ghodsinia and co-workers[202] synthesized a recyclable heterogeneous SBA-16/GPTMS-TSC-CuI catalytic system in which CuI was anchored onto a mesoporous material (SBA-16) functionalized by aminated 3-glycidyloxypropyltrimethoxysilane (GPTMS) with thiosemicarbazide (TSC). This novel mesostructured catalyst was investigated for the C–S coupling products of aryl halides with S8/thiourea under solvent-free conditions in high reaction yields (Scheme 85). The reaction was performed in notably reduced reaction times in comparison to the earlier reports.

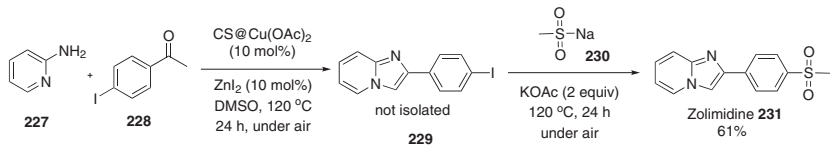


Scheme 85 Heterogeneous SBA-16/GPTMS-TSC-CuI -catalyzed C–S coupling[202]

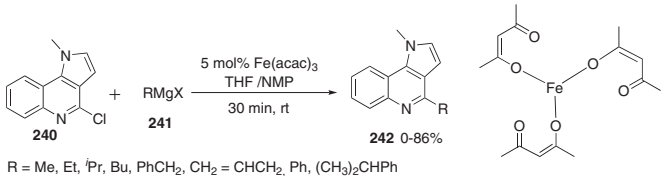
In 2021, the Ning group[203] developed an oxidative cross-coupling reaction between sodium sulfinates **225** and vinyl azides **238** to form β-ketosulfones **239** (Scheme 86).



Scheme 86 Cu-catalyzed oxidative cross-coupling reaction[203]



Scheme 82 Synthesis of antiulcer drug zolimidine using the method developed by Shen *et al.*[199]



Scheme 87 Fe-catalyzed Kumada cross-coupling of 4-chloropyrrolo[3,2-*c*]quinoline[204]

5 Fe-Catalyzed Reactions

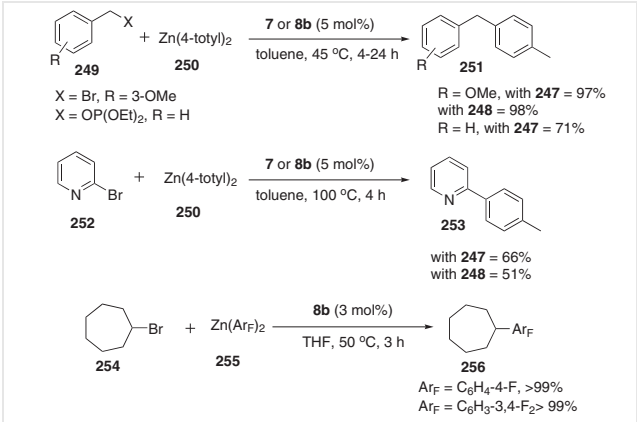
We will discuss in this section the use of iron catalysis in various cross-coupling reactions and is generally presented in chronological order.

5.1 C–C Cross-Coupling Reaction

In 2009, Colacino and co-workers[204] developed Fe-catalyzed cross-coupling reaction of 4-chloropyrrolo[3,2-*c*]quinoline (**240**) with aryl or alkyl magnesium halides in the presence of $Fe(acac)_3$ (Scheme 87). The reaction was performed in a mixture of THF and NMP in just 30 min. The coupled products are useful scaffolds for medicinal chemistry and were obtained moderate to excellent yields of 52–94%.

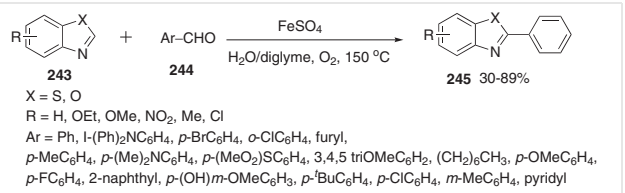
In 2012, Liu et al.[205] discovered a Fe-catalyzed arylation of benzoazoles **243** with aromatic aldehydes with oxygen as an oxidant in good to excellent yields under base-free conditions (Scheme 88). The reaction was achieved by using a mixture of water/diglyme instead of organic solvents and better yields were obtained when benzothiazoles were employed as substrates.

Adams and co-workers[206] demonstrated the synthesis of Fe(I) complexes, $[FeX_2(dpbz)_2]$ [X = 4-tolyl, Cl, Br, dpbz = 1,2-bis(diphenylphosphino)benzene] (Scheme 89) and investigated their catalytic efficiency in Negishi cross-coupling reactions with arylzinc reagents (Scheme 90).

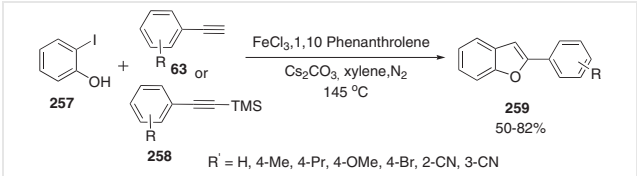


Scheme 90 Fe(I) complex-catalyzed Negishi cross-coupling reactions[206]

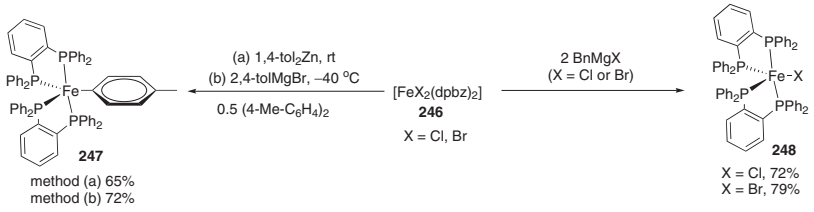
In 2013, Yang and co-workers[207] discovered a Fe-catalyzed method to achieve both the intramolecular O-arylation of *o*-iodophenols and Sonogashira cross-coupling and aryl acetylenes/1-substituted-2-trimethylsilyl acetylenes to afford the corresponding 2-arylbenzo[*b*]furans (**259**) in good reaction yields (Scheme 91). The procedure utilized 5% $FeCl_3$ and 10% 1,10-phenanthroline as a catalytic system.



Scheme 88 Fe-catalyzed arylation of benzoazoles[205]

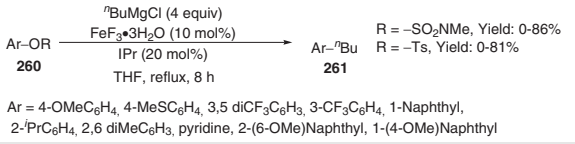


Scheme 91 Fe-catalyzed Sonogashira cross-coupling and intramolecular O-arylation[207]



Scheme 89 Synthesis of Fe(I) complexes $[FeX\text{-}(dpbz)_2]$[206]

Agrawal et al.[208] achieved the Fe-catalyzed cross-coupling of alkyl Grignard reagents using aryl sulfamates or tosylates **260** in quantitative yields (Scheme 92).

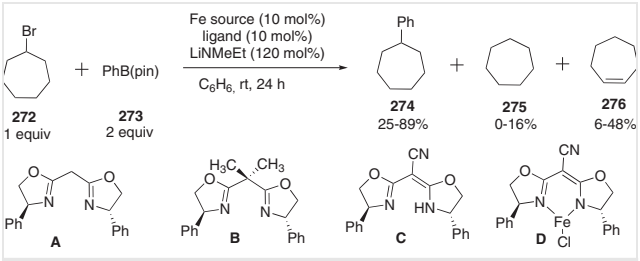


Scheme 92 Fe-catalyzed cross-coupling of aryl sulfamates or tosylates with alkyl Grignard reagents[208]

Hajipour and co-workers[209] prepared heterogeneous Fe-based catalyst supported on acac-functionalized silica, which was employed as a catalyst in Mizoroki–Heck reaction of aryl iodides and olefins in poly(ethylene glycol) as a green solvent (Scheme 93). Interestingly, this protocol allowed selective coupling reaction of aryl iodides in the presence of bromides. The catalyst could be recovered well from the reaction mixture and recycled up to five times.

In 2017, Bisz and co-workers[210] reported that benign cyclic ureas (DMI, DMPU) are efficient and sustainable ligands instead of hazardous NMP in Fe-catalyzed alkylations of aryl chlorides or tosylates with alkyl Grignard reagents (Scheme 94 and Scheme 95). Moreover, this protocol allowed C(sp^2)–C(sp^3) cross-coupling synthesis of a dual NK1/serotonin receptor antagonist.

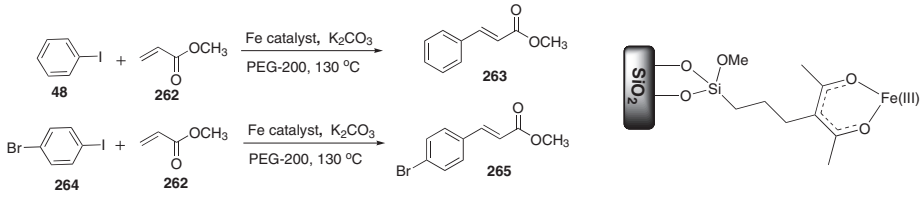
In 2018, Crockett and co-workers[211] discovered a Fe-catalyzed cross-coupling reaction between alkyl halides and arylboronic esters by employing lithium amide bases coupled with Fe complexes containing deprotonated cyano-bis(oxazoline) ligands (A–D) affording up to 89% yields of the coupled products (Scheme 96). Remarkably, the reaction required neither alkyllithium reagents for activation of the boronic ester nor magnesium additives. Moreover, the two-step synthesis of pharmaceutically important Cinacalcet (**281**) was shown by using this protocol (Scheme 97).



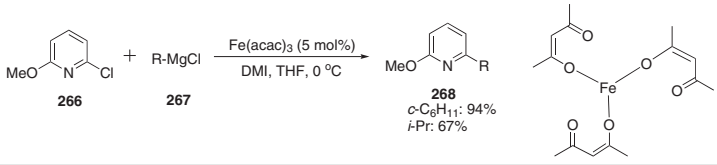
Scheme 96 Fe-catalyzed cross-coupling reaction between alkyl halides and arylboronic esters[211]

5.2 C–S Cross-Coupling Reaction

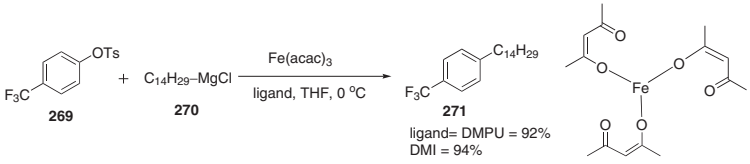
In their study In 2009, Wu and co-workers[212] developed a catalytic system that is greener and used reusable of Fe-Cl_3·$6H_2O$/cationic 2,2′-bipyridyl for the coupling of aryl iodides with thiols to make C–S bond in refluxed water under aerobic conditions (Scheme 98).



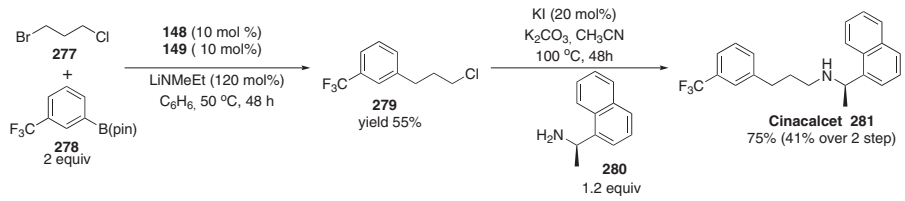
Scheme 93 Fe-catalyzed selective coupling reaction of aryl iodides[209]



Scheme 94 Fe-catalyzed alkylation of aryl chlorides[210]



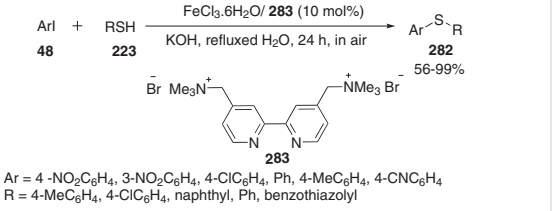
Scheme 95 Fe-catalyzed alkylation of aryl tosylates[210]



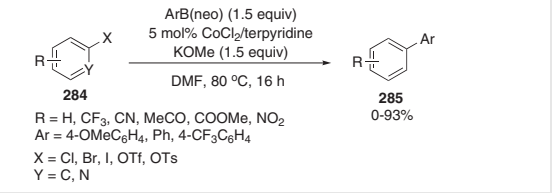
Scheme 97 Fe-catalyzed synthesis of pharmaceutical compound Cinacalcet[211]

6 Co-Catalyzed Reactions

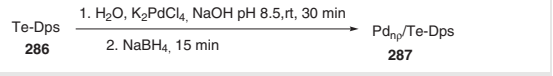
We will discuss in this section the use of cobalt catalysis in various cross-coupling reactions and is generally presented in chronological order.



Scheme 98 $FeCl_3{\cdot}6H_2O$/cationic 2,2′-bipyridyl catalytic system for C–S cross-coupling[212]



Scheme 99 Cobalt(II)/terpyridine-catalyzed SM cross-coupling reaction[213]



Scheme 100 Synthesis of Pd nanoparticles stabilized within the protein cavity of Dps protein[214]

In 2017, Duong and co-workers[213] described a Co-catalyzed Suzuki–Miyaura cross-coupling reaction of aryl halides and arylboronic esters by employing cobalt(II)/terpyridine catalyst and KOMe, generating the corresponding (hetero)biaryls in moderate to excellent yields (Scheme 99). This procedure tolerated the π-electron-rich and π-electron-deficient heteroaryl halides and electron-deficient aryl halides.

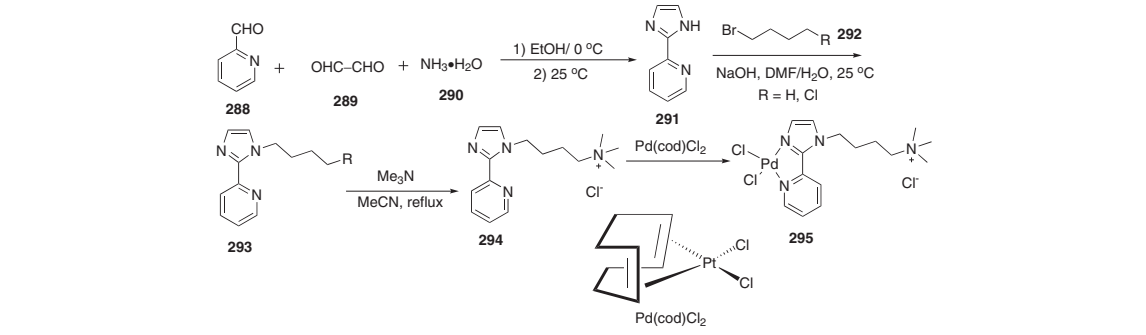
7 Transition-Metal Nanoparticle-Promoted Reactions

We will discuss in this section the use of nanoparticle catalysis in various cross-coupling reactions and is generally presented in chronological order.

7.1 Pd Nanoparticles

In 2009, Prastaro and co-workers[214] prepared a precatalyst consisting of Pd nanoparticles stabilized within the protein cavity of Dps protein (Pdnp/Te-Dps) (**287**) and tested its catalytic ability for Suzuki–Miyaura cross-coupling reactions under phosphine-free, aerobic conditions in water (Scheme 100).

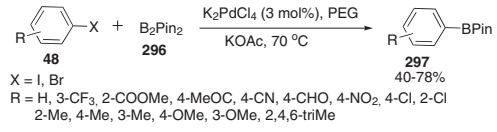
Based on the well-known fact that bacteria can recover Pd(0) in the form of nanoparticles, Søbjerg and co-workers[215] decided to investigate the scope of the reactions that could be catalyzed by bio-recovered palladium. They demonstrated that the Mizoroki–Heck and Suzuki–Miyaura reactions were catalyzed by bio-Pd(0) nanoparticles set up



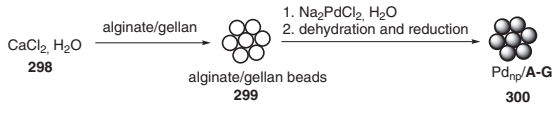
Scheme 101 Synthesis of water-soluble ammonium-functionalized bidentate nitrogen-containing ligand and its Pd chelating complex[216]

on the surface of Gram-negative bacteria such as *C. necator* and *P. putida*. In 2011, Zhou and co-workers[216] synthesized a water-soluble ammonium-functionalized bidentate nitrogen-containing ligand (**294**) and its Pd chelating complex (**295**) and utilized this for Suzuki–Miyaura cross-coupling reaction in neat water under aerobic conditions (Scheme 101).

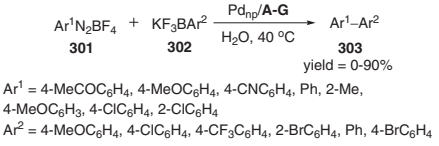
Khalafi-Nezhad and co-workers[217] published a report on the synthesis of a recyclable heterogeneous catalyst system in which they managed to immobilize Pd NPs on a silica-starch substrate (PNP-SSS) and found an effective catalyst in



Scheme 102 Synthesis of aryl boronates[218]



Scheme 103 Synthesis of Pd nanoparticles stabilized by natural beads of alginate/gellan mixture[219]



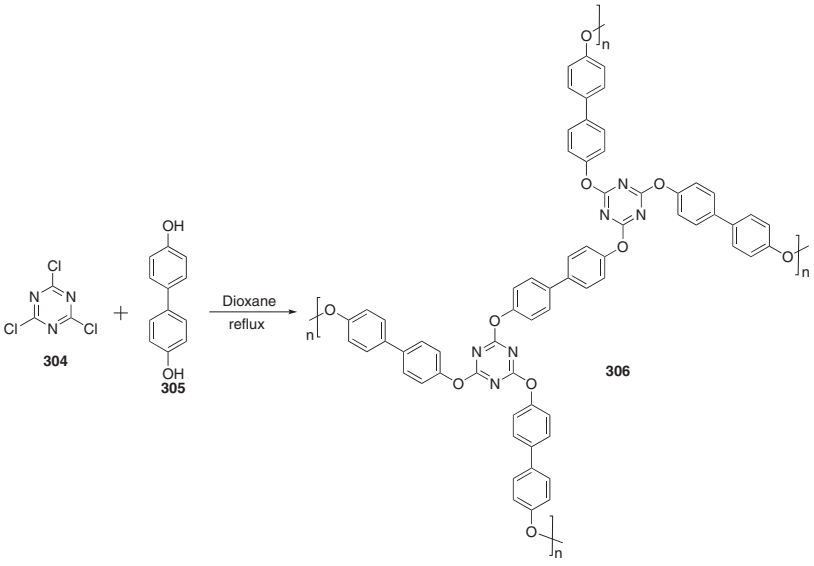
Scheme 104 Pdnp/A–G mediated Suzuki–Miyaura cross-coupling reaction[219]

Heck and copper-free Sonogashira reactions with water as an eco-friendly solvent. The silica-starch substrate effectively stabilized and provided a platform to Pd NPs and prevented their aggregation and separation from the SSS surface. In 2012, Bej and co-workers[218] generated Pd nanoparticles in PEG that catalyzed the reaction of aryl/benzyl halides with bis(pinacolato)diboron to furnish aryl/benzyl boronates in high yield, which, in turn, were used as a reaction partner in the solvent- and ligand-free Suzuki–Miyaura coupling reaction with different aryl/benzyl halides in 53–72% yield (Scheme 102).

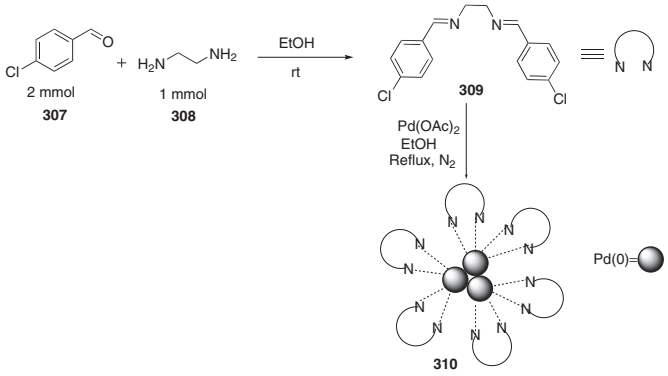
Cacchi and co-workers[219] made Pd nanoparticles stabilized by natural beads of an alginate/gellan mixture for the phosphine- and base-free Suzuki–Miyaura cross-coupling reaction of potassium aryltrifluoroborates and arenediazonium tetrafluoroborates in 1:1 molar ratio with catalyst loading of just 0.01–0.002 mol% under aerobic conditions in water (Scheme 103, Scheme 104).

In 2014, Huang and co-workers[220] reported a synthetic procedure of Pd nanocomposite by depositing palladium nanoparticles in the micropores of the SBA-15 with hydrophobic triphenylsilyl or trimethylsilyl groups grafted on the mesopores. The authors then allowed ligand-free Hiyama cross-couplings of aryl halides and various aryltriethoxysilanes at 100 °C in air. Puthiaraja and co-workers[221] synthesized a novel nitrogen-rich functional mesoporous covalent organic polymer (MCOP), which offered excellent support for Pd nanoparticles (Pd@MCOP) by nucleophilic substitution reaction of cyanuric chloride (**304**) and 4,4′-dihydroxybiphenyl (**305**) (Scheme 105).

In 2015, Mandegani and co-workers[222] developed the synthesis of a novel nano tetraimine Pd(0) complex (**310**) with the complexation of $Pd(OAc)_2$ with *N*,*N*-bisimine li-



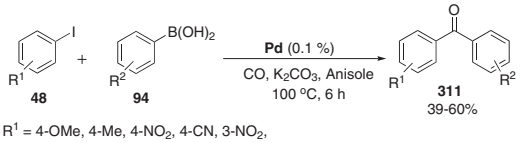
Scheme 105 Synthesis of functional mesoporous covalent organic polymer (MCOP)[221]



Scheme 106 Synthesis of a novel nano tetraimine Pd(0) complex[222]

gand (**309**) (Scheme 106). The catalytic efficiency of this heterogeneous nano-complex was investigated towards the Heck–Mizoroki reaction in water. The catalyst could be reusable and recycled without loss in catalytic activity.

In 2016, Gautam and co-workers[223] investigated the efficiency of PdNPs supported on fibrous nanosilica (KCC-1) towards carbonylative Suzuki–Miyaura cross-coupling reaction with a low Pd loading of 0.1% (Scheme 107). This KCC-1-PEI/Pd catalytic system displayed a TON 28-times and TOF 51-times bigger than already reported supported Pd catalyst in the literature for this reaction, probably owing to the fibrous nature of the KCC-1 support and because the PEI functionalization enhanced the stability.



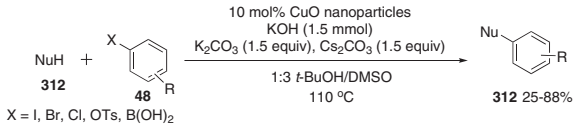
Scheme 107 PdNPs/KCC-1 mediated carbonylative SM cross-coupling reaction[223]

In 2019, Yamada and co-workers[224] investigated the effect of a co-existing metal in the ligand-free Suzuki–Miyaura coupling reaction of an aryl chloride under continuous irradiation microwave and a PdNPs catalyst (SGlPd), and established that the co-existing metal such as aluminum foil is involved in this reaction due to its microwave absorption ability in the reaction system. Mohazzab and co-workers[225] synthesized reusable mesh-GO/Pd catalyst by immobilization of Pd NPs on stainless-steel mesh. Dhara and co-workers[226] prepared glucose-stabilized palladium nanoparticles with recycling and reusing capability up to four times and explored its catalytic potential for both Suzuki and Heck reactions in aqueous medium supported by microwave irradiation. This procedure allowed the coupling of various electron-rich and electron-deficient aryl halides in high reaction yields.

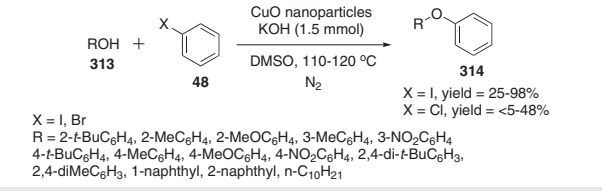
Blanco and co-workers[227] impregnated graphene acid (GA) with $Pd(OAc)_2$ yielding GA-Pd nanohybrids with a size ranging from 1 nm up to 9 nm and applied the material in the Suzuki–Miyaura cross-coupling reaction.

7.2 Cu Nanoparticles

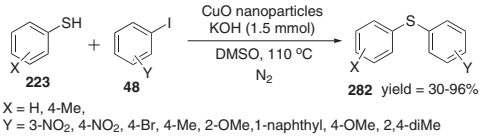
In 2009, Jammi and co-workers[228] studied the catalytic behavior of CuO nanoparticles for C–S, C–O, and C–N bond formations through ligand-free cross-coupling reactions of different nucleophiles such as imidazoles, amides, amines, alcohols, thiols, and phenols with aryl halides by using a base (i.e., KOH, K_2CO_3, and Cs_2CO_3) at moderate temperature to afford the cross-coupled products in high yield (Scheme 108, Scheme 109, and Scheme 110).



Scheme 108 CuO nanoparticle-mediated C–N cross-coupling reaction[228]

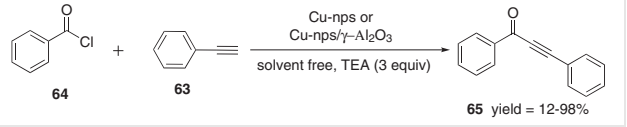


Scheme 109 CuO nanoparticle-mediated C–O cross-coupling reaction[228]



Scheme 110 CuO nanoparticle-mediated C–S cross-coupling reaction[228]

In 2013, Sun and co-workers[229] utilized supported copper NPs for the first time in Pd-, ligand-, and solvent-free coupling reactions of acyl chlorides with terminal alkynes to generate corresponding ynones in 12–98% yield (Scheme 111).



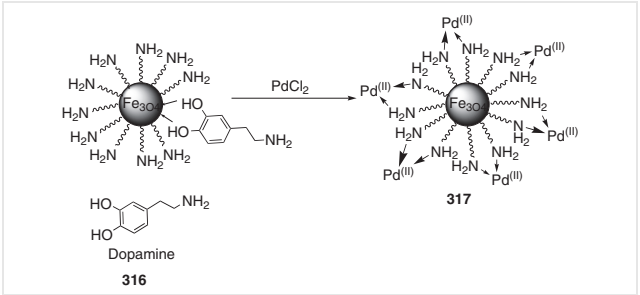
Scheme 111 CuNP-mediated Sonogashira cross-coupling of acyl chlorides[229]

In 2017, a similar report for the synthesis of ynones via solvent-free Sonogashira reactions was disclosed by Wang and co-workers[230] by employing mesoporous phenol-formaldehyde resin-supported copper nanoparticles catalyst (Cu NPs@MP) having wide surface areas and narrow pore-size distributions. The catalyst was synthesized by the melt infiltration of copper nitrate hydrates and subsequent *in-situ* reduction of Cu(II) by template pyrolysis. This catalyst displayed higher catalytic efficiency than copper powder and mesoporous silica SBA-15-supported Cu NPs.

8 Miscellaneous Reactions

In 2010, Guo and co-workers[231] reported a ligand-free iron/copper co-catalyzed amination of aryl halides affording the corresponding products under microwave irradiation in good yields (Scheme 112).

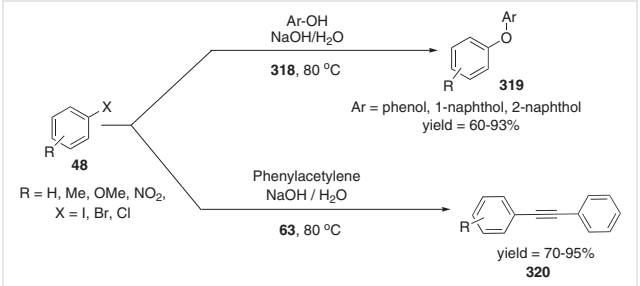
In 2012, Vaddula and co-workers[232] discovered a magnetically recoverable heterogenized Pd catalyst (**317**) for the Heck-type arylation of alkenes with diaryliodonium salts in aqueous polyethylene glycol using ultrasonication within 1–5 min (Scheme 113). The protocol was equally well suited for unactivated alkenes such as styrene, allyl alcohol, and allyl acetate.



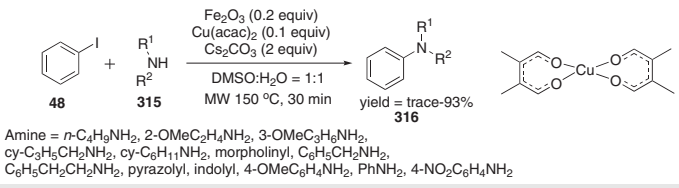
Scheme 113 Pd catalyst for the Heck-type arylation[232]

In 2013, Al-Amin and co-workers[233] utilized sulfur-modified gold-supported palladium material (SAPd) and two categories of microwave approaches, single- and multimode in conjunction with Suzuki–Miyaura cross-coupling reactions. The catalyst had very low leaching properties. In 2017, Akiyama and co-workers[234] described novel sulfur modified gold-supported ruthenium nanoparticles (SARu) via a three-step process involving immobilization of ruthenium using $Ru(acac)_3$ and 4-methoxybenzyl alcohol as a reductant via *in-situ* metal nanoparticle and nano space organization without requiring any preformed template to immobilize and stabilize metal nanoparticles. The catalyst was evaluated for ligand-free Suzuki–Miyaura coupling of arylboronic acids and aryl halides including aryl chlorides with low Ru leaching. In 2013, Zolfigol and co-workers[235] prepared a stable magnetically divisible Pd nanocatalyst [Fe_3O_4@SiO_2@PPh_2@Pd(0)] and explored in Sonogashira cross-coupling reactions and O-arylation of phenols using NaOH as a base in water (Scheme 114).

Steib and co-workers[236] disclosed ligand-free Cr-catalyzed cross-couplings of *N*-heterocyclic halides, alkenyl io-

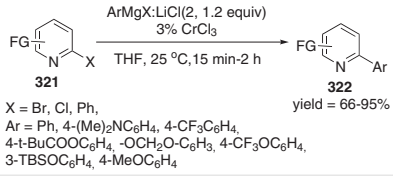


Scheme 114 Pd nanocatalyst [Fe_3O_4@SiO_2@PPh_2@Pd(0)] mediated Sonogashira cross-coupling reactions and O-arylation of phenols[235]



Scheme 112 Iron/copper co-catalyzed amination of aryl halides[231]

dides, aromatic halogenated ketones, or imines with various (hetero)arylmagnesium reagents by employing 3% of chromium(II) chloride at 25 °C (Scheme 115). This method produces lower amounts of homo-coupled products in comparison to the corresponding manganese, iron, or cobalt cross-couplings.



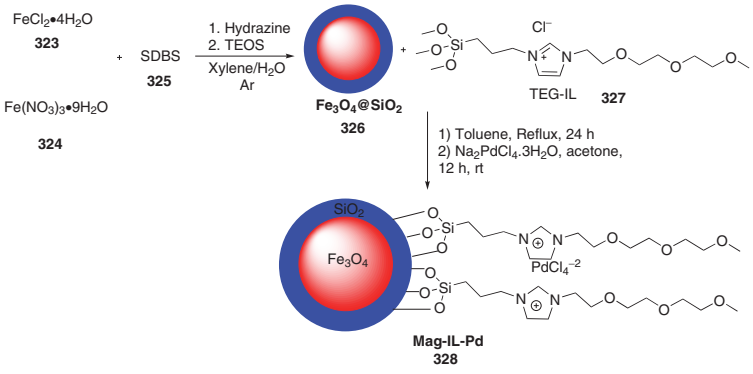
Scheme 115 Cr-catalyzed cross-couplings of *N*-heterocyclic halides[236]

In 2014, Karimi and co-workers[237] prepared a novel magnetically recoverable Pd catalyst (Mag-IL-Pd) (**328**) by anchoring an imidazolium ionic liquid in front of triethylene glycol motifs on the surface of silica-coated iron oxide nanoparticles (Scheme 116). This nanocomposite exhibited notable activity for the Suzuki–Miyaura coupling reaction in water. The protocol allowed the coupling of challenging substrates such as heteroaryl/*ortho*-substituted aryl halides and aryl chlorides efficiently in excellent yields.

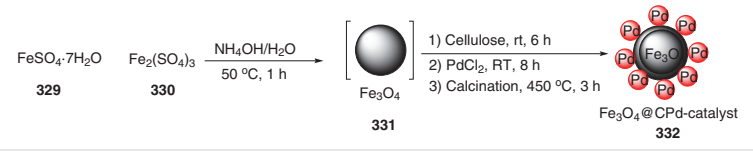
Baig and co-workers[238] synthesized a magnetically recoverable carbon-supported Pd catalyst (**332**) via *in-situ* generation of nano ferrites and fusion of carbon from naturally abundant biopolymer cellulose via calcinations (Scheme 117). The catalytic efficiency of this catalyst was investigated for various reactions such as oxidation of alcohols, arylation of aryl halides, and amination reactions.

In 2016, Rathi and co-workers[239] prepared a nanocatalyst made up of ultra-small Pd/PdO nanoparticles supported on maghemite via co-precipitation using inexpensive raw materials and was employed efficiently in various cross-coupling reactions such as Suzuki and Heck–Mizoroki reaction and the allylic oxidation of alkenes. In 2017, Mohammadinezhad and co-workers[240] reported the synthesis of a heterogeneous magnetically separable core-shell-like Fe_3O_4@Boehmite-NH_2-CoII NPs (**342**) of 13–54 nm size as an eco-friendly catalyst and explored its catalytic efficiency for the Suzuki and Heck cross-coupling reactions in water (Scheme 118). Moreover, the catalyst can be reused at least seven times without a significant and decrease in the catalytic activity.

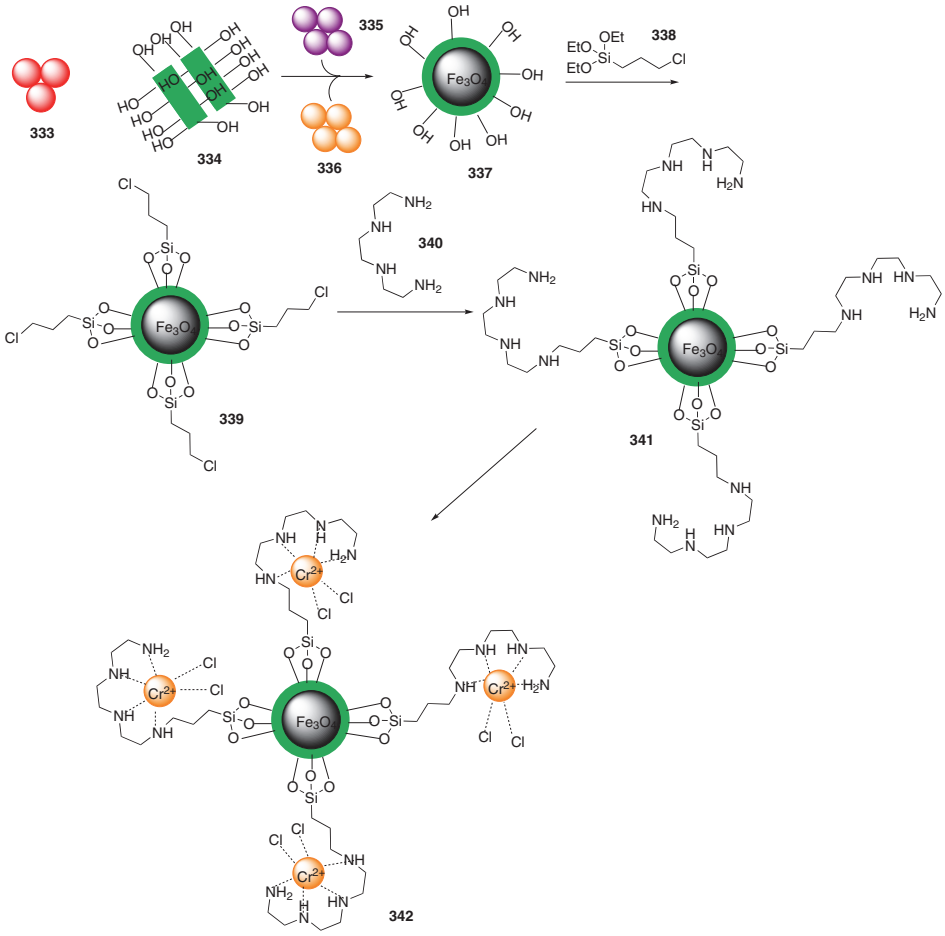
In 2018, Kaur and co-workers[241] presented a green catalytic approach involving zero-valent Pd-Ni alloy NPs using microwaves as an energy source and ethanol/water as a solvent system. Metallosurfactants were synthesized that capped the Pd-Ni alloy NPs. The synthesized NPs exhibited catalytic efficiency for the Heck coupling reaction. The main attributes of the protocol include short reaction time, wide substrate scope, mild reaction conditions, avoiding the use of toxic organic solvents, and reusability of the catalyst. In 2019, Kazemnejadi and co-workers[242] prepared a magnetically recoverable, heterogeneous Fe_3O_4@SiO_2@Im[Cl]Co(III)-melamine nanocomposite and investigated its efficiency for Sonogashira and Heck cross-coupling reactions. The coupling reaction was phosphine-, base-, and ligand-free, used ethanol as solvent and proceeded with high to excellent yields.



Scheme 116 Synthesis of Pd catalyst (Mag-IL-Pd)[237]



Scheme 117 Synthesis of carbon-supported Pd catalyst[238]



Scheme 118 Synthesis of Fe_3O_4@Boehmite-NH_2-CoII NPs[240]

9 Perspectives and Future Directions

This review included information and examples of an impressive and appealing range of past and recent developments of the various approaches to transition-metal-catalyzed cross-coupling reactions such as Suzuki, Heck, Sonogashira, Stille, Kumada, Kochi, Murahashi, Corriu, Hiyama, and Negishi reactions, as well as decarboxylative, carbonylative, and α-arylative, C–O, C–N, C–S bond-forming reactions for the synthesis of natural products and agrochemicals. The past more than 45 years have seen continuous growth in cross-coupling protocols, and plenty of new tools for cross-coupling have been reported by researchers. In the last twelve years especially, we have observed an explosive development of this chemistry. Modern synthetic organic chemistry has seen a marvelous advancement after the dawn of Pd and later by other transition metals (copper, iron, nickel, cobalt or Zr) catalyzed cross-coupling reactions for C–C and C–heteroatom bond formation. As reflected in this chapter, copper, in some cases, can replace the more traditional palladium systems because of some obvious reasons of being cheaper, readily available, and more efficient, using mild reaction conditions, offering high functional group tolerance, and involving less toxic oxygen or nitrogen ligands for selective C–O or C–N bond formation. Similarly, Ni has proved to be an amazingly versatile catalyst for such transformations in large-scale processes compared to the corresponding palladium catalysts, because of its lower cost, greater reactivity toward halocarbon electrophiles such as C–Cl and even inert C–F bonds, C–O-derived electrophiles such as less reactive mesylates and tosylates, phosphates, sulfamates, carbamates, phosphoramides, carbonates, certain esters, activated ethers, and phenols. Apart from the several interesting reports mentioned in this chapter, the field of cobalt and iron-catalyzed cross-coupling reactions is still immature. Since the renaissance of the field of iron-catalyzed cross-coupling in the early 2000s the metal has presented itself as a useful alternative to palladium-catalyzed cross-coupling reactions despite the significant challenges. Finally, it is expected that the recently developed early-transition-metal-catalyzed cross-coupling reactions will serve as an impetus for chemists to address

longstanding challenges in this field in the forthcoming years to enlarge the armory of tools available to them. Hence, the profound interest in early-transition-metals Cu, Ni, Fe, Zr in cross-coupling reactions for C–C and C–heteroatom bond formation and the significant collaborations between organic and inorganic chemists have emerged only in recent years and it will be exciting to see its full potential unlocked soon.

Conflict of Interest

The authors declare no conflict of interest.

Funding Information

S.K. is thankful to Council of Scientific and Industrial Research (CSIR) for his fellowship.

References

(1) Suzuki, A. *Pure Appl. Chem.* **1985**, *57*, 1749.
(2) Heck, R. F. *Org. React.* **1982**, *27*, 345.
(3) Sonogashira, K. In *Comprehensive Organic Synthesis, Vol. 3*; Trost, B. M.; Fleming, I., Ed.; Pergamon: Oxford, **1991**, 521.
(4) (a) Milstein, D.; Stille, J. K. *J. Am. Chem. Soc.* **1979**, *101*, 4981. (b) Milstein, D.; Stille, J. K. *J. Am. Chem. Soc.* **1979**, *101*, 4992. (c) Milstein, D.; Stille, J. K. *J. Org. Chem.* **1979**, *44*, 1613.
(5) (a) Negishi, E. *Acc. Chem. Res.* **1982**, *15*, 340. (b) Erdik, E. *Tetrahedron* **1992**, *48*, 9577. (c) Negishi, E. I.; Liu, F. In *Metal-Catalyzed Cross-Coupling Reactions*; Diederich, F.; Stang, P. J., Ed.; Wiley-VCH: Weinheim, **1998**, 1–47.
(6) Tamao, K.; Sumitani, K.; Kumada, M. *J. Am. Chem. Soc.* **1972**, *94*, 4374.
(7) Nakao, Y.; Hiyama, T. *Chem. Soc. Rev.* **2011**, *40*, 4893.
(8) Schlummer, B.; Scholz, U. *Adv. Synth. Catal.* **2004**, *346*, 1599.
(9) Beletskaya, I. P.; Ananikov, V. P. *Chem. Rev.* **2011**, *111*, 1596.
(10) Nicolaou, K. C.; Bulger, P. G.; Sarlah, D. *Angew. Chem. Int. Ed.* **2005**, *44*, 4442.
(11) Zweig, J. E.; Kim, D. E.; Newhouse, T. R. *Chem. Rev.* **2017**, *117*, 11680.
(12) Torborg, C.; Beller, M. *Adv. Synth. Catal.* **2009**, *351*, 3027.
(13) Magano, J.; Dunetz, J. R. *Chem. Rev.* **2011**, *111*, 2177.
(14) Bao, Z.; Chen, Y.; Cai, R.; Yu, L. *Macromolecules* **1993**, *26*, 5281.
(15) Yang, S.-H.; Hsu, C.-S. *J. Polym. Sci., Part A: Polym. Chem.* **2009**, *47*, 2713.
(16) Chang, N.-H.; Kinoshita, M.; Nishihara, Y. *Liquid Crystals*, In *Applied Cross-Coupling Reactions*; Nishihara, Y., Ed.; Lecture Notes in Chemistry; Springer: Berlin, **2013**, 111–135.
(17) Naso, F.; Babudri, F.; Farinola, G. M. *Pure Appl. Chem.* **1999**, *71*, 1485.
(18) Figueira-Duarte, T. M.; Müllen, K. *Chem. Rev.* **2011**, *111*, 7260.
(19) Ullmann, F.; Bielecki, J. *Ber. Dtsch. Chem. Ges.* **1901**, *34*, 2174.
(20) Tamao, K.; Sumitani, K.; Kumada, M. *J. Am. Chem. Soc.* **1972**, *94*, 4374.
(21) Corriu, R. J. P.; Masse, J. P. *J. Chem. Soc., Chem. Commun.* **1972**, 144a.
(22) Heck, R. F. *Palladium-Catalyzed Vinylation of Organic Halides*, In *Organic Reactions*; John Wiley & Sons Inc: Hoboken, **1982**, 345–390.
(23) Mizoroki, T.; Mori, K.; Ozaki, A. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 581.
(24) Heck, R. F.; Nolley, J. P. *J. Org. Chem.* **1972**, *37*, 2320.
(25) Sonogashira, K.; Tohda, Y.; Hagihara, N. *Tetrahedron Lett.* **1975**, 4467.
(26) Takahashi, S.; Kuroyama, Y.; Sonogashira, K.; Hagihara, N. *Synthesis* **1980**, 627.
(27) Baba, S.; Negishi, E. *J. Am. Chem. Soc.* **1976**, *98*, 6729.
(28) Negishi, E. *Acc. Chem. Res.* **1982**, *15*, 340.
(29) Milstein, D.; Stille, J. K. *J. Am. Chem. Soc.* **1978**, *100*, 3636.
(30) Milstein, D.; Stille, J. K. *J. Am. Chem. Soc.* **1979**, *101*, 4992.
(31) Miyaura, N.; Yamada, K.; Suzuki, A. *Tetrahedron Lett.* **1979**, 3437.
(32) Suzuki, A. *Pure Appl. Chem.* **1985**, *57*, 1749.
(33) Miyaura, N.; Yanagi, T.; Suzuki, A. *Synth. Commun.* **1981**, *11*, 513.
(34) Nakao, Y.; Hiyama, T. *Chem. Soc. Rev.* **2011**, *40*, 4893.
(35) Hatanaka, Y.; Hiyama, T. *J. Org. Chem.* **1988**, *53*, 918.
(36) Hiyama, T. *J. Organomet. Chem.* **2002**, *653*, 58.
(37) Old, D. W.; Wolfe, J. P.; Buchwald, S. L. *J. Am. Chem. Soc.* **1998**, *120*, 9722.
(38) Jana, R.; Pathak, T. P.; Sigman, M. S. *Chem. Rev.* **2011**, *111*, 1417.
(39) Kohei, T.; Miyaura, N. *Introduction to Cross-Coupling Reactions*, In *Cross-Coupling Reactions*; Miyaura, N., Ed.; Topics in Current Chemistry, Vol. 219; Springer: Berlin, **2002**, 1–9.
(40) Negishi, E. J. *Organomet. Chem.* **2002**, *653*, 34.
(41) Seechurn, J. C. C. C.; Kitching, M. O.; Colacot, T. J.; Snieckus, V. *Angew. Chem. Int. Ed.* **2012**, *51*, 5062.
(42) Suzuki A., Heck R. F., Negishi E., The Nobel Prize in Chemistry, **2010** (accessed Feb15, 2018): http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/10/20101062242515 68280.pd
(43) Wu, X.-F.; Anbarasan, P.; Neumann, H.; Beller, M. *Angew. Chem. Int. Ed.* **2010**, *49*, 9047.
(44) Johansson Seechurn, C. C. C.; Kitching, M. O.; Colacot, T. J.; Snieckus, V. *Angew. Chem. Int. Ed.* **2012**, *51*, 5062.
(45) Wu, X.-F.; Anbarasan, P.; Neumann, H.; Beller, M. *Angew. Chem. Int. Ed.* **2010**, *49*, 9047.
(46) Chen, X.; Engle, K. M.; Wang, D.-H.; Yu, J.-Q. *Angew. Chem. Int. Ed.* **2009**, *48*, 5094.
(47) Nicolaou, K. C.; Bulger, P. G.; Sarlah, D. *Angew. Chem. Int. Ed.* **2005**, *44*, 4442.
(48) Schlummer, B.; Scholz, U. *Adv. Synth. Catal.* **2004**, *346*, 1599.
(49) Czaplik W. M., Mayer M., Cvengros J., von Wangelin A. J. **2009**, *2*: 396
(50) Fürstner, A.; Leitner, A.; Mendez, M.; Krause, H. *J. Am. Chem. Soc.* **2002**, *124*, 13856.
(51) Piontek, A.; Bisz, E.; Szostak, M. *Angew. Chem. Int. Ed.* **2018**, *57*, 11116.
(52) Fürstner, A.; Martin, R. *Chem. Lett.* **2005**, *34*, 624.
(53) Li, B.; Wu, Z.-H.; Gu, Y.-F.; Sun, C.-L.; Wang, B.-Q.; Shi, Z.-J. *Angew. Chem.* **2011**, *123*, 1141.
(54) Gensch, T.; Klauck, F. J. R.; Glorius, F. *Angew. Chem. Int. Ed.* **2016**, *55*, 11287.
(55) Cahiez, G.; Moyeux, A. *Chem. Rev.* **2010**, *110*, 1435.
(56) Gosmini, C.; Moncomble, A. *Isr. J. Chem.* **2010**, *50*, 568.
(57) Netherton, M. R.; Fu, G. C. *Adv. Synth. Catal.* **2004**, *346*, 1525.
(58) Wang, Q.; Chen, C. *Tetrahedron Lett.* **2008**, *49*, 2916.
(59) Tobisu, M.; Chatani, N. *Acc. Chem. Res.* **2015**, *48*, 1717.
(60) Phapale, V. B.; Cardenas, D. J. *Chem. Soc. Rev.* **2009**, *38*, 1598.
(61) Rosen, B. M.; Quasdorf, K. W.; Wilson, D. A.; Zhang, N.; Resmerita, A.-M.; Garg, N. K.; Percec, V. *Chem. Rev.* **2011**, *111*, 1346.
(62) Thapa, S.; Shrestha, B.; Gurung, S. K.; Giri, R. *Org. Biomol. Chem.* **2015**, *13*, 4816.

(63) Beletskaya, I. P.; Cheprakov, A. V. *Coord. Chem. Rev.* **2004**, *248*, 2337.
(64) Yang, C.-T.; Zhang, Z.-Q.; Liu, Y.-C.; Liu, L. *Angew. Chem.* **2011**, *123*, 3990.
(65) Colby, D. A.; Tsai, A. S.; Bergman, R. G.; Ellman, J. A. *Acc. Chem. Res.* **2012**, *45*, 814.
(66) *Modern Rhodium-Catalyzed Organic Reactions*; Evans, P. A., Ed.; Wiley-VCH Verlag GmbH: Weinheim, **2005**.
(67) Colby, D. A.; Bergman, R. G.; Ellman, J. A. *Chem. Rev.* **2010**, *110*, 624.
(68) Kuhl, N.; Schröder, N.; Glorius, F. *Adv. Synth. Catal.* **2014**, *356*, 1443.
(69) Lautens, M.; Roy, A.; Fukuoka, K.; Fagnou, K.; Martín-Matute, B. *J. Am. Chem. Soc.* **2001**, *123*, 5358.
(70) Farrington, E. J.; Brown, J. M.; Barnard, C. F. J.; Rowsell, E. *Angew. Chem. Int. Ed.* **2002**, *41*, 169.
(71) Oi, S.; Aizawa, E.; Ogino, Y.; Inoue, Y. *J. Org. Chem.* **2005**, *70*, 3113.
(72) Koike, T.; Du, X.; Sanada, T.; Danda, Y.; Mori, A. *Angew. Chem. Int. Ed.* **2003**, *42*, 89.
(73) Su, B.; Cao, Z.-C.; Shi, Z.-J. *Acc. Chem. Res.* **2015**, *48*, 886.
(74) Cho, S. H.; Kim, J. Y.; Kwak, J.; Chang, S. *Chem. Soc. Rev.* **2011**, *40*, 5068.
(75) Miao, J.; Ge, H. *Eur. J. Org. Chem.* **2015**, 7859.
(76) Balasubramanian, M. *Industrial Scale Palladium Chemistry*, In *Palladium in Heterocyclic Chemistry, Vol. 26*; Li, J. J.; Gribble, G. W., Ed.; Elsevier: Amsterdam, **2007**, 587–620.
(77) Torborg, C.; Beller, M. *Adv. Synth. Catal.* **2009**, *351*, 3027.
(78) For recent reviews containing Ni-catalyzed cross-couplings involving C–O bond, see: (a) Yu, D.-G.; Li, B.-J.; Shi, Z.-J. *Acc. Chem. Res.* **2010**, *43*, 1486. (b) Mesganaw, T.; Garg, N. K. *Org. Process Res. Dev.* **2013**, *17*, 29.
(79) (a) Zhao, Y.-L.; Wu, G.-J.; Han, F.-S. *Chem. Commun.* **2012**, *48*, 5868; and references therein. (b) Zhao, Y.-L.; Wu, G.-J.; Li, Y.; Gao, L.-X.; Han, F.-S. *Chem. Eur. J.* **2012**, *18*, 9622; and references therein.
(80) (a) Rosen, B. M.; Quasdorf, K. W.; Wilson, D. A.; Zhang, N.; Resmerita, A.-M.; Garg, N. K.; Percec, V. *Chem. Rev.* **2011**, *111*, 1346. (b) Li, B.-J.; Yu, D.-G.; Sun, C.-L.; Shi, Z.-J. *Chem. Eur. J.* **2011**, *17*, 1728.
(81) (a) Bolm, C.; Legros, J.; Le Paih, J.; Zani, L. *Chem. Rev.* **2004**, *104*, 6217. (b) Shinokubo, H.; Oshima, K. *Eur. J. Org. Chem.* **2004**, 2071. (c) Fürstner, A.; Martin, R. *Chem. Lett.* **2005**, *34*, 624. (d) Sherry, B. D.; Fürstner, A. *Acc. Chem. Res.* **2008**, *41*, 1500. (e) Cahiez, G.; Duplais, C. *The Chemistry of Organomagnesium Compounds*; Wiley VCH: Weinheim, **2008**. (f) Plietker, B. *Iron Catalysis in Organic Synthesis*; Wiley-VCH: Weinheim, **2008**. (g) Correa, A.; Mancheno, O. G.; Bolm, C. *Chem. Soc. Rev.* **2008**, *37*, 1108. (h) Czaplik, W. M.; Mayer, M.; Cvengros, J.; von Wangelin, A. J. *ChemSusChem* **2009**, *2*, 396. (i) Sun, C.-L.; Li, B.-J.; Shi, Z.-J. *Chem. Rev.* **2011**, *111*, 1293. (j) Nakamura, E.; Hatakeyama, T.; Ito, S.; Ishizuka, K.; Ilies, L.; Nakamura, M. *Org. React.* **2014**, *83*, 1.
(82) (a) Bauer, E. *Iron Catalysis II*; Springer International Publishing: Switzerland, **2015**. (b) Plietker, B. *Iron Catalysis in Organic Chemistry: Reactions and Applications*; Wiley-VCH: New York, **2008**. (c) Su, B.; Cao, Z.-C.; Shi, Z.-J. *Acc. Chem. Res.* **2015**, *48*, 886. (d) Kuzmina, O. M.; Steib, A. K.; Moyeux, A.; Cahiez, G.; Knochel, P. *Synthesis* **2015**, 1696. (e) Bedford, R. B.; Brenner, P. B. *Top. Organomet. Chem.* **2015**, *50*, 19. (f) Bauer, I.; Knölker, H.-J. *Chem. Rev.* **2015**, *115*, 3170. (g) Gopalaiah, K. *Chem. Rev.* **2013**, *113*, 3248.
(83) Kharasch, M. S.; Fields, E. K. *J. Am. Chem. Soc.* **1941**, *63*, 2316.
(84) (a) Tamura, M.; Kochi, J. K. *J. Am. Chem. Soc.* **1971**, *93*, 1487. (b) Tamura, M.; Kochi, J. K. *Synthesis* **1971**, 303. (c) Kwan, C. L.; Kochi, J. K. *J. Am. Chem. Soc.* **1976**, *98*, 4903. (d) Smith, R. S.; Kochi, J. K. *J. Org. Chem.* **1976**, *41*, 502.
(85) Fabre, J.-L.; Julia, M.; Verpeaux, J.-N. *Tetrahedron Lett.* **1982**, *23*, 2469.
(86) Molander, G. A.; Rahn, B. J.; Shubert, D. C.; Bonde, S. E. *Tetrahedron Lett.* **1983**, *24*, 5449.
(87) (a) Cahiez, G.; Marquais, S. *Pure Appl. Chem.* **1996**, *68*, 53. (b) Cahiez, G.; Marquais, S. *Tetrahedron Lett.* **1996**, *37*, 1773. (c) Cahiez, G.; Avedissian, H. *Synthesis* **1998**, 1199. (d) Dohle, W.; Kopp, F.; Cahiez, G.; Knochel, P. *Synlett* **2001**, 1901.
(88) (a) Fürstner, A.; Leitner, A.; Mendez, M.; Krause, H. *J. Am. Chem. Soc.* **2002**, *124*, 13856. (b) Fürstner, A.; Leitner, A. *Angew. Chem. Int. Ed.* **2002**, *41*, 609. (c) Fürstner, A.; Leitner, A. *Angew. Chem. Int. Ed.* **2003**, *42*, 308. (d) Martin, R.; Fürstner, A. *Angew. Chem. Int. Ed.* **2004**, *43*, 3955.
(89) Nagano, T.; Hayashi, T. *Org. Lett.* **2004**, *6*, 1297.
(90) Nakamura, M.; Matsuo, K.; Ito, S.; Nakamura, E. *J. Am. Chem. Soc.* **2004**, *126*, 3686.
(91) Itami, K.; Higashi, S.; Mineno, M.; Yoshida, J.-i. *Org. Lett.* **2005**, *7*, 1219.
(92) (a) Bedford, R. B.; Bruce, D. W.; Frost, R. M.; Goodby, J. W.; Hird, M. *Chem. Commun.* **2004**, 2822. (b) Bedford, R. B.; Bruce, D. W.; Frost, R. M.; Hird, M. *Chem. Commun.* **2005**, 4161. (c) Bedford, R. B.; Betham, M.; Bruce, D. W.; Danopoulos, A. A.; Frost, R. M.; Hird, M. *J. Org. Chem.* **2006**, *71*, 1104.
(93) Dunet, G.; Knochel, P. *Synlett* **2006**, 407.
(94) Li, B.-J.; Xu, L.; Wu, Z.-H.; Guan, B.-T.; Sun, C.-L.; Wang, B.-Q.; Shi, Z.-J. *J. Am. Chem. Soc.* **2009**, *131*, 14656.
(95) Bouissane, L.; Sestelo, J. P.; Sarandeses, L. A. *Org. Lett.* **2009**, *6*, 1285.
(96) Raju, J.; Vance, R.; Goodman, J. A.; Reeder, A. J.; Liao, M. R.; Maust, J.; Irish, M. T.; Espino, P. A.; Andrews, P. D. R. *Org. Lett.* **2009**, *11*, 1463.
(97) Cai, M.; Zheng, G.; Ding, G. *Green Chem.* **2009**, *11*, 1687.
(98) Basu, B.; Biswas, K.; Kundu, S.; Ghosh, S. *Green Chem.* **2010**, *12*, 1734.
(99) Lee, D.-H.; Jung, J.-Y.; Jin, M.-J. *Green Chem.* **2010**, *12*, 2024.
(100) Milton, E. J.; Clarke, M. L. *Green Chem.* **2010**, *12*, 381.
(101) Peng, Y.-Y.; Liu, J.; Lei, X.; Yin, Z. *Green Chem.* **2010**, *12*, 1072.
(102) Marziale, A. N.; Faul, S. H.; Reiner, T.; Schneider, S.; Eppinger, J. *Green Chem.* **2010**, *12*, 35.
(103) Long, B.; Yang, Y.; Wang, L.; He, W.; Xiang, J. *Synth. Commun.* **2010**, *40*, 1202.
(104) Chen, Y.; Cao, W.; Wu, K.; Chen, J.; Lee, A. W. M. *Synth. Commun.* **2010**, *40*, 984.
(105) Molander, G. A.; Hiebel, M.-A. *Org. Lett.* **2010**, *2*, 4876.
(106) Liu, C.; Ni, Q.; Bao, F.; Qiu, J. *Green Chem.* **2011**, *13*, 1260.
(107) Islam, M.; Mondal, P.; Roy, A. S.; Tuhina, K.; Mondal, S.; Hossain, D. *Synth. Commun.* **2011**, *41*, 2583.
(108) Lee, P. H.; Mo, J.; Kang, D.; Eom, D.; Park, C.; Lee, C.-H.; Jung, Y. M.; Hwang, H. *J. Org. Chem.* **2011**, *76*, 312.
(109) Ali, H.; Ait-Mohand, S.; Gosselin, S.; van Lier, J. E.; Guerin, B. *J. Org. Chem.* **2011**, *76*, 1887.
(110) Modak, A.; Mondal, J.; Sasidharan, M.; Bhaumik, A. *Green Chem.* **2011**, *13*, 1317.
(111) Modak, A.; Mondal, J.; Bhaumik, A. *Green Chem.* **2012**, *14*, 2840.
(112) Susanto, W.; Chu, C.-Y.; Ang, W. J.; Chou, T.-C.; Lo, L.-C.; Lam, Y. *Green Chem.* **2012**, *14*, 77.
(113) Mondal, M.; Bora, U. *Green Chem.* **2012**, *14*, 1873.
(114) Wang, L.; Welker, M. E. *J. Org. Chem.* **2012**, *77*, 8280.
(115) Liu, C.; Zhang, Y.; Liu, N.; Qiu, J. *Green Chem.* **2012**, *14*, 2999.

(116) Keller, M.; Hameau, A.; Spataro, G.; Ladeira, S.; Caminade, A.-M.; Majoral, J.-P.; Ouali, A. *Green Chem.* **2012**, *14*, 2807.
(117) Saha, P.; Mondal, S.; Hazra, A.; Naskar, S.; Maity, A.; Sahu, K. B.; Paira, R.; Das, S.; Banerjee, S.; Mondal, N. B. *Synth. Commun.* **2012**, *42*, 3166.
(118) Lu, G.-P.; Caib, C.; Lipshutz, B. H. *Green Chem.* **2013**, *15*, 105.
(119) Oberholzera, M.; Frech, C. M. *Green Chem.* **2013**, *15*, 1678.
(120) Yang, J.; Wang, D.; Liu, W.; Zhang, X.; Bian, F.; Yu, W. *Green Chem.* **2013**, *15*, 3429.
(121) Ding, G.; Wang, W.; Jiang, T.; Han, B. *Green Chem.* **2013**, *15*, 3396.
(122) Yan, H.; Sun, P.; Qu, X.; Lu, L.; Zhu, Y.; Yang, H.; Mao, J. *Synth. Commun.* **2013**, *43*, 2773.
(123) Zhang, G.; Luan, Y.; Han, X.; Wang, Y.; Wen, X.; Ding, C.; Gao, J. *Green Chem.* **2013**, *15*, 2081.
(124) Yuan, H.; Shen, Y.; Yu, S.; Shan, L.; Sun, Q.; Zhang, W. *Synth. Commun.* **2013**, *43*, 2817.
(125) Cheng, K.; Hu, S.; Zhao, B.; Zhang, X.-M.; Qi, C. *J. Org. Chem.* **2013**, *78*, 5022.
(126) Sakashita, S.; Takizawa, M.; Sugai, J.; Ito, H.; Yamamoto, Y. *Org. Lett.* **2013**, *15*, 4308.
(127) Ojha, D. P.; Prabhu, K. R. *J. Org. Chem.* **2013**, *78*, 12136.
(128) Liu, L.; Dong, Y.; Tang, N. *Green Chem.* **2014**, *16*, 2185.
(129) Zhao, H.; Cheng, M.; Zhang, J.; Cai, M. *Green Chem.* **2014**, *16*, 2515.
(130) Nehra, P.; Khungar, B.; Pericherla, K.; Sivasubramanian, S. C.; Kumar, A. *Green Chem.* **2014**, *16*, 4266.
(131) Li, L.; Zhao, S.; Joshi-Pangu, A.; Diane, M.; Biscoe, M. R. *J. Am. Chem. Soc.* **2014**, *136*, 14027.
(132) Nadaf, R. N.; Seapy, D. G. *Synth. Commun.* **2014**, *44*, 2012.
(133) Strueben, J.; Gates, P. J.; Staubitz, A. *J. Org. Chem.* **2014**, *79*, 1719.
(134) Sun, F.; Lv, L.; Huang, M.; Zhou, Z.; Fang, X. *Org. Lett.* **2014**, *16*, 5024.
(135) Tan, J.; Chen, Y.; Li, H.; Yasuda, N. *J. Org. Chem.* **2014**, *79*, 8871.
(136) Shen, C.; Shen, H.; Yang, M.; Xia, C.; Zhang, P. *Green Chem.* **2015**, *17*, 225.
(137) da Silva, J. F. M.; Perez, A. F. Y.; de Almeida, N. P. *Synth. Commun.* **2015**, *17*, 1995.
(138) Boruah, P. R.; Ali, A. A.; Saikia, B.; Sarma, D. *Green Chem.* **2015**, *17*, 1442.
(139) Strappaveccia, G.; Ismalaj, E.; Petrucci, C.; Lanari, D.; Marrocchi, A.; Drees, M.; Facchetti, A.; Vaccaro, L. *Green Chem.* **2015**, *17*, 365.
(140) Strappaveccia, G.; Luciani, L.; Bartollini, E.; Marrocchi, A.; Pizzo, F.; Vaccaro, L. *Green Chem.* **2015**, *17*, 1071.
(141) Khana, R. I.; Pitchumani, K. *Green Chem.* **2016**, *18*, 5518.
(142) Jadhav, S. N.; Kumbhar, A. S.; Rode, C. V.; Salunkhe, R. S. *Green Chem.* **2016**, *18*, 1898.
(143) Wang, J.; Zong, Y.; Wang, X.; Hu, Y.; Yuea, G.; Pan, Y. *Green Chem.* **2016**, *18*, 967.
(144) Nambo, M.; Keske, E. C.; Rygus, J. P. G.; Yim, J. C.-H.; Crudden, C. M. *Green Chem.* **2017**, *7*, 1108.
(145) Pulipati, Y.; Gurram, V.; Laxmi, S. V.; Satyanarayana, Y.; Singh, K.; Kumar, V.; Sharma, S.; Pottabathini, N.; Iska, V. B. R. *Synth. Commun.* **2017**, *12*, 1142.
(146) Chen, X.; Zhou, X.-Y.; Wang, L.-G. *Synth. Commun.* **2017**, *22*, 2096.
(147) Jadhav, S. N.; Rode, C. V. *Green Chem.* **2017**, *19*, 5958.
(148) Clavé, G.; Pelissier, F.; Campidellib, S.; Grison, C. *Green Chem.* **2017**, *19*, 4093.
(149) Markovic, T.; Murray, P. R. D.; Rocke, B. N.; Shavnya, A.; Blakemore, D. C.; Willism, M. C. *J. Am. Chem. Soc.* **2018**, *140*, 15916.
(150) Sharma, S.; Sarkar, B. R. *Synth. Commun.* **2018**, *8*, 906.
(151) Landstrom, E. B.; Handa, S.; Aue, D. H.; Gallou, F.; Lipshutz, B. H. *Green Chem.* **2018**, *20*, 3436.
(152) Gong, X.; Wu, J.; Meng, Y.; Zhang, Y.; Ye, L.-W.; Zhu, C. *Green Chem.* **2019**, *21*, 995.
(153) Chehrouri, M.; Othman, A. A.; Jiménez-Cecilia, S.; Cabrerizo, C. M.; Sansano, J. M. *Synth. Commun.* **2019**, *10*, 1301.
(154) Balfour, M. N.; Zukerman-Schpector, J.; Rodriguez, M. J. D.; Reis, J. S.; Esteves, C. H. A.; Stefani, H. A. *Synth. Commun.* **2019**, *3*, 351.
(155) Liu, X.; Xu, W.; Xiang, D.; Zhang, Z.; Chen, D.; Hu, Y.; Li, Y.; Ouyang, Y.; Lin, H. *New J. Chem.* **2019**, *43*, 12206.
(156) Zhou, Z.; Li, J.; Xu, Z.; Cai, M. *Synth. Commun.* **2020**, *13*, 2015.
(157) Bangar, P. G.; Nahide, P. D.; Meroliya, H. K.; Waghmode, S. A.; Iyer, S. *Synth. Commun.* **2021**, *2*, 308.
(158) Fors, B. P.; Buchwald, S. L. *J. Am. Chem. Soc.* **2010**, *45*, 15914.
(159) Dooleweerdt, K.; Fors, B. P.; Buchwald, S. L. *Org. Lett.* **2010**, *10*, 2350.
(160) Perez, F.; Minatti, A. *Org. Lett.* **2011**, *8*, 1984.
(161) Tambade, P. J.; Patil, Y. P.; Qureshi, Z. S.; Dhake, K. P.; Bhanage, B. M. *Synth. Commun.* **2012**, *1*, 176.
(162) Zhang, G.; Wang, X.; Li, C.; Yin, D. *Synth. Commun.* **2013**, *43*, 456.
(163) Wagner, P.; Bollenbach, M.; Doebelin, C.; Bihel, F.; Bourguignon, J.-J.; Salomé, C.; Schmitt, M. *Green Chem.* **2014**, *16*, 4170.
(164) Fan, A.; Peng, J.; Zhou, D.; Li, X.; Chen, C. *Synth. Commun.* **2020**, *20*, 3113.
(165) Feng, T.; Luo, X.; Dong, J.; Mo, J. *Synth. Commun.* **2021**, *51*,08 1284.
(166) Patel, J. J.; Patel, A. P.; Chikhalia, K. H. *Synth. Commun.* **2021**, *1*, 81.
(167) Xu, K.; Yang, F.; Zhang, G.; Wu, Y. *Green Chem.* **2013**, *15*, 1055.
(168) Hayashi, M.; Matsuura, T.; Tanaka, I.; Ohta, H.; Watanabe, Y. *Org. Lett.* **2013**, *13*, 628.
(169) Taylor, B. L. H.; Jarvo, E. R. *J. Org. Chem.* **2011**, *76*, 7573.
(170) Liu, H.; Liang, Z.; Qian, Q.; Lin, K. *Synth. Commun.* **2014**, *20*, 2999.
(171) Tollefson, E. J.; Dawson, D. D.; Osborne, C. A.; Jarvo, E. R. *J. Am. Chem. Soc.* **2014**, *136*, 14951.
(172) Tollefson, E. J.; Hanna, L. E.; Jarvo, E. R. *Acc. Chem. Res.* **2015**, *48*, 2344.
(173) Dawson, D. D.; Jarvo, E. R. *Org. Process Res. Dev.* **2015**, *19*, 1356.
(174) Funicello, M.; Chiummiento, L.; Lupattelli, P.; Tramutola, F. *Synth. Commun.* **2015**, *15*, 1799.
(175) Liao, Y.; Yang, W.; Wei, T.; Cai, M. *Synth. Commun.* **2019**, *9*, 1134.
(176) Yalavarty, M.; Paradise, C. M.; Klein, T. A. L.; Miles, D. H. *Synth. Commun.* **2010**, *11*, 1625.
(177) Chen, G.; Xie, J.; Weng, J.; Zhu, X.; Zheng, Z.; Cai, J.; Wan, Y. *Synth. Commun.* **2011**, *21*, 3123.
(178) Wang, B.; Zhu, J.; Wei, Y.; Luo, G.; Qu, H.; Liu, L. X. *Synth. Commun.* **2015**, *24*, 2841.
(179) Sagadevan, A.; Lyub, P.-C.; Hwang, K. C. *Green Chem.* **2016**, *18*, 4526.
(180) Ali, I.; Nighot, D.; Lone, M. N.; Jain, A. *Synth. Commun.* **2017**, *12*, 1175.
(181) Charpe, V. P.; Hande, A. A.; Sagadevan, A.; Hwang, K. C. *Green Chem.* **2018**, *20*, 4859.
(182) Xu, H.; Wu, K.; Tian, J.; Zhu, L.; Yao, X. *Green Chem.* **2018**, *20*, 793.
(183) Liao, M.; Zhang, X.; Yue, P. *Synth. Commun.* **2018**, *13*, 1694.
(184) Baig, R. B. N.; Vaddula, B. R.; Nadagouda, M. N.; Varma, R. S. *Green Chem.* **2015**, *17*, 1243.
(185) Zhang, J.; Chen, J.; Liu, M.; Zheng, X.; Ding, J.; Wu, H. *Green Chem.* **2012**, *14*, 912.
(186) Xiong, W.; Qi, C.; Guo, T.; Zhang, M.; Chena, K.; Jiang, H. *Green Chem.* **2017**, *19*, 1642.

(187) Saranya, T. V.; Sruthi, P. R.; Anas, S. *Synth. Commun.* **2019**, *2*, 297.
(188) Chen, J.; Wang, J.; Chen, X.; Huang, Y.; Shouzhi, P. *Synth. Commun.* **2019**, *6*, 836.
(189) Wang, F.-H.; Liu, Z.-Y.; Yang, S.; Shi, L.; Lin, D.-Z.; Liu, H.-Y.; Yuan, G.-Q. *Synth. Commun.* **2021**, *13*, 2053.
(190) Li, X.; Yang, D.; Jiang, Y.; Fu, H. *Green Chem.* **2010**, *12*, 1097.
(191) Liu, Z.-J.; Vors, J.-P.; Gesing, E. R. F.; Bolm, C. *Green Chem.* **2011**, *13*, 42.
(192) Cao, C.; Lu, Z.; Cai, Z.; Pang, G.; Shi, Y. *Synth. Commun.* **2012**, *2*, 279.
(193) Sagadevan, A.; Ragupathi, A.; Lin, C.-C.; Hwu, J. R.; Hwang, K. C. *Green Chem.* **2015**, *17*, 1113.
(194) Wang, X.; Wang, M.; Xie, J. *Synth. Commun.* **2017**, *19*, 1797.
(195) Bai, D.-X.; Lim, R. S.-E.; Ng, H.-F.; Teo, Y.-C. *Synth. Commun.* **2021**, *9*, 1398.
(196) Chen, W.; Ma, D.; Hu, G.; Hong, Z.; Gao, Y.; Zhao, Y. *Synth. Commun.* **2016**, *14*, 1175.
(197) Ricordi, V. G.; Freitas, C. S.; Perin, G.; Lenardão, E. J.; Jacob, R. G.; Savegnago, L.; Alves, D. *Green Chem.* **2012**, *14*, 1030.
(198) Yang, Y.; Dong, W.; Guo, Y.; Rioux, R. M. *Green Chem.* **2013**, *15*, 3170.
(199) Shen, C.; Xu, J.; Yu, W.; Zhang, P. *Green Chem.* **2014**, *16*, 3007.
(200) Yu, S.; Liu, Y.; Wang, J.; Luo, F.; Cheng, H.; Jiang, Y. *Synth. Commun.* **2018**, *19*, 2566.
(201) Chen, J.; Zhang, K.; Zhao, Y.; Pu, S. *Synth. Commun.* **2018**, *11*, 1316.
(202) Ghodsinia, S. S. E.; Akhlaghinia, B. *Green Chem.* **2019**, *21*, 3029.
(203) Ning, Z.; Xu, Z.; Liu, R.; Du, Z. *Synth. Commun.* **2021**, *51*, 3492.
(204) Colacino, E.; Benakki, H.; Guenoun, F.; Martinez, J.; Lamaty, F. *Synth. Commun.* **2009**, *9*, 1583.
(205) Liu, S.; Chen, R.; Guo, X.; Yang, H.; Deng, G.; Li, C.-J. *Green Chem.* **2012**, *14*, 1577.
(206) Adams, C. J.; Bedford, R. B.; Carter, E.; Gower, N. J.; Haddow, M. F.; Harvey, J. N.; Huwe, M.; Cartes, M. Á.; Mansell, S. M.; Mendoza, C.; Murphy, D. M.; Neeve, E. C.; Nunn, J. *J. Am. Chem. Soc.* **2012**, *134*, 10333.
(207) Yang, J.; Shen, G.; Chen, D. *Synth. Commun.* **2013**, *6*, 837.
(208) Agrawal, T.; Cook, S. P. *Org. Lett.* **2013**, *1*, 96.
(209) Hajipour, A. R.; Azizi, G. *Green Chem.* **2013**, *15*, 1030.
(210) Bisz, E.; Szostak, M. *Green Chem.* **2017**, *19*, 5361.
(211) Crockett, M. P.; Tyrol, C. C.; Wong, A. S.; Li, B.; Byers, J. A. *Org. Lett.* **2018**, *20*, 5233.
(212) Wu, W.-Y.; Wang, J.-C.; Tsai, F.-Y. *Green Chem.* **2009**, *11*, 326.
(213) Duong, H. A.; Wu, W.; Teo, Y.-Y. *Organometallics* **2017**, *36*, 4363.
(214) Prastaro, A.; Ceci, P.; Chiancone, E.; Boffi, A.; Cirilli, R.; Colone, M.; Fabrizi, G.; Stringaroe, A.; Cacchi, S. *Green Chem.* **2009**, *11*, 1929.
(215) Søbjerg, L. S.; Gauthier, D.; Lindhardt, A. T.; Bunge, M.; Finster, K.; Meyer, R. L.; Skrydstrup, T. *Green Chem.* **2009**, *11*, 2041.
(216) Zhou, C.; Wang, J.; Li, L.; Wang, R.; Hong, M. *Green Chem.* **2011**, *13*, 2100.
(217) Khalafi-Nezhad, A.; Panahi, F. *Green Chem.* **2011**, *13*, 2408.
(218) Bej, A.; Srimani, D.; Sarkar, A. *Green Chem.* **2012**, *14*, 661.
(219) Cacchi, S.; Caponetti, E.; Casadei, M. A.; Giulio, A. D.; Fabrizi, G.; Forte, G.; Goggiamani, A.; Moreno, S.; Paolicelli, P.; Petrucci, F.; Prastaroa, A.; Saladino, M. L. *Green Chem.* **2012**, *14*, 317.
(220) Huang, S.-H.; Liua, C.-H.; Yang, C.-M. *Green Chem.* **2014**, *16*, 2706.
(221) Puthiaraja, P.; Pitchumani, K. *Green Chem.* **2014**, *16*, 4223.
(222) Mandegani, Z.; Asadi, M.; Asadi, Z.; Mohajeri, A.; Iranpoor, N.; Omidvar, A. *Green Chem.* **2015**, *17*, 3326.
(223) Gautam, P.; Dhiman, M.; Polshettiwar, V.; Bhanage, B. M. *Green Chem.* **2016**, *18*, 5890.
(224) Yamada, M.; Shio, Y.; Akiyama, T.; Honma, T.; Ohki, Y.; Takahashi, N.; Murai, K.; Arisawa, M. *Green Chem.* **2019**, *21*, 4541.
(225) Mohazzab, B. F.; Jaleh, B.; Issaabadi, Z.; Nasrollahzadeh, M.; Varma, R. S. *Green Chem.* **2019**, *21*, 3319.
(226) Dhara, K.; Parasar, B.; Patil, A. J.; Dash, J. *Synth. Commun.* **2019**, *6*, 859.
(227) Blanco, M.; Mosconi, D.; Tubaro, C.; Biffis, A.; Badocco, D.; Pastore, P.; Otyepka, M.; Bakandritsos, A.; Liu, Z.; Ren, W.; Agnoli, S.; Granozzi, G. *Green Chem.* **2019**, *21*, 5238.
(228) Jammi, S.; Sakthivel, S.; Rout, L.; Mukherjee, T.; Mandal, S.; Mitra, R.; Saha, P.; Punniyamurthy, T. *J. Org. Chem.* **2009**, *74*, 1971.
(229) Sun, W.; Wang, Y.; Wu, X.; Yao, X. *Green Chem.* **2013**, *15*, 2356.
(230) Wang, K.; Yang, L.; Zhao, W.; Cao, L.; Sun, Z.; Zhang, F. *Green Chem.* **2017**, *19*, 1949.
(231) Guo, D.; Huang, H.; Zhou, Y.; Xu, J.; Jiang, H.; Chena, K.; Liu, H. *Green Chem.* **2010**, *12*, 276.
(232) Vaddula, B. R.; Saha, A.; Leazer, J.; Varma, R. S. *Green Chem.* **2012**, *14*, 2133.
(233) Al-Amin, M.; Akimoto, M.; Tameno, T.; Ohki, Y.; Takahashi, N.; Hoshiya, N.; Shutoa, S.; Arisawa, M. *Green Chem.* **2013**, *15*, 1142.
(234) Akiyama, T.; Taniguchi, T.; Saito, N.; Doi, R.; Honma, T.; Tamenori, Y.; Ohki, Y.; Takahashi, N.; Fujioka, H.; Sato, Y.; Arisawa, M. *Green Chem.* **2017**, *19*, 3357.
(235) Zolfigol, M. A.; Khakyzadeh, V.; Moosavi-Zare, A. R.; Rostami, A.; Zare, A.; Iranpoor, N.; Beyzavie, M. H.; Luque, R. *Green Chem.* **2013**, *15*, 2132.
(236) Steib, A. K.; Kuzmina, O. M.; Fernandez, S.; Flubacher, D.; Knochel, P. *J. Am. Chem. Soc.* **2013**, *135*, 15346.
(237) Karimi, B.; Mansouria, F.; Valib, H. *Green Chem.* **2014**, *16*, 2587.
(238) Baig, R. B. N.; Nadagouda, M. N.; Varma, R. S. *Green Chem.* **2014**, *19*, 4333.
(239) Rathi, A. K.; Gawande, M. B.; Pechousek, J.; Tucek, J.; Aparicio, C.; Petr, M.; Tomanec, O.; Krikavova, R.; Travnicek, Z.; Varma, R. S.; Zboril, R. *Green Chem.* **2016**, *18*, 2363.
(240) Mohammadinezhad, A.; Akhlaghinia, B. *Green Chem.* **2017**, *19*, 5625.
(241) Kaur, N.; Kaur, G.; Bhalla, A.; Dhaub, J. S.; Chaudhary, G. R. *Green Chem.* **2018**, *20*, 1506.
(242) Kazemnejadi, M.; Alavi, S. A.; Rezazadeh, Z.; Nasseri, M. A.; Allahresania, A.; Esmaeilpour, M. *Green Chem.* **2019**, *21*, 1718.

Environ Monit Assess (2023) 195:1457
https://doi.org/10.1007/s10661-023-12020-z

RESEARCH

A novel approach for forecasting PM2.5 pollution in Delhi using CATALYST

Abhishek Verma · Virender Ranga · Dinesh Kumar Vishwakarma

Received: 1 September 2023 / Accepted: 23 October 2023
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract Air pollution is one of the main environmental issues in densely populated urban areas like Delhi. Predictions of the PM2.5 concentration must be accurate for pollution reduction strategies and policy actions to succeed. This research article presents a novel approach for forecasting PM2.5 pollution in Delhi by combining a pre-trained CNN model with a transformer-based model called CATALYST (Convolutional and Transformer model for Air Quality Forecasting). This proposed strategy uses a mixture of the two models. To derive attributes of the PM2.5 timeline of data, a pre-existing CNN model is utilized to transform the data into visual representations, which are analyzed subsequently. The CATALYST model is trained to predict future PM2.5 pollution levels using a sliding window training approach on extracted features. The model is utilized for analyzing temporal dependencies in PM2.5 time-series data. This model incorporates the advancements in the transformer-based architecture initially designed for natural language processing applications. CATALYST combines positional encoding with the Transformer architecture to capture intricate patterns and variations resulting from diverse meteorological, geographical, and anthropogenic factors. In addition, an innovative approach is suggested for building input–output couples, intending to address the problem of missing or partial data in environmental time-series datasets while ensuring that all training data blocks are comprehensive. On a PM2.5 dataset, we analyze the proposed CATALYST model and compare its performance with other standard time-series forecasting approaches, such as ARIMA and LSTM. The outcomes of the experiments demonstrate that the suggested model works better than conventional methods and is a potential strategy for accurately forecasting PM2.5 pollution. The applicability of CATALYST to real-world scenarios can be tested by running more experiments on real-world datasets. This can help develop efficient pollution mitigation measures, impacting public health and environmental sustainability.

Keywords PM2.5 · Transformer · Forecasting · Delhi · Time-Series

A. Verma (✉) · V. Ranga · D. K. Vishwakarma
Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Bawana Road, Delhi -110042, India
e-mail: abhishekcms08@gmail.com

V. Ranga
e-mail: virenderranga@dtu.ac.in

D. K. Vishwakarma
e-mail: dvishwakarma@gmail.com

Introduction

The deleterious effect of environmental human health effects of pollution, animals, and plants renders it a pressing global concern. Increasingly, air pollution

Published online: 11 November 2023

is a prominent problem among the different types of pollution, especially in urban regions undergoing swift economic expansion, industrialization, and urbanization. The decline in air quality presents significant hazards to human welfare, resulting in respiratory illnesses and other health complexities. Additionally, it impedes the achievement of environmental sustainability and societal advancement. The issue has prompted significant interest in environmental conservation, governmental policy-making, and individual health maintenance, focusing on developing accurate air quality prediction models. In the following sub-sections, the impact of sub-sectional pollution is discussed along with the importance of air quality forecasting followed by human activities and air pollution in Delhi.

Impact of environmental pollution and health risks

Environmental pollution is a widespread issue with wide-ranging effects that hurt populations of people, animals, and plants. Examining the consequences of air, water, and land/soil waste contamination is vital because of pollution's considerable influence on human health (Khan & Ghouri, 2011). The swift progress of the economy, industrialization, and air pollution has significantly increased due to urbanization in numerous urban areas worldwide (Darçın, 2014). The circumstance mentioned above has presented significant peril to human beings' physical and mental welfare, the ecosystem, and the enduring advancement of civilization. The escalation and amplification of industrialization and urbanization have resulted in an increasingly severe problem of atmospheric pollution. The circumstance above presents a special peril to human welfare, given its correlation with diverse respiratory afflictions such as bronchial asthma, chronic pharyngitis, and chronic bronchitis. (Chang et al., 2020). The asthma condition has been studied by Chang et al. (2020). In addition, significant air pollution can give rise to a haze, which can cause reduced visibility in the lower atmosphere, traffic accidents, flight delays, and other related consequences. In recent years, the importance of achieving precise air quality forecasting has garnered significant attention, owing to its relevance in environmental conservation (X. Liao et al., 2015), governmental decision-making (Zheng et al., 2015), and the daily health of individuals (Ha Chi & Kim Oanh, 2021).

Remote sensing is of utmost significance within the framework of the United Nations' 2030 Agenda for Sustainable Development (SDG-UN 2030). Remote sensing, due to its ability to gather scientific and environmental data across large areas through the use of satellites, light aircraft, or unmanned aerial vehicles (UAVs), is recognized as a crucial tool for obtaining vital information about our environment. This is especially relevant in areas where there are limited networks for monitoring air pollution on the ground. The distinctive perspective offered by remote sensing enables us to perceive and analyze the ecological consequences of human actions in unprecedented ways. This skill is in perfect harmony with the overarching goals of the Sustainable Development Goals (SDGs), which aim to tackle urgent global issues such as poverty, inequality, climate change, and environmental degradation. The significance of remote sensing in enhancing our understanding and preparedness for environmental concerns is emphasized by this alignment. Through the incorporation of these valuable ideas and methods into our research endeavors and the wider scientific community, we actively contribute to the collaborative pursuit of achieving sustainable development and ensuring the preservation of our planet for future generations (C. A. Varotsos & Cracknell, 2020).

Many urban areas have established air quality monitoring stations within metropolitan regions to observe the immediate concentrations of PM2.5 and other atmospheric pollutants, such as PM10, CO, O3, NO2, SO2, and others. Air pollution harms human populations and the natural environment across all global regions (L. Wang et al., 2020)

Importance of air quality forecasting

Air pollution is a significant factor in developing hazardous cardiovascular and respiratory diseases, with delicate particulate matter (PM2.5) being a primary culprit (Chu et al., 2021). The inhalation of PM2.5 significantly contributes to premature death in numerous urban areas. (Dai et al., 2014) ahead of time, authorities must issue concerns about excessive levels of PM2.5 in the nearby air to allow sensitive people to take safeguards (Varotsos et al., 2021). To achieve this goal, persistent forecasting of PM2.5 air concentrations is necessary for both the short-term (a few hours ahead) and the medium-term (a few days early) periods. Despite this, predicting PM2.5

concentrations is challenging since PM2.5 concentrations vary substantially over time and space and are affected by a wide range of elements related to weather, geography, and anthropology, some of which result in abrupt changes in PM2.5 concentrations (Liou et al., 2020). It is easier to choose measures that, if one is aware of the medium- and long-term trends in PM2.5 concentrations at specific sites, will improve air quality over time and lower the possible health risks related to poor air quality, as well as the primary contributing variables (Tian et al., 2021). Over time, monitoring the air quality helps detect anthropogenic factors that harm air quality at local levels. Some examples of these causes are road traffic, construction projects, and industrial expansion (Zhu et al., 2019). A dispatching strategy for power systems that optimizes multiple objectives in a high-dimensional space. The strategy focuses on the spatial and temporal distribution of various pollutants, such as volatile organic compounds (VOCs), sulfur dioxide (SO2), and nitrogen dioxide (NO2). Through the integration of a comprehensive model that accounts for pollutant dispersion in thermal power plants, along with the consideration of environmental capacity, this study successfully achieves significant reductions in air pollutant concentrations. Specifically, the study demonstrates a noteworthy decrease of 27.5% in air pollutant concentrations on days characterized by high levels of pollution (Dai, Huang, & Zeng, 2023b).

Human activities and air pollution

Sources of PM2.5 caused by human activity include using fossil fuels in various transportation modes, generating electricity, heating and cooling indoor spaces, and several industrial-scale manufacturing activities. On the other hand, significant sources of PM2 include burning crop leftovers, partially caused forest fires, and dust storms brought on by deforestation, regularly 5 inputs in certain places (Cusworth et al., 2018). Air pollutant buildup and dissipation are impacted by complicated atmospheric conditions that are challenging to identify in traditional meteorological data measures, such as thermal updrafts and periodic temperature inversion layers. These atmospheric conditions affect the accumulation and dissipation of air contaminants (Murthy et al., 2020). Significant weather occurrences that occur sporadically and without warning throughout severe seasonal meteorological cycles result in abrupt PM2.5 air concentrations that have changed influence in many places. Despite this varied range of impacts, several communities have significantly improved air quality and lowered the levels of PM2.5 (Fowler et al., 2020).

Delhi, situated within the Indian Geography Plain (IGP), has an estimated population of approximately 16.5 million. As per the report by the Registrar General of India in 2011, the National Capital Region of Delhi, comprising regions of Haryana, Uttar Pradesh, Rajasthan, and Delhi, is estimated to have a populace exceeding 46 million. The average daily levels of surface particulate matter (PM2.5) pollution in Delhi frequently surpass the unhealthy air threshold established by the World Health Organization, which is a 24-h average of 25 μg m^{-3}, as well as the daily mean point designated by the Indian Central Pollution Control Board (CPCB), which is 60 μg m^{-3}. Throughout the year, PM2.5 standards in Delhi are surpassed, with a yearly average PM2.5 concentration exceeding 100 μg m^{-3} (Tiwari et al., 2013). Ambient PM2.5 concentrations are prone to significant episodic surges during post-monsoon season (October–November). Anthropogenic pollution has been linked to many health problems in Delhi (Dey et al., 2012; Guttikunda & Jawahar, 2014; Nagpure et al., 2014). Between 2000 and 2010, the air quality in Delhi deteriorated, and researchers (Nagpure et al., 2014) predicted a 60% rise in mortality. Air pollution–related illness is 12 times more common in Delhi than in the rest of India (Kandlikar & Ramachandran, 2000). A crucial inquiry pertains to how the smoke discharges resulting from agricultural fires in rural areas after the monsoon season affect the pre-existing elevated levels of urban air pollution in the Indo-Gangetic Plain. This research aims to evaluate the extent to which fire emissions contribute to PM2.5 pollution in Delhi during the post-monsoon burning season from 2012 to 2016. Identifying the source of surface PM2.5, whether fires or other anthropogenic factors cause it, is of utmost importance in devising effective measures to mitigate pollution exposure.

Long-term air quality monitoring

Numerous different PM2.5 forecasting approaches are used and suggested, utilizing various data sources and methodologies. Both short-term predictions, up to a few hours in the future, and medium-to-long-term

forecasts, out to days or months, respectively, are often the focus of the methodologies. The medium- to long-term predictions frequently use a variety of data-driven deterministic, stochastic, and statistical approaches to analyze historical data and are based on patterns of several exogenous variables observed at ground-level weather and air-quality recording stations (Chu et al., 2021; Liao et al., 2020). A dispatching strategy for power systems that optimizes multiple objectives in a high-dimensional space. The strategy focuses on the spatial and temporal distribution of various pollutants, such as volatile organic compounds (VOCs), sulfur dioxide (SO2), and nitrogen dioxide (NO2) (Gariazzo et al., 2020). Through the integration of a comprehensive model that accounts for pollutant dispersion in thermal power plants, along with the consideration of environmental capacity, this study successfully achieves significant reductions in air pollutant concentrations. Specifically, the study demonstrates a noteworthy decrease of 27.5% in air pollutant concentrations on days characterized by high levels of pollution (Dai, Huang, & Zeng, 2023b).

The enduring presence of long-range correlations in Sahel precipitation anomalies over the period from 1900 to 2010, as well as their possible implications for precipitation forecasting, highlights the significance of integrating historical data patterns into forecasting models, providing vital insights to improve our comprehension and readiness for precipitation-related environmental concerns (Efstathiou & Varotsos, 2012). The accuracy of such PM2.5 forecasting algorithms is significantly influenced by the variables included as predictors. Physio-chemical approaches are also used to simulate the dynamics of the atmosphere (Lv et al., 2016). The study proposes a novel PCA-MEE-ISPO-LightGBM model to evaluate the risks associated with urban haze, considering multiple dimensions of sensitivity, component hazards, and vulnerability. The findings of this study indicate that the implementation of a comprehensive approach leads to a notable enhancement in the accuracy of evaluating haze risk in urban areas, in comparison to solely considering causative factors. Consequently, this comprehensive approach provides a more realistic and dependable assessment of haze risk (Dai et al., 2022). Nevertheless, owing to the numerous uncertainties, these methods are often less accurate than those of data-driven approaches. Aerosol optical depth (AOD) and its almost linear association with ground-level PM2.5 (Lee et al., 2011) are two satellite data points that are increasingly used (Hoff & Christopher, 2009). As of late, AOD has also been employed in some regions to supplement meteorological data variables in some models and to produce real-time, short-term PM2.5 forecasts (Zhao et al., 2018). (Zamani Joharestani et al., 2019) Several models created to forecast PM2.5 in the short run concentrate on the past univariate PM2.5 collected data at particular places. In combination with other models, statistical models, such as the autoregressive integrated moving average model (ARIMA) and its components (AR and MA), are frequently utilized (L. Wang et al., 2020; Zhao et al., 2018). Using multiple linear regression (MLR) and ARIMA, some models demonstrate correlations between PM2.5 historical trends and a few variables (Wei et al., 2021). However, the fickle nature and sharp peaks that characterize PM2.5 data trends present problems for these models and force them to concentrate on residual trends or highly pre-processed data. Another helpful technique to extract information from historical PM2.5 time-series trends is wavelet decomposition, transformation, and power spectrum analysis (Chen et al., 2020; Graupe et al., 1975; Feng et al., 2015), sometimes in combination with fractal analysis (L. Wang et al., 2020). To deal with geographical and temporal fluctuations, several methods combine this with cluster analysis (Mahajan et al., 2018), adaptive iterative forecasting techniques (Liou et al., 2020), and ARMA (Fan, 2017). A dispatching strategy for power systems that optimizes multiple objectives in a high-dimensional space. The strategy focuses on the spatial and temporal distribution of various pollutants, such as volatile organic compounds (VOCs), sulfur dioxide (SO2), and nitrogen dioxide (NO2). Through the integration of a comprehensive model that accounts for pollutant dispersion in thermal power plants, along with the consideration of environmental capacity, this study successfully achieves significant reductions in air pollutant concentrations. Specifically, the study demonstrates a noteworthy decrease of 27.5% in air pollutant concentrations on days characterized by high levels of pollution (Dai et al., 2023a, b)

Novelty of the approach

The field of air quality forecasting is of utmost importance in academic research due to its wide-ranging implications for both public health and environmental

management. This study introduces a novel methodology, referred to as CATALYST (Convolutional and Transformer model for Air Quality Forecasting), which presents a notable progression in the domain of air quality prediction. Our study presents a comprehensive theoretical framework that incorporates state-of-the-art methodologies and tackles the urgent requirement for precise and effective air quality prediction.

The novelty of our paper is evident in several crucial aspects:

- Integration of Convolutional Neural Networks (CNNs) and Transformer Networks: In contrast to conventional methodologies that tend to concentrate on a singular form of neural network, our approach leverages the capabilities of both Convolutional Neural Networks (CNNs) and Transformer Networks (Ragab et al., 2020; Vaswani et al., 2017). The integration of short-term and long-term contextual feature extraction enables a more comprehensive comprehension of air quality patterns.
- Efficient computational framework: The primary focus is on achieving computational efficiency while maintaining high levels of prediction accuracy. The utilization of a single encoder and a linear layer in our model results in a streamlined processing pipeline, leading to a notable reduction in the computational workload. The approach demonstrates practicality in real-time and large-scale applications.
- In-depth data pre-processing: The model places significant emphasis on the significance of data pre-processing techniques, specifically the utilization of sliding windows and positional encoding. These techniques not only contribute to the improvement of prediction accuracy but also enhance the efficiency of data processing, resulting in time and computational resource savings.
- Performance evaluation: The proposed methodology is conducted using a comprehensive framework that incorporates various metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). This rigorous evaluation approach allows for a thorough assessment of the methodology's performance. The comparative analysis conducted against state-of-the-art models demonstrates the enhanced predictive accuracy and reliability of CATALYST.

In brief, the model presents an innovative methodology that combines the advantages of Convolutional Neural Networks (CNNs) and Transformer Networks, with a focus on enhancing computational efficiency and emphasizing meticulous data pre-processing. The distinctive amalgamation of these constituents distinguishes our methodology, rendering it a significant contribution to the domain of air quality prediction. The objective of this paper is to offer a thorough comprehension of our innovative methodology and its potential implications for enhancing air quality forecasting systems.

Literature review

Air pollution is a significant problem affecting millions of people's quality of life globally. The World Health Organization (WHO) estimates that air pollution causes 7 million premature deaths each year, with particulate matter (PM2.5) being one of the most harmful pollutants (Puthumana et al., 2021). The investigation concerning historical data emphasizes the importance of comprehending the patterns of air pollution over a span of time, providing valuable insights into the progression of air quality in densely populated urban areas (Varotsos and Cartalis, 1991). Predicting PM2.5 concentrations accurately is critical for mitigating air pollution's adverse effects. Researchers have been exploring various machine learning models to estimate PM2.5 concentrations accurately. In recent years, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been the most popular models used (L. Li, 2020; Liu et al., 2020; Shi et al., 2020; Yao et al., 2021; Zhang, 2020). This literature review aims to provide an overview of the recent research on predicting PM2.5 concentrations using machine learning models. CNNs, extensively employed in image processing, have been utilized in studies to forecast air quality. (Liu et al., 2020) This paper outperformed standard methods such as ARIMA (Jian et al., 2012) by utilizing a pre-trained CNN model to anticipate PM2.5 concentrations using meteorological data. Similarly, Yao et al. (2021) employed a CNN-based model with multi-scale and multi-channel characteristics. Li et al. (2019) extracted spatial data from PM2.5 concentration maps and captured temporal features using an LSTM network (Hochreiter & Schmidhuber, 1997; H. Li et al., 2019).

Both models outperformed traditional approaches such as SVR and ANN (Agarwal et al., 2020). RNNs, which can model sequential data, have been used to forecast time series. Zhang et al. (2020) predicted PM2.5 concentrations using a hybrid model that combined a GRU and a CNN (Zhang, 2020), surpassing standard models such as ARIMA and SVM. The study by Shi et al. (2020) utilized meteorological data to train an attention-based Recurrent Neural Network (RNN) model. The study's results indicated that the attention-based RNN model's performance surpassed conventional methods such as Autoregressive Integrated Moving Average (ARIMA) and Random Forest (RF). Hybrid models have also been proposed, combining machine learning techniques with classical statistical methodologies. Liu et al. (2020) created a hybrid model combining a CNN and a seasonal decomposition approach, outperforming classic methods like ARIMA and SVM (Liu et al., 2020). Wang et al. (2019) suggested a hybrid model that beat established procedures such as ARIMA and SVM by combining a wavelet transform with a multivariate adaptive regression splines (MARS) approach (Niu et al., 2021). Incorporating additional data sources, such as meteorological and traffic data, has been demonstrated to enhance PM2.5 forecasting accuracy. Feng et al. (2015) used meteorological and traffic data to estimate PM2.5 concentrations in the Beijing–Tianjin–Hebei region, outperforming established methods such as ARIMA and SVM. P. Wang et al. (2019) projected PM2.5 concentrations using meteorological data and a MARS technique, surpassing established methods such as ARIMA and SVM. Anthropogenic factors, air quality, and PM2.5 concentrations are also influenced by anthropogenic causes such as industrial growth and transportation. The study by Zhu et al. (2019) aimed to examine the influence of industrial emissions on the concentrations of PM2.5 in China. The study's findings revealed that the high levels of PM2.5 in the country were primarily attributed to the emissions from the industrial sector. Similarly, transportation emissions from vehicles such as cars and trucks are also a significant source of air pollution, particularly in urban areas with high traffic congestion. In addition, land use changes such as deforestation, urbanization, and agricultural practices can also contribute to air pollution. Deforestation can potentially elevate the levels of dust and smoke in the atmosphere, whereas urbanization may result in amplified emissions from industrial and transportation activities. Ultimately, climate change can exert a substantial influence on the quality of the air. Elevated temperatures have the potential to augment the occurrence and severity of wildfires, resulting in escalated concentrations of smoke and particulate matter within the atmosphere. The phenomenon of climate change has the potential to cause alterations in precipitation patterns, thereby influencing the dispersion of pollutants in the atmosphere and their conveyance to diverse geographical locations.

The proposed model in this research paper offers several advantages over the previous studies:

- Integration of CNN and Transformer: In contrast to prior literature that predominantly concentrated on either Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) architectures, the present study amalgamates the respective advantages of both models. The proposed model achieves a more comprehensive input time-series data analysis by integrating a pre-trained CNN for feature extraction and a Transformer Network for capturing long-term contextual information. Incorporating this integration results in enhanced precision in forecasting PM2.5 levels instead of the models that exclusively depend on CNN or RNN.
- Enhanced prediction accuracy: The model under consideration exhibits enhanced performance compared to prior methodologies. It achieves a lower RMSE of 21.01 and MAE of 13.37, indicating its ability to make more precise predictions. Additionally, the coefficient of determination (R^2) value of 0.83 highlights the model's explanatory solid power in capturing the variability of PM2.5 concentrations. These improved accuracy metrics demonstrate the superiority of the proposed model over traditional methods such as SVR, ARIMA, LSTM, and other hybrid models (Verma et al., 2023).
- Computational efficiency: The model architecture prioritizes computational efficiency while ensuring the preservation of prediction accuracy. Adopting a solitary encoder and a linear layer reduces model intricacy, thereby accelerating processing times in both the training and inference stages. The computational efficiency of the proposed model makes it better suited for real-time

applications or large-scale implementation, distinguishing it from previous models that may require more computational resources.

- Comprehensive methodology: The scholarly article outlines a thorough approach that encompasses the gathering of data, the preparation of data for analysis, and the assessment of the model's effectiveness. The effectiveness of the proposed model in capturing spatial and temporal dependencies is attributed to its consideration of multiple monitoring stations and incorporation of techniques such as sliding windows and positional encoding. The incorporation of a range of evaluation metrics, such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and *R*-squared (R^2), offers a thorough evaluation of the model's efficacy. The proposed model is deemed to possess enhanced reliability and applicability compared to prior research due to its holistic approach.

Study area and dataset

This section presents a comprehensive overview of the study area, Delhi, encompassing its geographic coordinates, demographic statistics, climatic conditions, and atmospheric pollution levels. The text delineates the dataset utilized in the research, encompassing hourly PM2.5 data from five air quality monitoring stations spanning the period from November 2016 to March 2023. The data underwent meticulous cleaning and pre-processing procedures to guarantee its quality and reliability for the research investigation.

Study area—Delhi

The National Capital Territory (NCT) of Delhi is located in the northern region of India, with geographical coordinates of north latitudes 28° 0.21′ and 28° 0.53′ and east longitudes 76° 0.20′ and 76° 0.37′. The altitude of this region is 216 m above the mean sea level. The National Capital Region (NCR) is a noteworthy area of industrial activity. As per the Indian census, the Capital Region of India's central and peripheral urban areas collectively accommodate an estimated population of 46 million individuals. The climatic conditions of the area are classified as subtropical semi-arid (steppe), featuring an annual mean precipitation of 611.8 mm and an annual mean temperature of 31.5 °C. The air quality in Delhi, situated in the Northern Hemisphere, is notably poor owing to its unique topographical features, weather patterns, and significant urban development. In addition, it is worth noting that PM2.5 concentrations experience a sudden surge due to natural sources, such as dust originating from the Thar Desert and its environs. On average, the urban area encounters 15 instances of pollution annually and over 200 days of PM2.5, surpassing the NAAQS standards.

As per the World Health Organization's report of 2018, the levels of PM2.5 in Delhi are significantly elevated in comparison to other cities across the globe. Numerous epidemiological studies have determined that extended exposure to fine particulate matter (PM2.5) in ambient air pollution has played a substantial role in exacerbating chronic respiratory diseases and causing mortality in Delhi.

The hourly PM2.5 dataset from 3rd November 2016 to 23rd March 2023 (Fig. 1) for Delhi city is compiled from the CPCB. It is recorded from five different air quality monitoring stations. The data is obtained from DPCC, IMD, and CPCB stations installed in various locations, as mentioned in Fig. 2. The dataset was split randomly into training (70%) and testing subsets (30%) areas. It is one of the largest urban areas in the North Indian region of India. In the present research study, a thorough evaluation of the dataset was conducted using a diverse range of monitoring systems. Specifically, five different monitoring systems are utilized, including DTU Delhi-CPCB, NSUT Dwarka-CPCB, Anand Vihar-DPCC, Pusa Delhi-IMD, and IGI-T3 Delhi-IMD. Each monitoring system was essential in generating valuable data, which is analyzed and processed to extract meaningful insights. The dataset under consideration is comprised of mean values, and any absent data points are substituted with the mean value of the corresponding attribute. Furthermore, the dataset is subjected to various data pre-processing methods to address any data quality concerns or anomalies prior to commencing the analysis. This step is pivotal in guaranteeing the data's utmost quality and the research outcomes' precision and dependability. The utilization of monitoring systems and meticulous data pre-processing techniques served to establish a robust and superior foundational dataset for the research study.

$$M_{\mathrm{i}} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} \tag{1}$$

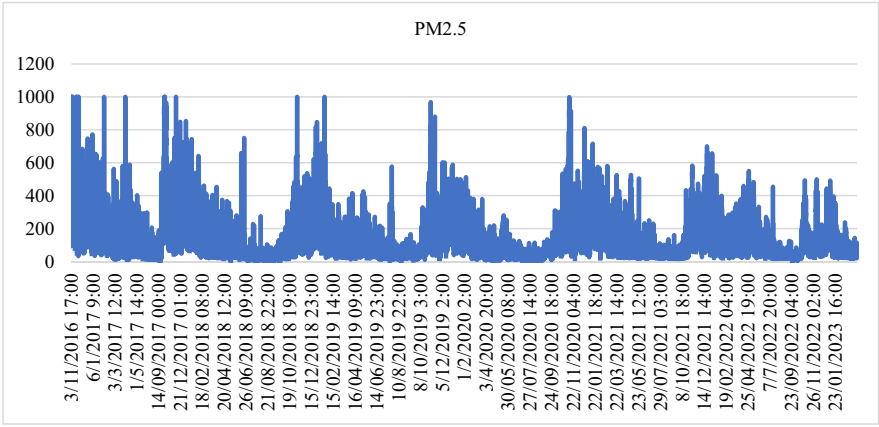


Fig. 1 Cleaned time-series data for PM2.5 pollutant level in Delhi

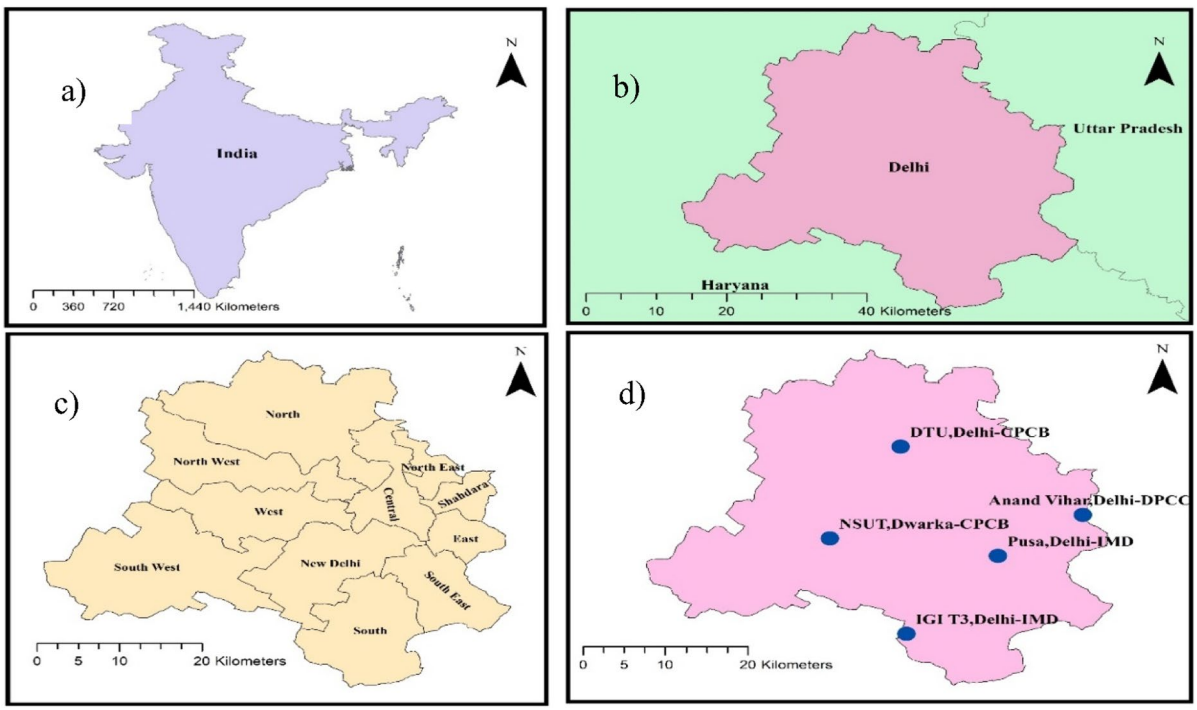


Fig. 2 Study area at different geographical scales. **a** Country level, **b** state level, **c** district level, and **d** selected monitoring stations

If $x_i \rightarrow \phi$, then replace x_i with M_{i-1} (2)

Dataset analysis

The dataset has been thoroughly analyzed using the Jupyter Notebook and the Python library Seaborn and Matplot to make a graph explaining the trend followed by PM2.5 in the National Capital Region. The dataset is further broken into five seasons by defining months in which the seasons occur and ruining the whole dataset into five categories. This section is part of a research paper that aims to analyze the trends of PM2.5 concentration in our study area.

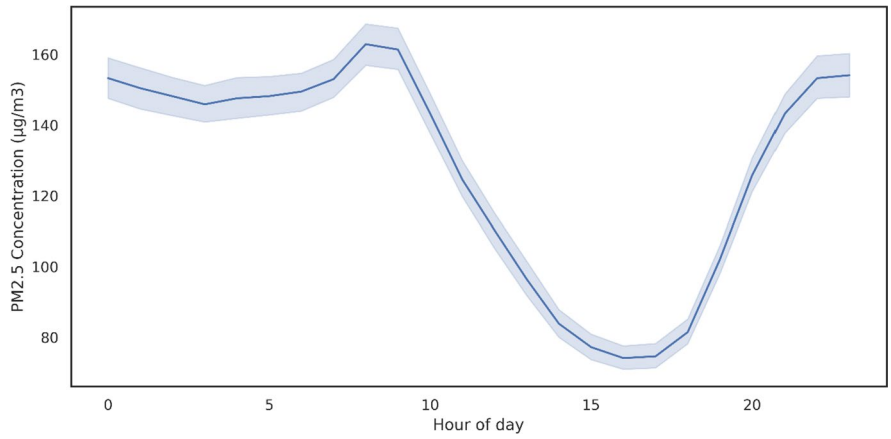


Fig. 3 Line plot of trend on hourly bases

The analysis does exploratory data analysis (EDA) on a dataset containing information about PM2.5 concentration. The dataset is imported using pandas, and the necessary packages (NumPy, Matplotlib, and Seaborn) are imported. The data is then cleaned and transformed to make it suitable for analysis.

The EDA starts with plotting PM2.5 concentration over time, which shows the overall trend of the data. The experiment then creates time-series features from the date–time index, which includes the hour, day of the week, month, year, and quarter. The seasons are also added to the dataset.

The study employs boxplots as a graphical tool to illustrate the correlation between the concentration of PM2.5 and both the day of the week and the season. These plots reveal that winter and spring have the highest PM2.5 concentration, while the monsoon season has the least. Weekends have less PM2.5 than weekdays. The analysis in Fig. 3 indicates a noticeable disparity in PM2.5 attention between weekdays and weekends. This can be attributed to the reduced traffic and industrial activity during weekends, as many industries remain closed on Saturdays and Sundays or operate at a reduced capacity. Figures 4 and 5 season-wise PM2.5 concentration tell that the analysis elucidates the phenomenon of a significant decrease in concentration levels during the monsoon season, in contrast to a substantial increase during the winter season. The primary contributing factor to this trend is the heightened usage of coal and wood for heating purposes during winter. However, a more pertinent explanation is that the air tends to mix over a reduced surface area during winter compared to summer. The following data represents the mean annual air mixing height. During the summer, atmospheric air mixing occurs at a relatively high altitude. This implies that the pollution generated by human activities is dispersed across a wider geographical region. During the winter, the atmospheric air tends to remain closer to the Earth's surface, increasing exposure to vehicular emissions and other pollutants and respiratory discomfort.

The analysis also creates line plots to show the trend of PM2.5 concentration over the months, seasons, and hours. These plots help to identify the hourly trends in PM2.5 concentration. The level of PM2.5 attention is subject to variability based on various factors, including wind velocity, moisture content, temperature, and anthropogenic actions such as vehicular traffic, industrial discharges, and residential heating. The causative factors behind the low levels of PM2.5 in the daytime and its high concentration at night, as you can observe in Figs. 3, 6 and 7 may differ depending on the geographical region and time of the year.

Typically, the atmospheric boundary layer exhibits greater height during daylight hours, resulting in enhanced dispersion of pollutants. Furthermore, diurnal variations in solar radiation and temperature can facilitate the production of secondary organic aerosols, which can subsequently disperse and mitigate the levels of PM2.5 in the ambient atmosphere. In contrast, reducing the boundary layer height during night-time results in decreased dispersion of pollutants and increased accumulation of PM2.5 in the atmosphere. Lower temperatures and reduced sunlight can impede the creation of secondary organic aerosols, resulting in elevated levels of PM2.5.

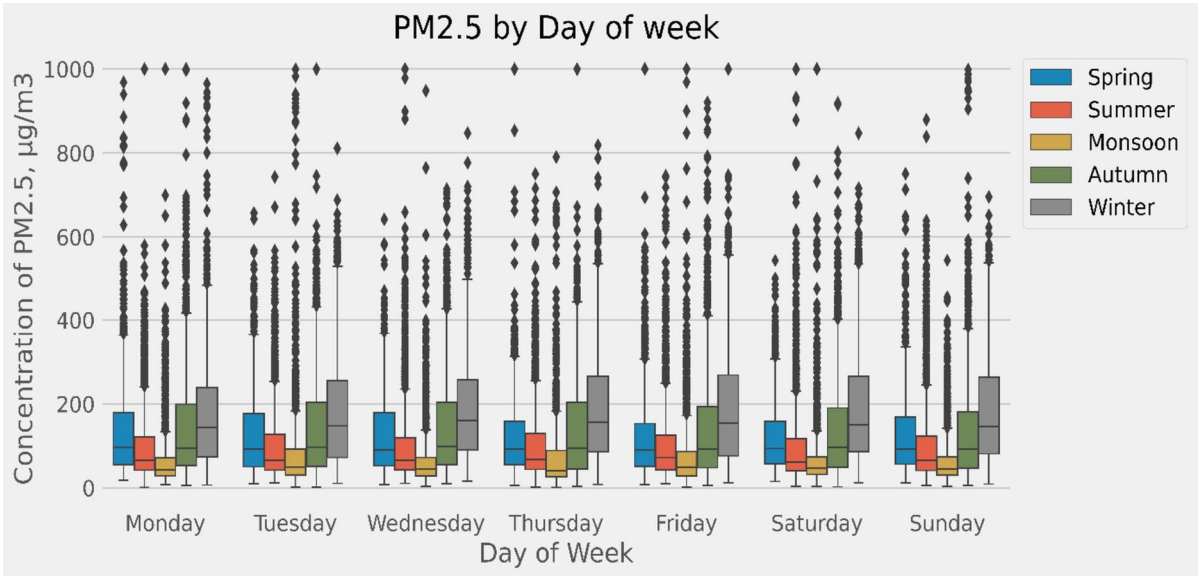


Fig. 4 Boxplot visualization of data in terms of season and days

As previously stated, the winter season, specifically December, January, and February, demonstrates the most significant concentrations of PM2.5, as depicted in Figs. 3, 6 and 8. There are several scientific justifications for this rationale, as enumerated below.

- Temperature inversion**:** It is a phenomenon that occurs frequently during the winter season. The phenomenon occurs when a thermal inversion is present, wherein a layer of relatively warm air overlays a layer of relatively cool air, leading to the confinement of pollutants in the lower atmospheric region. Consequently, the PM2.5 and other contaminants' concentrations escalate, resulting in elevated pollution levels.
- Atmospheric stability**:** During winter, atmospheric stability increases, accumulating pollutants in the lower atmosphere. A stable atmosphere inhibits the dispersion of pollutants, resulting in elevated levels of PM2.5.
- Human activities: The winter tends to involve a heightened usage of heating and cooking appliances, thereby, increasing the emission of PM2.5 and other pollutants. Furthermore, the augmented utilization of automobiles and industrial operations throughout this period also adds to elevated concentrations of PM2.5.

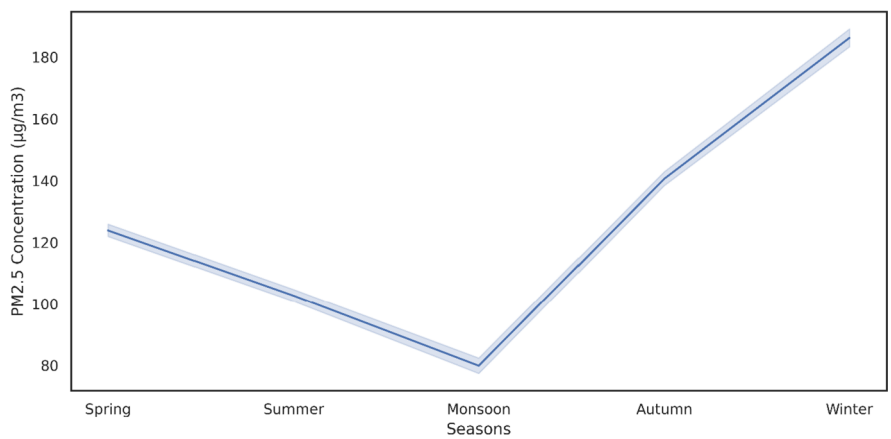


Fig. 5 Season-wise PM2.5 concentration

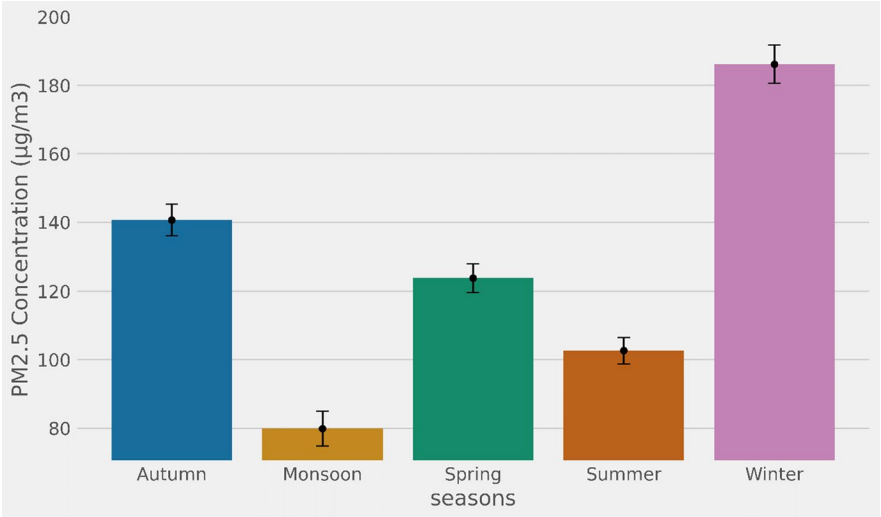


Fig. 6 Season-wise bar graph

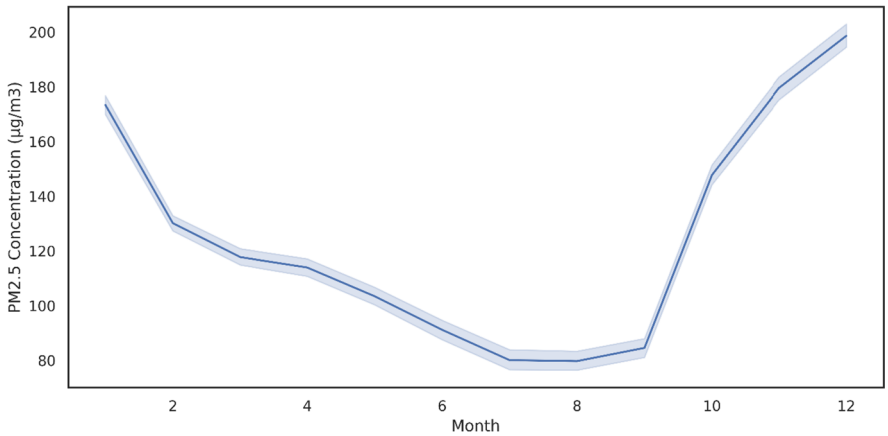


Fig. 7 Month-wise trend of PM2.5 concentration

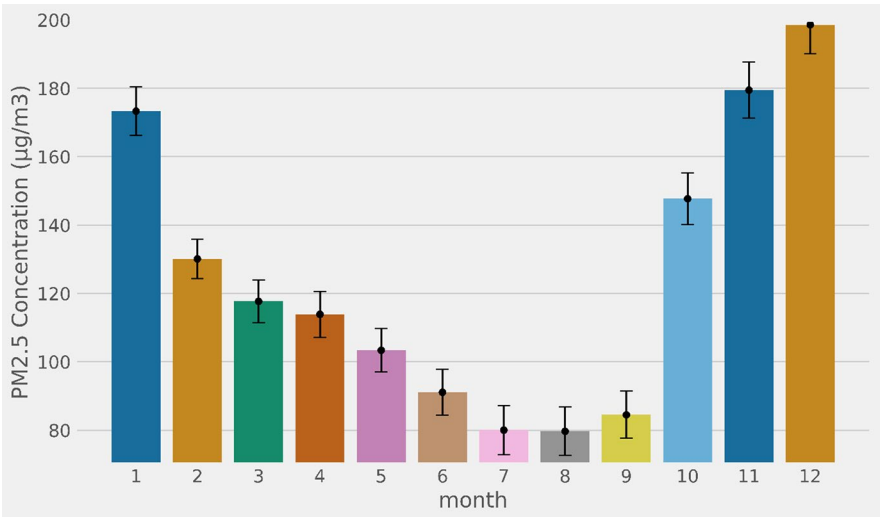


Fig. 8 Bar graph of month-wise trend PM2.5 concentration

In general, the confluence of these variables renders the winter season, particularly December, the apex period for elevated PM2.5 concentrations, as seen in Fig. 6.

Overall, this EDA is essential for data analysis as it helps to visualize and analyze the data, providing insights into the trend of PM2.5 concentration. The technique tells various patterns in the data that can be further explored as the causes of high PM2.5 concentration in the area.

Proposed methodology

This section provides a detailed discussion of the proposed methodology. The proposed methodology consists of three main phases. The first stage of the study entails the collection and pre-processing of obtained data from five distinct monitoring stations situated in Delhi, namely DTU Delhi-CPCB, NSUT Dwarka-CPCB, Anand Vihar-DPCC, Pusa Delhi-IMD, and IGI-T3 Delhi-IMD. Subsequently, the gathered information is processed to facilitate subsequent examination. The following phase involves the utilization of a pioneering transformer-based methodology to predict air quality. The data that has been gathered and undergone pre-processing from the monitoring stations are utilized as input for the Transformer model. By using its attention mechanism, the Transformer model can effectively capture extended dependencies and acquire intricate patterns within the data, ultimately facilitating precise predictions regarding air quality.

During the third stage, the outcomes derived from the predictive model are evaluated and scrutinized through the application of performance metrics, including but not limited to mean absolute error (MAE), coefficient of determination (R^2), and root mean square error (RMSE). The present study examines the precision and efficacy of the suggested approach in forecasting atmospheric conditions. The proposed methodology integrates various components such as data collection, preprocessing, a unique transformer-based approach, and result analysis to offer a comprehensive framework for air quality prediction (Fig. 9).

Algorithm 1 shows the algorithm of the proposed architecture. The algorithm's objective is to facilitate the training and prediction of air quality by utilizing the CATALYST model. The system receives input in the form of data points obtained from monitoring stations and uses this information to forecast air quality values. The iterative process of the algorithm is executed for a predetermined number of epochs. The process involves converting the input data into a vector image and extracting features through a Convolutional Neural Network (CNN). The Transformer block is used to extract enduring characteristics. The application of dropout is observed in the convolutional neural network features, while the Transformer features are subjected to both dropout and batch normalization. The final predicted output is derived by concatenating the enhanced features.

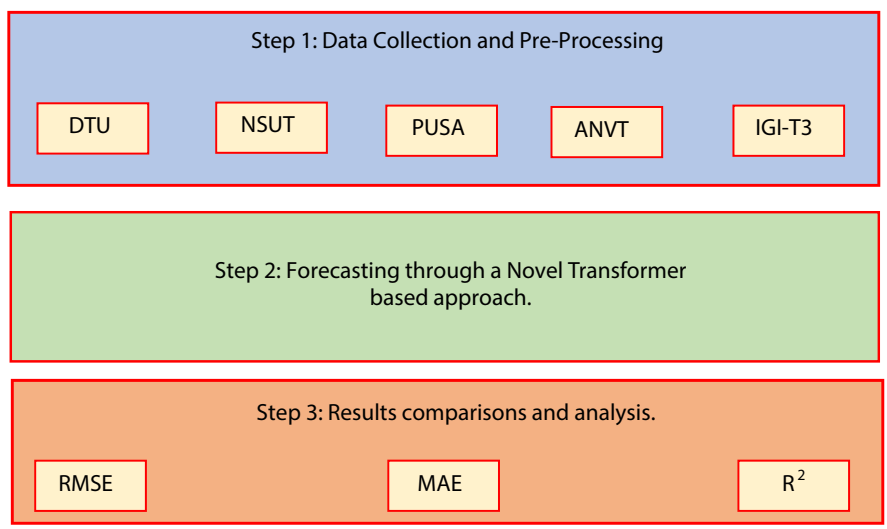


Fig. 9 Steps for PM2.5 forecasting

Aim: To learn a mapping function $F:(X_j,Y_j)\longrightarrow$ from data points obtained from five monitoring stations

Input: Set of data points $X_j = \{x_{j1},x_{j2},\ldots\ldots\ldots,x_{jSW}\}$ where, SW is the size of the sliding window used for input layer partitioning

Output: Air Quality Prediction, $Y_j \in$ forcasted value

1. $D = \{(X_0,Y_0), (X_1,Y_1),\ldots\ldots\ldots\ldots,(X_B,Y_B)\}$ represents the dataset consisting of hourly data points X and Y as the forecasted value with B number of blocks into which the data is partitioned for both input and label data.
   for $E\leftarrow 1$ to *Epochs* do
2. Input data X_j is converted into a vector image.
3. $F(X_j)$ be the feature representation obtained by CNN from vector image.
4. $T(F(X_j))\leftarrow F(X_j)$ be the long-term features obtained from Transformer block.
5. $D(F(X_j))\leftarrow F(X_j)$ improve model's performance by applying dropout to feature representation obtained by CNN.
6. $B(D(F(X_j)))\leftarrow T(F(X_j))$ improve model's performance by applying dropout and batch-normalization to feature representation obtained by transformer.
7. $P(X_j) = D(F(X_j)) \oplus B(D(F(X_j)))$ concatenation of both the enhanced features to obtain forecasted output.
   end

Algorithm 1 Training and prediction algorithm for CATALYST

Data pre-processing

This research work employs a dataset consisting of 48,362 hourly data points that have undergone a process of cleaning and organization. The above data were obtained from five monitoring stations and merged into a unified column. To facilitate analysis, the data underwent a process of partitioning into blocks encompassing input and label data. Utilizing the sliding window technique in the input layer boosted the streamlining of data processing by decreasing nested loops and consolidating them into a singular circle. This methodology conserves both time and computational resources. Furthermore, using the positional encoding layer facilitated the integration of sinusoidal positional embedding with the input data. The process involves converting the PM2.5 data into a vector image, which is subsequently utilized as input for a pre-existing Convolutional Neural Network (CNN) to extract features.

Short-term contextual feature learning using CNN

So, in the proposed methodology, input is fed as an image to a pre-trained Convolutional Neural Network (CNN). That is employed for feature extraction from the input time series before their integration into the Transformer. Utilizing a pre-existing Convolutional Neural Network (CNN) for feature extraction from PM2.5 data is a viable methodology because CNNs are tailored toward image-processing tasks and can proficiently capture spatial patterns and dependencies within the data. By converting PM2.5 data into a vector image, the convolutional neural network (CNN) can extract pertinent features from the data while decreasing its dimensionality. This facilitates the subsequent transformer model in carrying out temporal modeling with greater efficiency.

Moreover, utilizing a pre-existing model conserves time and computational resources, since the model has already undergone comprehensive training on vast datasets. It can be adapted to the particular task at hand through fine-tuning. Finally, using the sliding window technique in the input layer diminishes the occurrence of nested loops and streamlines the temporal intricacy of the model. In general, the rationales, as mentioned earlier, render the utilization of a pre-trained convolutional neural network a viable methodology for predicting PM2.5 data.

Long-term contextual features using a transformer

Long-term contextual characteristics are just as crucial for the effective forecasting of air quality data as the short-term contextual variables provided by the pre-trained CNN. The suggested solution uses a transformer network to capture these long-term relationships. The transformer network may model the input data sequence's long-term relationships well, which can also capture the context necessary for precise prediction. The primary task of the transformer encoder layers involves using self-attention and feed-forward neural network operations to carry out the prognosis. Ultimately, the decoder's linear layer produces conclusive prognostications. A full graphical representation of the architecture model can be seen in Fig. 10.

The transformer network computes a sequence of outputs that may be utilized for predicting based on a series of inputs. It comprises several encoder layers, each with a feed-forward neural network and a self-attention mechanism. The feed-forward neural network gives the model non-linearity, and the self-attention tool enables the transformer to recognize the significance of each input piece.

The Mean Squared Error (MSE) loss function and Adam optimizer used in CNN training are also used in the transformer. During training, a validation set is used to keep track of the model's progress while a batch size of 64 is set. A cyclical learning rate schedule is utilized to improve convergence, and the learning rate is adjusted during training using the learning rate annealing approach.

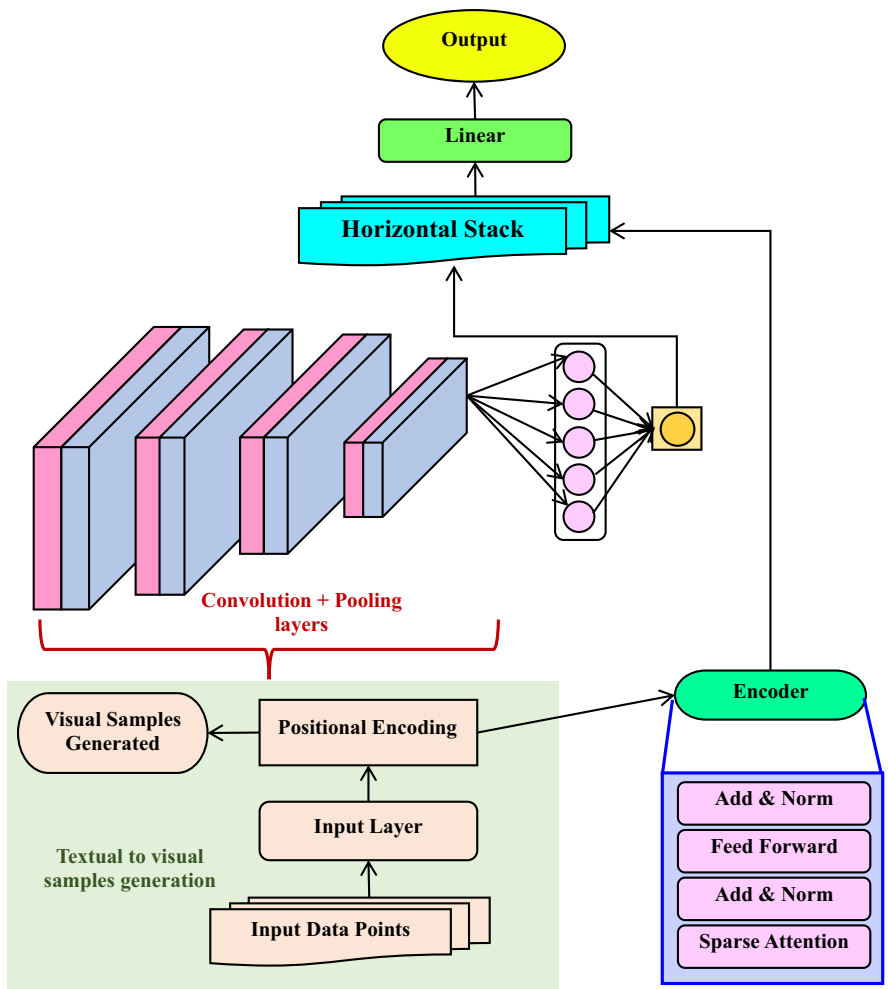


Fig. 10 The architecture of CATALYST—a novel Convolutional and Transformer model for Air Quality Forecasting

In general, integrating a pre-trained Convolutional Neural Network (CNN) and a Transformer network facilitates the extraction of contextual features from the input time series, encompassing both short-term and long-term information. This enables precise prediction of atmospheric quality information with reduced computational resources.

Final prediction

The ultimate forecast is generated through the amalgamation of the pre-existing Convolutional Neural Network and the Transformer network's outputs. The Convolutional Neural Network that has been pre-trained can extract contextual features that are short-term in nature. On the other hand, the Transformer is used to capture the long-term relationships in the input data. The final prediction is derived by integrating the outputs of both models.

To enhance the model's performance, several techniques, including dropout and batch normalization, are employed to mitigate the issue of overfitting and augment the model's capacity to generalize the novel data. The evaluation of the model's performance is conducted through the utilization of diverse evaluation metrics, including but not limited to Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and *R*-squared (R^2).

In brief, the ultimate forecast is derived by amalgamating the results obtained from the pre-existing Convolutional Neural Network (CNN) and the Transformer Network, followed by implementation methodologies such as dropout and batch normalization to enhance the efficacy of the model. The evaluation of the model's performance is conducted through diverse evaluation metrics. The full methodology can be seen in Fig. 3.

Computational efficiency

Various techniques are investigated and assessed with the aim of improving computational efficacy and mitigating the risk of overfitting. The selected model utilizes a solitary encoder and a linear layer, resulting in a notable decrease in complexity while still achieving proficient performance.

The utilization of a solitary encoder in the model results in optimization of the processing pipeline, thereby, reducing the computational burden. The utilization of this methodology obviates the necessity for numerous encoder layers, thereby, diminishing the computational workload during both the training and inference stages. Furthermore, the model integrates a linear layer, which enhances the computational efficiency. Linear layers exhibit computational efficiency in comparison to more intricate layers, such as fully connected layers, due to their utilization of simpler matrix operations. Through the implementation of these measures, the model can enhance the efficiency of computational resources while maintaining the precision of its predictions. The acceleration of training and inference times facilitates the model's practicality for real-time or large-scale applications. In general, the adoption of a solitary encoder and a linear layer amplifies computational efficacy, thereby, expediting the processing and mitigating the likelihood of overfitting. The enhancements facilitate a heightened level of efficacy and pragmatic in the forecasting framework.

Results and discussion

This section outlines the technical aspects of the study's implementation, encompassing the hardware configuration, employment of machine learning frameworks, and the particular model architecture utilized. The present discourse delves into the process of optimizing the model, the parameters involved in training, and the metrics employed for evaluation. Furthermore, the section presents a comprehensive examination of the dataset, demonstrating patterns in PM2.5 levels across various temporal dimensions such as time, seasons, and hours. The discourse underscores the association between the concentration of PM2.5 and diverse factors, including but not limited to the day of the week, season, and atmospheric conditions. Additionally, the aforementioned section conducts a comparative analysis between the CATALYST model and other baseline models, showcasing its heightened predictive accuracy and dependability.

Implementation details

The experiments are conducted on a PC server equipped with an NVIDIA QUADRO RTX A5000 graphics card featuring a memory capacity of 24 GB and 3042 NVIDIA Cuda cores. Implementing machine learning techniques for air quality forecasting involves

the utilization of open-source frameworks such as Pytorch (https://pytorch.org) and Sklearn (https://scikit-learn.org/). The open-source Tensorflow library, available at https://github.com/tensorflow/, configures deep learning and Transformer Models. The implementation that has been presented is founded on a time-series forecasting model that utilizes transformers. This model combines a Convolutional Neural Network (CNN) and a Transformer Network.

The model optimization process involves utilizing the Mean Squared Error (MSE) loss function alongside the Adam optimizer for updating the model parameters. The model's training is conducted with a batch size of 64. The training process is terminated after 1000 epochs or when the validation loss fails to demonstrate improvement for ten consecutive periods. The code employs an automatic technique, learning rate annealing, to modify the learning rate in response to the training progress. Additionally, cyclical learning-rate schedules vary the learning rate to enhance convergence cyclically. The starting learning rate varies from 0.000001 to 0.000200 according to requirement. The proposed framework's training, validation, and testing ratio (Fig. 11) has been set to 80%, 10%, and 10%, respectively. The optimal model is selected employing the validation loss metric and subsequently employed to generate predictions on the test dataset.

To assess the efficacy of the model, three evaluation metrics are employed, namely the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Fundamental Percentage Error (MAPE). Metrics are utilized to quantify the disparity between the anticipated values and the actual values. The assessment is conducted on the designated test dataset, and the outcomes are documented.

The implementation presented showcases a time-series forecasting model that utilizes a transformer-based approach, integrating both a Convolutional Neural Network and a Transformer Network. The model's performance on the test set is satisfactory, indicating its potential applicability to diverse time-series forecasting tasks. The details about the implementation can serve as a valuable reference for scholars and professionals who intend to utilize this model for their individual time-series prediction endeavors.

Evaluation metrics

The Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 are statistical techniques utilized to determine the discrepancy between predicted and observed values. This study compares the error metrics of RMSE, MAE, and R^2 for different models.

Root Mean Square Error (RMSE) is a statistical metric that quantifies the discrepancy between the predicted and observed values. Its primary application involves the computation of values over distinct time intervals. The degree of misalignment in the prediction increases proportionally with the magnitude of the value as shown in Eq. 3.

The Mean Absolute Error (MAE) is a widely employed assessment metric in scholarly literature, particularly within the domain of data analysis and prediction. This is a metric that quantifies the mean absolute deviation between the projected values and the factual values. In order to determine the Mean Absolute Error (MAE), the initial step involves computing the absolute discrepancies between the anticipated values and their corresponding factual values. The dissimilarities are subsequently computed to derive the comprehensive mean absolute error (MAE) metric as shown in Eq. 4. The Mean Absolute Error (MAE) is denominated in the identical units as the predicted variable, thereby rendering it readily interpretable and facilitating comparisons.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=0}^{l-1} \left(z_i - f_i\right)^2}{l}} \tag{3}$$

$$\mathrm{MAE} = \frac{\sum_{i=0}^{l-1} \left|z_i - f_i\right|}{l} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{x} \left(z_i - f_i\right)^2}{\sum_{i=1}^{x} \left(z_i - f_i^{\mathrm{mean}}\right)^2} \tag{5}$$

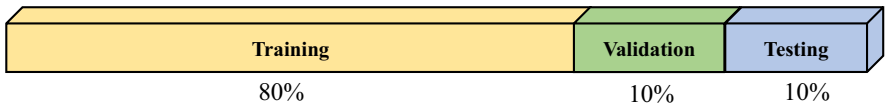


Fig. 11 Training, validation, and testing ratio

Ablation experiments

In this section, we further perform the ablation experiments. The experiments were conducted by modifying the CATALYST by reducing different modules to measure the performance gain of different components in the CATALYST. For this purpose, the constructed three comparative versions of the CATALYST are as follows:

(1) CATALYST-No-CNN: The ablation experiment entails the extraction of the Convolutional Neural Network (CNN) component from the CATALYST model. The CATALYST-No-CNN model places exclusive emphasis on the transformer's functionalities while disregarding the feature extraction capabilities of the CNN. However, despite these challenges, the system is able to attain a satisfactory level of performance, as evidenced by its root mean square error (RMSE) of 35.83, mean absolute error (MAE) of 22.77, and an *R*-squared value of 0.81. This finding illustrates the substantial contribution of the transformer module in the CATALYST model toward capturing intricate temporal connections.
(2) Lightweight CATALYST: The primary objective of the Lightweight CATALYST version is to minimize complexity and computing demands. The aforementioned objective can be accomplished by potentially reducing the number of layers or hidden units in the CATALYST model while preserving the presence of both the CNN and transformer components. The aforementioned lightweight iteration demonstrates encouraging outcomes, as seen by an RMSE (Root Mean Square Error) of 33.79, MAE (Mean Absolute Error) of 22.21, and an *R*-squared value of 0.81. These metrics indicate that a simpler design may effectively retain strong predictive powers.
(3) CATALYST no Positional Encoder: In the CATALYST No Positional Encoder experiment, the model does not include the positional encoder component. Positional encoders play a crucial role in comprehending the sequential arrangement inside the transformer architecture. Despite the lack of positional encoding, the model has commendable performance, with an RMSE (Root Mean Square Error) of 35.27, MAE (Mean Absolute Error) of 21.44, and an *R*-squared value of 0.82. This observation indicates that the transformer model has the ability to address the absence of positioning information in this particular activity by using its intrinsic self-attention capabilities.

The results obtained from the ablation experiments demonstrate that the CATALYST model has a high degree of robustness and adaptability across different configurations as can be observed in Table 1. The exclusion of the CNN component does not substantially impede the model's performance, suggesting that the transformer is proficient at capturing temporal patterns. The implementation of a less burdensome iteration of the CATALYST framework effectively preserves its operational efficiency while concurrently mitigating its intricacy. Moreover, the model's capacity to achieve high performance even in the absence of positional encoding implies that the self-attention processes employed by the transformer are adept at effectively managing the order of sequences. The results demonstrate the adaptability and robust prediction capabilities of the CATALYST model, even when subjected to various architectural alterations.

Comparison of performance with baseline models

The study's performance metrics for a range of models, namely SVR-RBF, SVR-Linear, ARIMA, Boost, LSTM, Transformer, and the proposed CATALYST, are presented in Table 1. The assessment criteria encompass Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and *R*-squared (R^2) score as you can see models Table 2.

The CATALYST model exhibited noteworthy outcomes, as evidenced by its RMSE value of 21.01, MAE value of 13.37, and R^2 score of 0.83. The aforementioned metrics demonstrate that CATALYST exhibits higher levels of predictive precision and dependability when compared to alternative models.

Table 1 Comparison of prediction performances in Ablation experiments

Models	RMSE	MAE	R^2
CATALYST-No-CNN	35.83	22.77	0.81
Light weight CATALYST	33.79	22.21	0.81
CATALYST-No-Positional-Encoder	35.27	21.44	0.82

Table 2 Comparison of different models

MODELS	RMSE	MAE	R^2
SVR-RBF	36.46	28.78	0.65
SVR-linear	42.37	34.24	0.53
Arima	55.51	48.28	0.82
Boost	25.72	12.53	0.84
LSTM	25.71	14.61	0.81
Transformer	24.02	13.92	0.84
CATALYST (Proposed)	**21.01**	**13.37**	**0.89**

Bold entries in table to hold significance that the bold one is our proposed model

It is noteworthy that CATALYST exhibits superior performance compared to Boost, LSTM, and Transformer models, which are widely recognized for their efficacy in time-series prediction tasks.

Comparison with available datasets

To enhance the credibility of CATALYST, an evaluation of its efficacy was conducted on two distinct publicly accessible datasets, namely Winchester. The datasets offer a range of geographical and environmental settings, thereby enabling an assessment of the models.

Relationsgeship analysis between datasets and performance metrics

To evaluate the resilience of the constructed model, a supplementary dataset sourced from Westminster, London is employed. A comparative analysis is conducted on the model's performance, with a specific focus on the time intervals of t-3, t-6, t-12, t-24, and t-48, between the Delhi dataset, the London Winchester dataset, and the Mumbai dataset. The objective of the evaluation is to investigate the potential correlations and ascertain the model's consistent performance across diverse datasets.

- London Dataset: The information pertaining to London is a useful resource that has been gathered from two separate stations located throughout the city. The dataset consists of hourly observations of PM2.5 particulate matter concentrations, spanning from December 30, 2018, to January 1, 2020. Prior to utilizing this dataset for analysis, a meticulous data pretreatment methodology was implemented. The process of data cleaning is a crucial and foundational step in order to guarantee the quality and accuracy of the data. All instances of outliers, missing numbers, or abnormalities have been recognized and appropriately resolved. This methodology improves the dependability of the dataset and guarantees its appropriateness for subsequent research. The calculation of mean values for PM2.5 concentrations serves to provide a measure of central tendency within the dataset, facilitating the analysis of air quality changes in London. The utilization of statistics facilitates the concise representation and enhanced accessibility of data, hence rendering it more suitable for scholarly investigation. It is essential to compare the performance of the model on the London dataset with its initial development for Delhi due to several reasons. London and Delhi have significant disparities in their urban landscapes, encompassing divergences in the origins of pollution, traffic flow, industrial operations, and climatic circumstances. Therefore, the evaluation of the model's suitability for various metropolitan environments is contingent upon conducting tests in London. Moreover, this comparative analysis functions as a significant means of verifying the model's resilience and applicability. The model's effective performance in both Delhi and London serves as evidence of its broad applicability in tackling air quality concerns on a global scale, irrespective of distinct geographical contexts. The comprehension of these variances has the potential to enhance the precision and applicability of the model by incorporating city-specific attributes, hence assuring its pertinence and reliability across diverse geographical regions.
- Mumbai Dataset: The dataset utilized in this research pertains to Mumbai and was obtained from three separate monitoring stations situated inside the urban area of Mumbai, India. Air quality data was collected at each site during the duration of 1 year, with observations recorded at intervals of 15 min. The focal variable in this dataset is "PM2.5", denoting the concentration of tiny particulate matter measuring 2.5 μm or smaller in the atmosphere. PM2.5 serves as a crucial measure of air quality and possesses substantial consequences for both public health and environmental integrity. Before doing the study, the dataset underwent a thorough data-cleaning procedure to handle missing values, outliers, and inconsistencies.

The inclusion of this preprocessing phase was of utmost importance in order to guarantee the integrity and dependability of the data. Furthermore, the timestamps that have been collected, spanning from January 1, 2021, to December 24, 2021, offer crucial temporal data that is important for doing time-series analysis. It is noteworthy to mention that the dataset under consideration pertains to the monitoring of air quality in Mumbai, a metropolitan region in India characterized by high population density. Although the reviewer expressed reservations over the dataset's applicability to the specific context, it is important to underscore the considerable significance of Mumbai's air quality challenges. In brief, the Mumbai dataset provides significant insights into the air quality conditions in Mumbai, India, specifically emphasizing the concentrations of PM2.5 particles. The accurate depiction of the dataset and the implementation of data-cleaning procedures are crucial in guaranteeing the integrity and appropriateness of the dataset for analysis within the specified context.

The analysis of the model's performance on the Delhi, Mumbai, and London Winchester datasets provides significant insights into the complexities and disparities involved in forecasting air quality in diverse geographical areas. It is worth mentioning that the model demonstrated the lowest Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) values for the London Winchester dataset, suggesting its superior performance when compared to the Delhi and Mumbai datasets as it can also be observed with Fig. 12.

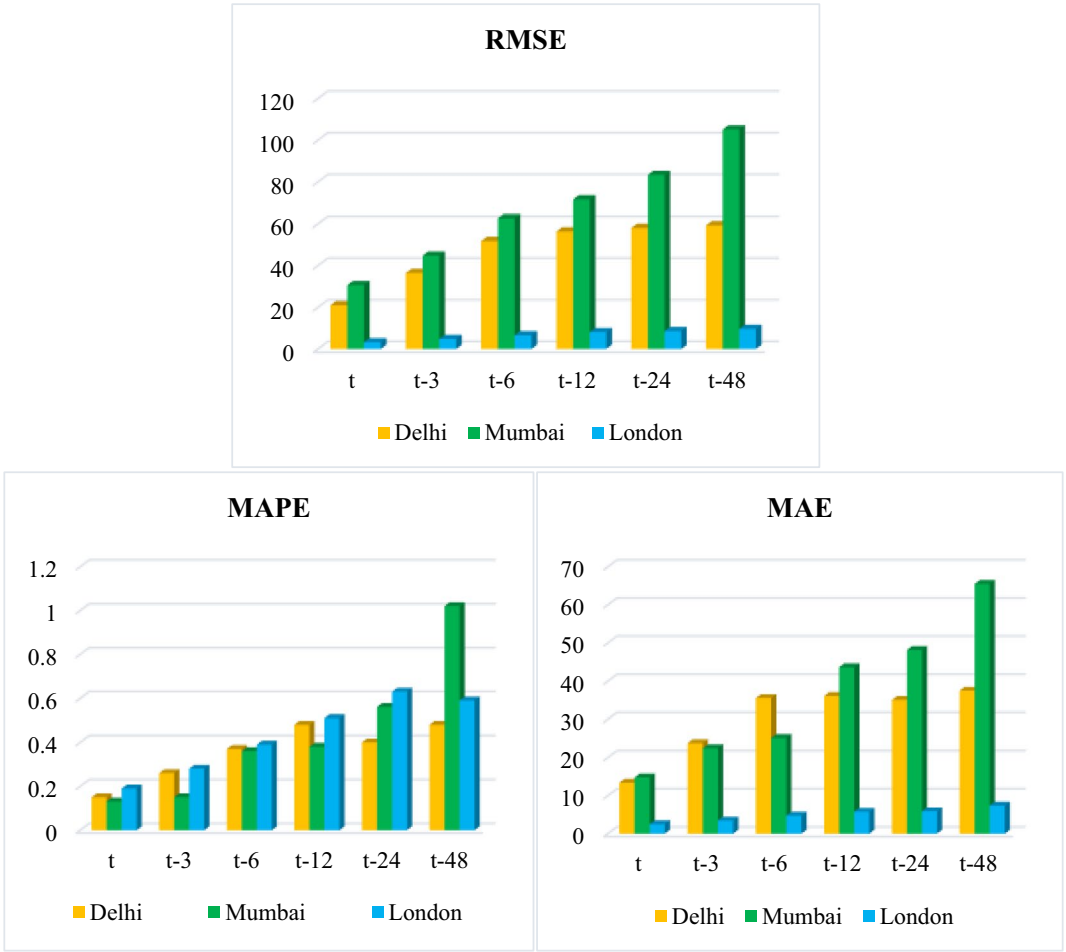


Fig. 12 Performance comparison across mega cities

The disparity in performance can be attributed to the unique air quality attributes exhibited by each respective site. Delhi, renowned for its significant air pollution, encounters notable variations in PM2.5 levels, encompassing both exceedingly elevated and diminished amounts. On the other hand, London exhibits reduced levels of pollution with a more limited range of fluctuation. The notable disparity in pollution patterns is a significant obstacle for the model, since properly forecasting extreme levels intrinsically proves to be more challenging, leading to elevated values of RMSE, MAE, and MAPE for Delhi as it can be observed in Table 3.

Moreover, the performance of the model is affected by the constrained quantity of data accessible for these datasets. The model's predictive accuracy may be compromised due to a lack of comprehensive historical data, which might impede its capacity to identify enduring patterns and seasonal variations. As a result, the model's predictions may be less exact. It is crucial to highlight that these findings underscore the need of employing a geographically tailored methodology for air quality modeling and the significance of modifying models to tackle the distinct air quality complexities encountered by certain cities or regions. It is important for policymakers and academics to consider these variances when formulating measures aimed at mitigating air pollution and safeguarding public health.

The following analysis of obtained results based on time variations as seen in the Delhi dataset, London Winchester dataset, and Mumbai dataset shown in Table 3 is:

- The analysis of the results based on different time variations provides valuable insights into the model's performance.
- Current time point (t): At present, the model attains its optimal precision, exhibiting the least values of RMSE, MAE, and MAPE. This suggests that the model is proficient in generating precise forecasts utilizing the latest available data.
- Shorter time intervals (t-3, t-6): The model's performance exhibits a relatively high degree of accuracy for shorter time intervals, albeit with slightly elevated error rates when compared to the present time point. This implies that the model remains proficient in effectively capturing short-term patterns and dependencies. At present, the model attains its optimal precision, exhibiting the least values of RMSE, MAE, and MAPE. This suggests that the model exhibits efficacy in generating precise prognostications grounded on the latest available data.
- Medium time intervals (t-12, t-24): As the temporal delay extends to moderate intervals, the efficacy of the model begins to deteriorate, resulting in elevated error metrics. This suggests that the model could encounter difficulties in precisely capturing extended-term patterns and dependencies.
- Longer time interval (t-48): At the maximum time interval, the model's performance exhibits the greatest discrepancies across all metrics. This underscores the constraints of the model in accurately capturing prolonged temporal dynamics.

The results underscore the importance of considering different time intervals for robust predictions and understanding the temporal capabilities of the model. By analyzing the performance at various time variations, we gain valuable insights into the model's strengths and limitations in capturing temporal dynamics.

Table 3 Comparison of performance metrics between Delhi dataset, London Winchester dataset, and Mumbai dataset at different time variations

Time variation	Delhi dataset			Mumbai dataset			London Winchester dataset		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
T	21.01	13.37	0.15	30.67	14.74	0.13	3.19	2.46	0.19
T-3	36.45	23.74	0.26	44.74	22.42	0.15	4.81	3.44	0.28
T-6	51.78	35.53	0.37	62.61	25.13	0.36	6.55	4.72	0.39
T-12	56.33	36.09	0.48	71.61	43.59	0.38	8.09	5.75	0.51
T-24	58.025	35.01	0.4	83.33	48.1	0.56	8.56	5.86	0.63
T-48	59.33	37.44	0.48	105.14	65.49	1.02	9.61	7.36	0.59

Discussion

In this particular section new insights have emerged as a result of the development and evaluation of the CATALYST model, which is utilized for the purpose of forecasting air quality. The CATALYST model exhibits exceptional performance in comparison to many state-of-the-art models on the Delhi dataset. In addition, it demonstrates positive performance on the Mumbai and London Winchester datasets, despite the distinctive air quality constraints presented by these urban areas. In the context of model comparisons, the evaluation of the CATALYST model involves its comparison with many other air quality forecasting models, namely SVR-RBF, SVR-Linear, ARIMA, Boost, LSTM, and Transformer. The CATALYST model demonstrates superior performance across all parameters, including RMSE, MAE, and R^2, when evaluated the Delhi dataset. Also, we examine the performance of the CATALYST model across several time intervals, specifically t-3, t-6, t-12, t-24, and t-48. The model demonstrates optimal performance at the present time point (t), with a decrease in predicted accuracy as the time gap expands. This suggests that the model has superior performance in capturing short-term patterns and relationships in comparison to long-term ones. The importance of these findings is noteworthy as they emphasize the potential of the CATALYST model as a proficient tool for forecasting air quality in various geographical locations. The aforementioned approach outperforms several contemporary models and exhibits encouraging results on datasets acquired from diverse locales, each presenting distinct air quality obstacles.

The research paper by Xu et al. (2023) involved the utilization of a Graph Neural Network (GNN) model for the purpose of predicting PM2.5 concentrations in Beijing, China. The GNN model attained a Root Mean Square Error (RMSE) of 19.8 μg/m^3. When the CATALYST model is implemented on the Delhi dataset, it demonstrates superior performance compared to the GNN model, resulting in an RMSE (Root Mean Square Error) of 21.01 μg/m3.In a separate research conducted by (Guo et al., 2023), a unique approach was utilized to forecast the levels of PM2.5 in London, UK. The methodology employed in this study incorporated the integration of a Long Short-Term Memory (LSTM) network with a wavelet transform. The hybrid model demonstrated a Root Mean Square Error (RMSE) value of 3.0 μg/m^3. The CATALYST model has comparable performance on the London Winchester dataset, with an RMSE value of 3.19 μg/m^3.

In general, the CATALYST model demonstrates significant potential as a novel air quality forecasting model, with superior or comparable performance to many contemporary state-of-the-art models across diverse datasets. The model demonstrates the capacity to function as a valuable instrument for policymakers and scholars in their endeavors to address air pollution and protect public health.

Conclusion

The present study introduced a comprehensive methodology for predicting air quality through the utilization of a pre-trained Convolutional Neural Network (CNN) in combination with a transformer-based approach. The research methodology comprised three primary phases, namely, data acquisition and preprocessing, learning of short-term contextual features through CNN, and extraction of long-term contextual features using a Transformer.

A dataset of hourly air quality measurements was obtained by utilizing data from five monitoring stations in Delhi, which was subsequently subjected to cleaning and organization. The utilization of the sliding window methodology facilitated the division of the data into input and label blocks, resulting in a more efficient data processing approach and conservation of computational resources. Furthermore, the technique of positional encoding was employed to incorporate sinusoidal positional embeddings into the input data.

During the initial phase of contextual feature acquisition, the input data was presented as an image and processed through a pre-existing Convolutional Neural Network (CNN) to extract pertinent features. This methodology utilized the convolutional neural network's aptitude for detecting spatial patterns and interdependencies present in the dataset. The reduction of dimensionality reduction by converting the data on air quality into a vector image. This enabled the subsequent Transformer Model to efficiently carry out temporal modeling and capture enduring contextual associations; a Transformer Network was employed. The Transformer encoder layers were

successful in modeling the input data sequence and capturing the requisite context for precise predictions through the integration of self-attention and feedforward neural network operations. The ultimate predictions were generated by the linear layer of the transformer decoder.

The ultimate prediction was attained through the integration of the pre-trained Convolutional Neural Network and the Transformer outputs. The model's performance was enhanced, and overfitting was mitigated through techniques such as dropout and batch normalization. The evaluation of the model's performance was conducted through the utilization of various metrics, including but not limited to Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and *R*-squared (R^2).

The methodology proposed in this study presented a comprehensive framework for predicting air quality. It involved a series of steps including data collection, pre-processing, feature extraction, and prediction analysis. The findings indicated the efficacy of the methodology in precisely predicting atmospheric phenomena. By incorporating both short-term and long-term contextual features, a comprehensive comprehension of the dynamics of air quality can be achieved.

Future scope

Although our proposed methodology has shown promising results, there are several opportunities for further research and improvement.

The model's predictive ability may be enhanced by initially expanding the dataset through the inclusion of data from a larger number of monitoring stations and the integration of additional attributes, such as meteorological data and traffic patterns. The integration of diverse data sources would result in a more comprehensive understanding of the factors that influence atmospheric conditions. Moreover, exploring various architectures and alterations of the Transformer Model could potentially yield benefits. A study aimed at investigating the impact of different attention mechanisms or the adoption of Transformer variants, such as the Vision Transformer (ViT), could potentially lead to improvements in the capacity to capture extended dependencies and spatial associations.

In addition, it would be beneficial to evaluate the efficacy of our proposed methodology through comparative analyses with other contemporary models and ensemble techniques. The evaluation of the effectiveness of the model in various geographical locations or incorporating temporal variations in air quality patterns could provide further insights into its general suitability. An analysis of the model's interpretability and understanding of the importance of input features would improve the performance of the factors that affect air quality. The data mentioned above can assist policymakers in formulating reasonable judgements geared toward augmenting the standard of air purity.

The practical advancement of integrating the functional forecasting instrument developed into pre-existing air quality monitoring frameworks would be a significant achievement. The capability to provide instantaneous forecasts and notifications has the capacity to offer noteworthy perspectives for the management of public health and the implementation of policy interventions.

In brief, the research presents a dependable methodology for predicting atmospheric conditions by incorporating contextual features of both short-term and long-term duration. The proposed approach demonstrates promising results and offers potential for further exploration to enhance accuracy, comprehensiveness, and applicability in real-life situations. Through the exploration of these research avenues, a significant contribution can be made toward the improvement of air quality management and the promotion of a healthier environment for the general public.

Author contribution All authors contributed significantly to the development and completion of this research study. The specific contributions of each author are as follows:

- **Abhishek Verma**: software, validation, investigation, data curation, writing— original draft, and visualization.
- **Virender Ranga**: conceptualization, methodology, formal analysis, writing—review, and supervision.
- **Dinesh Kumar Vishwakarma**: writing—review and editing, and supervision.

All authors have reviewed and approved the final version of the manuscript submitted for publication. Correspondence regarding this research study and manuscript should be directed to Abhishek Verma (Abhishekcms08@gmail.com).

Data availability The data to support the conclusions of this study is available upon request from the corresponding author.

Declarations

Ethics approval and consent to participate This research study, titled "A Novel Approach for Forecasting PM2.5 Pollution in Delhi Using CATALYST", conducted by Abhishek Verma, Virender Ranga, and Dinesh Kumar Vishwakarma, was undertaken with due consideration of ethical principles and in compliance with the guidelines set forth by relevant ethical bodies. Consent to participate is not applicable.

Consent for publication We, the undersigned authors of the manuscript titled "A Novel Approach for Forecasting PM2.5 Pollution in Delhi Using CATALYST", submitted for consideration to "Environmental Monitoring and Assessment", hereby confirm our consent for the publication of this manuscript in the journal.

We acknowledge that by providing this consent, we are granting the publisher**,** "Springer Nature", the right to publish, reproduce, distribute, and communicate our work to the scholarly community and the public. We understand that this includes both printed and electronic formats. We affirm that:
1. All authors have made substantial contributions to the conception, design, data acquisition, analysis, interpretation, and drafting of the manuscript.
2. All authors have reviewed and approved the final version of the manuscript submitted for publication.
3. The manuscript is an original work, and its submission to "Environmental Monitoring and Assessment" does not infringe upon the rights of any third party.
4. Any conflict of interest or financial disclosures related to the research have been accurately disclosed in the manuscript.
We understand that "Environmental Monitoring and Assessment" may perform peer review and editorial processes to ensure the quality and integrity of the published work. We also acknowledge that the manuscript may be subject to revisions as suggested by the peer reviewers and editors.
Furthermore, we acknowledge that the corresponding author, "Abhishek Verma", will serve as the main point of contact between the authors and "Environmental Monitoring and Assessment" during the review and publication process. We agree to promptly respond to any queries or requests from the journal's editorial team.

Competing interests The authors declare no competing interests.

References

Agarwal, S., Sharma, S., Suresh, R., Rahman, M. H., Vranckx, S., Maiheu, B., Blyth, L., Janssen, S., Gargava, P., Shukla, V. K., & Batra, S. (2020). Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions. *Science of the Total Environment, 735*, 139454. https://doi.org/10.1016/j.scitotenv.2020.13945

Chang, Q., Zhang, H., & Zhao, Y. (2020). Ambient air pollution and daily hospital admissions for respiratory system–related diseases in a heavy polluted city in Northeast China. *Environmental Science and Pollution Research, 27*(9), 10055–10064. https://doi.org/10.1007/s11356-020-07678-8

Chen, X., Yin, L., Fan, Y., Song, L., Ji, T., Liu, Y., Tian, J., & Zheng, W. (2020). Temporal evolution characteristics of PM2.5 concentration based on continuous wavelet transform. *Science of the Total Environment, 699*, 134244. https://doi.org/10.1016/j.scitotenv.2019.134244

Chu, J., Dong, Y., Han, X., Xie, J., Xu, X., & Xie, G. (2021). Short-term prediction of urban PM2.5 based on a hybrid modified variational mode decomposition and support vector regression model. *Environmental Science and Pollution Research, 28*(1), 56–72. https://doi.org/10.1007/s11356-020-11065-8

Cusworth, D. H., Mickley, L. J., Sulprizio, M. P., Liu, T., Marlier, M. E., DeFries, R. S., Guttikunda, S. K., & Gupta, P. (2018). Quantifying the influence of agricultural fires in northwest India on urban air pollution in Delhi, India. *Environmental Research Letters, 13*(4), 044018. https://doi.org/10.1088/1748-9326/aab303

Dai, H., Huang, G., Zeng, H., & Yu, R. (2022). Haze risk assessment based on improved PCA-MEE and ISPO-LightGBM model. *Systems, 10*(6), 6. https://doi.org/10.3390/systems10060263

Dai, H., Huang, G., Wang, J., & Zeng, H. (2023a). VAR-tree model based spatio-temporal characterization and prediction of O3 concentration in China. *Ecotoxicology and Environmental Safety, 257*, 114960. https://doi.org/10.1016/j.ecoenv.2023.114960

Dai, H., Huang, G., & Zeng, H. (2023b). Multi-objective optimal dispatch strategy for power systems with spatio-temporal distribution of air pollutants. *Sustainable Cities and Society, 98*, 104801. https://doi.org/10.1016/j.scs.2023.104801

Dai, L., Zanobetti, A., Koutrakis, P., & Schwartz, J. D. (2014). Associations of fine particulate matter species with mortality in the United States: A multicity time-series analysis. *Environmental Health Perspectives, 122*(8), 837–842. https://doi.org/10.1289/ehp.1307568

Darçın, M. (2014). Association between air quality and quality of life. *Environmental Science and Pollution Research, 21*(3), 1954–1959. https://doi.org/10.1007/s11356-013-2101-3

Dey, S., Di Girolamo, L., van Donkelaar, A., Tripathi, S. N., Gupta, T., & Mohan, M. (2012). Variability of outdoor fine particulate (PM2.5) concentration in the Indian subcontinent: A remote sensing approach. *Remote Sensing of Environment, 127*, 153–161. https://doi.org/10.1016/j.rse.2012.08.021

Efstathiou, M. N., & Varotsos, C. A. (2012). Intrinsic properties of Sahel precipitation anomalies and rainfall. *Theoretical and Applied Climatology, 109*(3), 627–633. https://doi.org/10.1007/s00704-012-0605-2

Fan, J. L., Li, Q., Hou, J., Feng, X., Karimian, H., & Lin, S. (2017). A spatiotemporal prediction framework for air pollution based on deep RNN. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 4*, 15–22. https://doi.org/10.5194/isprs-annals-iv-4-w2-15-2017

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., & Wang, J. (2015). Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment, 107*, 118–128. https://doi.org/10.1016/j.atmosenv.2015.02.030

Fowler, D., Brimblecombe, P., Burrows, J., Heal, M. R., Grennfelt, P., Stevenson, D. S., Jowett, A., Nemitz, E., Coyle, M., Liu, X., Chang, Y., Fuller, G. W., Sutton, M. A., Klimont, Z., Unsworth, M. H., & Vieno, M. (2020). A chronology of global air quality. *Philosophical Transactions of the Royal Society a: Mathematical, physical and engineering sciences, 378*(2183), 20190314. https://doi.org/10.1098/rsta.2019.0314

Gariazzo, C., Carlino, G., Silibello, C., Renzi, M., Finardi, S., Pepe, N., Radice, P., Forastiere, F., Michelozzi, P., Viegi, G., & Stafoggia, M. (2020). A multi-city air pollution population exposure study: Combined use of chemical-transport and random-forest models with dynamic population data. *Science of the Total Environment, 724*, 138102. https://doi.org/10.1016/j.scitotenv.2020.138102

Graupe, D., Krause, D., & Moore, J. (1975). Identification of autoregressive moving-average parameters of time series. *IEEE Transactions on Automatic Control, 20*(1), 104–107. https://doi.org/10.1109/TAC.1975.1100855

Guo, Q., He, Z., & Wang, Z. (2023). Predicting of Daily PM2.5 Concentration employing wavelet artificial neural networks based on meteorological elements in Shanghai, China. *Toxics, 11*(1), 1. https://doi.org/10.3390/toxics11010051

Guttikunda, S. K., & Jawahar, P. (2014). Atmospheric emissions and pollution from the coal-fired thermal power plants in India. *Atmospheric Environment, 92*, 449–460. https://doi.org/10.1016/j.atmosenv.2014.04.057

Ha Chi, N. N., & Kim Oanh, N. T. (2021). Photochemical smog modeling of PM2.5 for assessment of associated health impacts in crowded urban area of Southeast Asia. *Environmental Technology & Innovation, 21*, 101241. https://doi.org/10.1016/j.eti.2020.101241

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoff, R. M., & Christopher, S. A. (2009). Remote sensing of particulate pollution from space: Have we reached the promised land? *Journal of the Air & Waste Management Association, 59*(6), 645–675. https://doi.org/10.3155/1047-3289.59.6.645

Jian, L., Zhao, Y., Zhu, Y.-P., Zhang, M.-B., & Bertolatti, D. (2012). An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. *Science of the Total Environment, 426*, 336–345. https://doi.org/10.1016/j.scitotenv.2012.03.025

Kandlikar, M., & Ramachandran, G. (2000). The causes and consequences of particulate air pollution in urban India: A synthesis of the science. *Annual Review of Energy and the Environment, 25*(1), 629–684. https://doi.org/10.1146/annurev.energy.25.1.629

Khan, M. A., & Ghouri, A. M. (2011). *Environmental pollution: Its effects on life and its remedies* (SSRN Scholarly Paper No. 1981242). https://papers.ssrn.com/abstract=1981242

Lee, H. J., Liu, Y., Coull, B. A., Schwartz, J., & Koutrakis, P. (2011). A novel calibration approach of MODIS AOD data to predict $PM_{2.5}$ concentrations. *Atmospheric Chemistry and Physics, 11*(15), 7991–8002. https://doi.org/10.5194/acp-11-7991-2011

Li, H., Li, Y., Zhang, Q., Zhang, Y., & Sun, X. (2019). PM2.5 prediction using a convolutional neural network and long short-term memory network based on air quality data. *International Journal of Environmental Research and Public Health, 16*(7), 1243. https://doi.org/10.3390/ijerph16071243

Li, L. (2020). A robust deep learning approach for spatiotemporal estimation of satellite AOD and PM2.5. *Remote Sensing, 12*(2), 264-NA. https://doi.org/10.3390/rs12020264

Liao, Q. Z., Zhu, M., Wu, L., Pan, X., Tang, X., & Wang, Z. (2020). Deep learning for air quality forecasts: A review. *Current Pollution Reports, 6*(4), 399–409. https://doi.org/10.1007/s40726-020-00159-z

Liao, X., Tu, H., Maddock, J. E., Fan, S., Lan, G., Wu, Y., Yuan, Z. K., & Lu, Y. (2015). Residents' perception of air quality, pollution sources, and air pollution control in Nanchang, China. *Atmospheric Pollution Research, 6*(5), 835–841. https://doi.org/10.5094/APR.2015.092

Liou, N.-C., Luo, C.-H., Mahajan, S., & Chen, L.-J. (2020). Why is short-time PM2.5 forecast difficult? The effects of sudden events. *IEEE Access, 8*, 12662–12674. https://doi.org/10.1109/ACCESS.2019.2963341

Liu, J., Xu, C., Liu, Y., Zhang, Z., & Lu, J. (2020). Hybrid PM2.5 prediction model based on convolutional neural network and seasonal decomposition method. *Environmental Science and Pollution Research, 27*(8), 8699–8711. https://doi.org/10.1007/s11356-019-07594-7

Lv, B., Cobourn, W. G., & Bai, Y. (2016). Development of nonlinear empirical models to forecast daily PM2.5 and ozone levels in three large Chinese cities. *Atmospheric Environment, 147*, 209–223. https://doi.org/10.1016/j.atmosenv.2016.10.003

Mahajan, S., Liu, H.-M., Tsai, T.-C., & Chen, L.-J. (2018). Improving the accuracy and efficiency of PM2.5 forecast service using cluster-based hybrid neural network model. *IEEE Access, 6*, 19193–19204. https://doi.org/10.1109/ACCESS.2018.2820164

Murthy, B. S., Latha, R., Tiwari, A., Rathod, A., Singh, S., & Beig, G. (2020). Impact of mixing layer height on air quality in winter. *Journal of Atmospheric and Solar-Terrestrial Physics, 197*, 105157. https://doi.org/10.1016/j.jastp.2019.105157

Nagpure, A. S., Gurjar, B. R., & Martel, J. (2014). Human health risks in national capital territory of Delhi due to air pollution. *Atmospheric Pollution Research, 5*(3), 371–380. https://doi.org/10.5094/APR.2014.043

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing, 452*, 48–62. https://doi.org/10.1016/j.neucom.2021.03.091

Puthumana, J. S., Ngaage, L. M., Borrelli, M. R., Rada, E. M., Caffrey, J., & Rasko, Y. (2021). Risk factors for cooking-related burn injuries in children, who global burn registry. *Bulletin of the World Health Organization*, *99*(6). https://doi.org/10.2471/BLT.20.279786

Ragab, M. G., Abdulkadir, S. J., Aziz, N., Al-Tashi, Q., Alyousifi, Y., Alhussian, H., & Alqushaibi, A. (2020). A novel one-dimensional CNN with exponential adaptive gradients for air pollution index prediction. *Sustainability, 12*(23), 23. https://doi.org/10.3390/su122310090

Shi, Z., Ye, Y., Chen, X., & Wang, H. (2020). Attention-based recurrent neural network for PM2.5 concentration prediction. *Environmental Science and Pollution Research, 27*(22), 27839–27850. https://doi.org/10.1007/s11356-020-08912-2

Tian, M., Gao, J., Zhang, L., Zhang, H., Feng, C., & Jia, X. (2021). Effects of dust emissions from wind erosion of soil on ambient air quality. *Atmospheric Pollution Research, 12*(7), 101108. https://doi.org/10.1016/j.apr.2021.101108

Tiwari, S., Srivastava, A. K., Bisht, D. S., Parmita, P., Srivastava, M. K., & Attri, S. D. (2013). Diurnal and seasonal variations of black carbon and PM2.5 over New Delhi, India: Influence of meteorology. *Atmospheric Research, 125–126*, 50–62. https://doi.org/10.1016/j.atmosres.2013.01.011

Varotsos, C. A., & Cracknell, A. P. (2020). Remote sensing letters contribution to the success of the sustainable development goals—UN 2030 agenda. *Remote Sensing Letters, 11*(8), 715–719. https://doi.org/10.1080/2150704X.2020.1753338

Varotsos, C. A., Mazei, Y., Saldaev, D., Efstathiou, M., Voronova, T., & Xue, Y. (2021). Nowcasting air pollution episodes in megacities: A case study for AthensGreece. *Atmospheric Pollution Research, 12*(7), 101099. https://doi.org/10.1016/j.apr.2021.101099

Varotsos, C., & Cartalis, C. (1991). Re-evaluation of surface ozone over Athens, Greece, for the period 1901–1940. *Atmospheric Research, 26*(4), 303–310. https://doi.org/10.1016/0169-8095(91)90024-Q

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Verma, A., Ranga, V., & Vishwakarma, D. K. (2023). Forecasting of satellite based carbon-monoxide time-series data using a deep learning approach. In *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT)* (pp. 1–4). https://doi.org/10.1109/ICITIIT57246.2023.10068609

Wang, L., Zhang, H., Mao, L., Li, S., & Wu, H. (2020). Assessing spatiotemporal characteristics of urban PM2.5 using fractal dimensions and wavelet analysis. *Mathematical Problems in Engineering*, *2020*, e8091515. https://doi.org/10.1155/2020/8091515

Wang, P., Zhang, G., Chen, F., & He, Y. (2019). A hybrid-wavelet model applied for forecasting PM2.5 concentrations in Taiyuan City, China. *Atmospheric Pollution Research, 10*. https://doi.org/10.1016/j.apr.2019.08.002

Wei, P., Xie, S., Huang, L., Zhu, G., Tang, Y., & Zhang, Y. (2021). Prediction of PM2.5 concentration in Guangxi region, China based on MLR-ARIMA. *Journal of Physics: Conference Series, 2006*(1), 012023. https://doi.org/10.1088/1742-6596/2006/1/012023

Xu, J., Wang, S., Ying, N., Xiao, X., Zhang, J., Jin, Z., Cheng, Y., & Zhang, G. (2023). Dynamic graph neural network with adaptive edge attributes for air quality prediction: A case study in China. *Heliyon, 9*(7), e17746. https://doi.org/10.1016/j.heliyon.2023.e17746

Yao, L., He, J., Sun, J., Gao, F., & Li, W. (2021). A multi-scale and multi-channel convolutional neural network for PM2.5 prediction. *Atmospheric Environment, 244*, 117962. https://doi.org/10.1016/j.atmosenv.2020.117962

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 prediction based on random foresT, XGBoost, and deep learning using multi-source remote sensing data. *Atmosphere, 10*(7), 7. https://doi.org/10.3390/atmos10070373

Zhang, L. L., Li, D., & Guo, Q. (2020). Deep learning from spatio-temporal data using orthogonal regularization residual CNN for air prediction. *IEEE Access, 8*(NA), 66037–66047. https://doi.org/10.1109/access.2020.2985657

Zhao, R., Gu, X., Xue, B., Zhang, J., & Ren, W. (2018). Short period PM2.5 prediction based on multivariate linear regression model. *PLOS ONE, 13*(7), e0201011. https://doi.org/10.1371/journal.pone.0201011

Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., & Li, T. (2015, August 1). Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th SIGKDD Conference on Knowledge Discovery and Data Mining* https://www.microsoft.com/en-us/research/publication/forecasting-fine-grained-air-quality-based-on-big-data/

Zhu, L., Hao, Y., Lu, Z.-N., Wu, H., & Ran, Q. (2019). Do economic activities cause air pollution? Evidence from China's major cities. *Sustainable Cities and Society, 49*, 101593. https://doi.org/10.1016/j.scs.2019.101593

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.

A Novel Real-Time Parametric Tracking Approach for Robust Microwave Filter Tuning

2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings) | 979-8-3503-2234-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIBThings58340.2023.10292473

Even Sekhri*
Department of Electrical Power Engieering and Mechatronics
Tallinn University of Technology
Tallinn, Estonia
0000-0002-2578-1910

Mart Tamre
Department of Electrical Power Engieering and Mechatronics
Tallinn University of Technology
Tallinn, Estonia
0000-0002-7489-9683

Rajiv Kapoor
Department of Electronics and Communication Engieering
Delhi Technological University
New Delhi, India
0000-0003-3020-1455

Rahul Kumar
Department of Electronics and Communication Engieering
Delhi Technological University
New Delhi, India
0000-0002-5068-5488

***Abstract*—The robust tracking of tuning parameters is crucial in enhancing the efficiency of automated filter tuning. The existing automated filter tuning methods use the phase plot information at the center frequency of filter and extract tuning parameters represented by the marker position. This paper presents a vision-based approach for tracking the tuning parameters over the Smith chart display of Vector Network Analyzer (VNA) using Deep Learning (DL). Using a camera to detect the marker allows the ease in presenting the proof of concept without accessing or modifying the commercial VNA's software. The realm of microwave cavity filter tuning has yet to incorporate real-time parametric tracking of tuning parameters on a Smith chart. For training the DL network, a labeled dataset has been curated. Manual tuning of a 5th order cavity filter (a filter with 5 tuning screws) was used to test the proposed method. Results obtained in this research demonstrates that the proposed method outperforms the state-of-the-art techniques in terms of tracking performance. The mAP value was determined to be 98.2% using YOLOv5. The robustness of the suggested method has been demonstrated through rigorous testing of the proposed technique on the polar plot display of the VNA. This methodology holds the potential for complete automation of the filter tuning process.**

Keywords—filter tuning, marker detection, microwave filter tuning, parametric tracking, robust tracking, Faster-CNN, Single-Shot Detector (SSD), YOLOv5

I. INTRODUCTION

Microwave (MW) cavity bandpass filters are commonly employed to isolate the desired signal amidst a crowded frequency spectrum. These filters require tuning to compensate for design imperfections and manufacturing errors [1]. Tuning a filter using the phase response of the reflection coefficient (S_{ii}, where i = number of ports) is a common way of tuning all-pole filters [2]. To determine the phase change, the tuning parameter variations at the center frequency of the filter are observed. Using the Smith chart plot on the Vector Network Analyzer (VNA) equipment is another approach to monitor the phase change. Researchers striving to automate the filter tuning process often overlook Smith chart-based tuning due to two primary reasons: a) the complexity involved in dealing with the "infinity" term on the short-circuit side of the Smith chart; and, b) the simultaneous variation of the resistance value and reactance value (capacitive reactance or inductive reactance) while tuning the filter. However, by precisely tracking the marker movement on the Smith chart, there is a possibility for achieving autonomous filter tuning.

The research community has put forth numerous techniques for Computer-Aided Tuning (CAT) of MW filters, offering a wide array of options for the tuning process including filter diagnosis [2][3][4][5][6][7], Time Domain (TD) [8][9], Group Delay (GD) [10], Space Mapping (SM) [11][12][13][14], Fuzzy Logic (FL) [15][16][17][18], Artificial Neural Network (ANN) [19][20][21][22], etc. TD-based tuning solutions needed a human operator to actually tune the filter. The filter diagnosis-based solutions rely on circuit models, which are inherently inexact. The automated filter tuning was attempted on data-driven algorithms based on SM [12] and FL [18]. The accuracy of SM-based filter tuning approach is dependent on the complexity of the coarse model and the fine model. The commercialization of FL-based filter tuning was hampered by the necessity of a large number of 'IF-THEN' linguistic fuzzy rules. Despite the wide range of CAT-based filter tuning techniques and automated filter tuning solutions documented in the literature, none of the existing approaches have addressed the tracking of tuning parameters variation at the filter's center frequency. Also, the utilization of Smith chart for automated filter tuning has not been considered.

With the significant boost in computational power over the recent years, DL-based algorithms have gained prevalent traction in the fields of object detection and tracking [23]. DL-based approaches not only possess powerful learning capabilities but also excel at handling challenges such as occlusion, scale transformation, and varying background etc. The DL-based object detection approaches can be divided into two categories i.e., region-based methods; and regression model-based methods. The region-based methods include Region-based Convolutional Neural Network (R-CNN) [24], Spatial Pyramid Pooling network (SPP-net) [25], Fast R-CNN [26], Faster R-CNN [27], Region-based Fully Convolutional Networks (R-FCN) [28], Feature Pyramid Networks (FPN) [29], Mask R-CNN [30] etc. Each region-based method follows a two-step process. In the initial step, an attention mechanism scans the entire image, mimicking the operation of a human brain. In the second step, focus is directed towards the Region of Interest (ROI). In contrast, regression-based methods like MultiBox [31], AttentionNet [32], G-CNN [33], Different versions of YOLO [34], Single Shot Detector (SSD) [35], Deconvolutional Single Shot Detector (DSSD) [36], and Deep Supervised Object Detectors (DSOD) [37] use regressive technique in a single stage for drawing the bounding boxes. However, aforementioned algorithms have primarily been developed for the detection and/or tracking of specific entities such as persons, cars, Unmanned Aerial Vehicles (UAVs), animals, and more. The use of any of these tracking techniques as a VNA marker detector (a representative entity of tuning

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:51:25 UTC from IEEE Xplore. Restrictions apply.

parameter variation) still remains unexplored and lacks published research.

The current research proposes a novel and robust method for tracking the variation in the tuning parameters on the Smith chart using YOLOv5. To train the DL model, a meticulously labeled dataset containing approximately 5000 images was specifically created (uploaded at IEEE dataport [38]). To increase the accuracy, a pre-processing stage was carried out to select the Region of Interest (ROI). The accuracy was further improved by incorporating multilevel thresholding in the pre-processing phase. The performance of the proposed method was compared with state-of-the-art tracking techniques. The accuracy of the results serves as compelling evidence for the robustness of the proposed methodology.

The remainder of the paper is structured as follows: Section II presents the proposed marker tracking methodology. The results of the current research work are discussed in Section III, along with a comparative analysis with other state-of-the-art tracking methodologies. Section IV provides the conclusion of the research conducted, along with a delineation of future prospects and potential areas for further exploration.

II. Proposed Methodology

A. Dataset generation

For tuning the filter using the marker tracking approach, a labeled marker dataset was created using Keysight Fieldfox RF Analyzer 9914A – 6.5 GHz [39]. The dataset images were taken whilst the cavity filter was being manually tuned. After collecting the images, they were labeled. The total dataset comprises of 4967 labeled images. The procedure for gathering the images for dataset is depicted in Fig. 1.

Fig. 1. Dataset generation while manual tuning the filter. A smartcamera looking at VNA is capturing the frames

The bifurcation of the dataset in terms of number of images and training-validation-test dataset split is presented in Table I.

TABLE I. Dataset Details

Category	Number	Split
Training Dataset	3475	70%
Validation Dataset	992	20%
Testing Dataset	500	10%
Total	4967	100%

B. Deep Network Training

Before training the network, the labeled images were not preprocessed. Cross-validation stage was not utilized in this work, either. The implementation of the deep networks was performed using the system possessing the configuration Ryzen-9 5800 HS CPU, 16 GB RAM and NVIDIA - RTX3060 with 6 GB GPU.

With a batch size of 16; 640 x 640 sized images; and 100 epochs, the model was trained using YOLOv5. This model was trained with Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01 and a decaying factor of 0.005. Completion of one epoch meant that both Feed Forward (FF) and Back Propagation (BP) have finished for the entire dataset for training. With a learning rate of 0.01 it took 1 h 37 min to train the YOLOv5 model. The precision-recall curve obtained after training the deep network is illustrated in Fig. 2. The descending slope confirms the effectiveness of the network training.

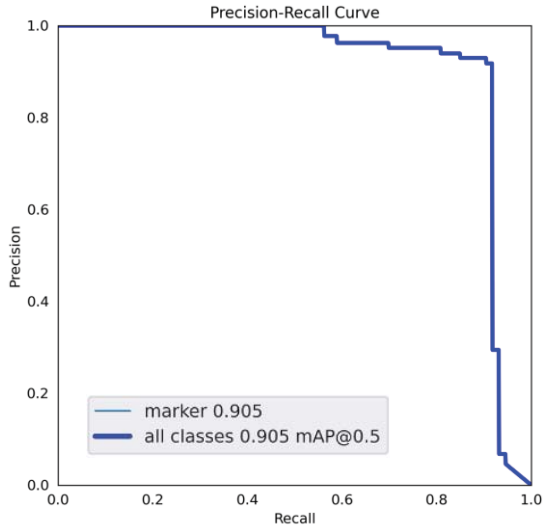


Fig. 2. The Precision-Recall Curve

C. Detection of Marker

After training of the deep network, the trained model was evaluated on the Smith chart of the VNA. A 5-pole cavity filter (filter having 5 tuning screws) that was being manually tuned was connected to the VNA. The phase change of the input reflection coefficient (S_{11}) on Smith chart plot was tracked in real-time.

For tracking the marker, a smartphone's camera was used as an IP web-camera to acquire the real-time video of a VNA's Smith chart plot. Adding a camera into the system allows us to present the proof of concept without accessing or changing the commercial VNA's software. As mentioned in Section II B, the dataset was trained on the raw images. However, for testing the model, the ROI was selected to remove the noise, thus, increasing the accuracy of results. Though some incorrectly classified marker occurrences were seen, the tracking results were considered to be encouraging. Therefore, another segmenting approach i.e. a multi-level thresholding step was added as an additional pre-processing step to further improve the accuracy of the marker tracking. The overall tracking scheme is presented in Fig. 3.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:51:25 UTC from IEEE Xplore. Restrictions apply.

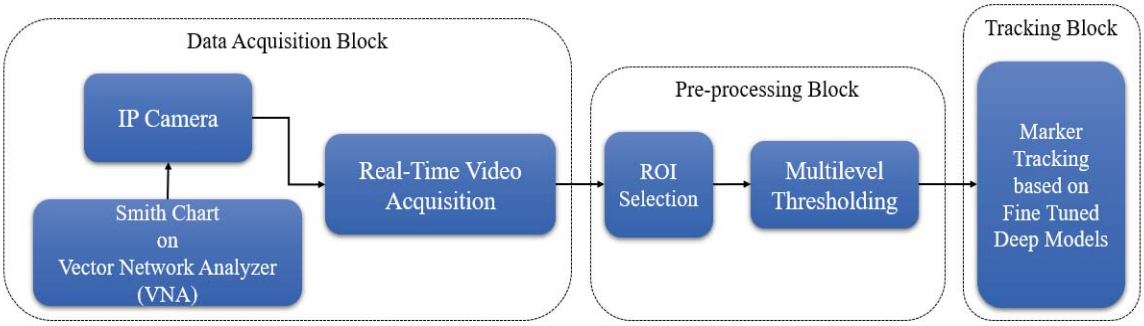


Fig. 3. Tuning Parameter Variation Detection Process

Adding multi-level thresholding assisted in tracking the marker accurately. The multi-thresholding technique also ensured that the marker is clearly visible throughout the filter tuning and is not obscured by several circles on the VNA. The resultant images after applying the pre-processing steps on the acquired images are presented in Fig. 4 and Fig. 5.

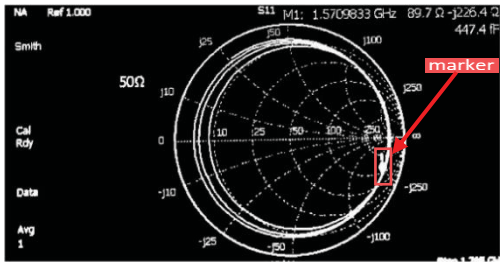


Fig. 4. Defining the marker entity and its position before initiating the tuning (VNA not calibrated)

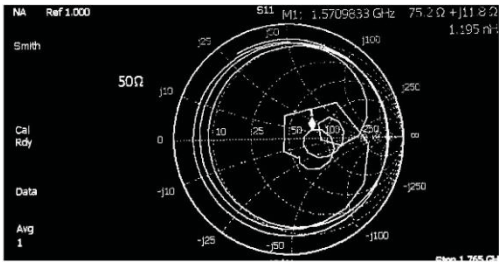


Fig. 5. Marker position when the last screw is about to be tuned

Fig. 4 presents the initial position of the marker prior to tuning the filter and Fig. 5 presents the scenario when the last screw is nearly tuned.

The overall setup for tracking the marker in real-time is displayed in Fig. 6. The marker is being tracked on the considered system while the filter is being manually tuned.

Fig. 6. The overall setup for tracking the marker in real-time

III. Results and Comparative Analysis

After video acquisition and preprocessing, the movement of marker within the ROI was tracked while tuning the filter. The trained DL-based model drew a bounding box around the center frequency marker as depicted in Fig. 7 and Fig. 8. The marker tracking result shown in Fig. 7 is for the case when the first screw was tuned starting from the marker's initial position (shown in Fig. 4). All the tuning screws were tuned sequentially in this work, starting from the filter's input side to its output side i.e. from first resonator to the last resonator. Fig. 8 depicts the marker being tracked while the filter's final tuning screw is turned (all the previous screws have been tuned). A video showing the real-time marker tracking on the Smith chart is publicly available[1].

Fig. 7. Marker being tracked while tuning the first screw is getting tuned

Fig. 8. Marker tracking when last screw is tuned

The green lines in Fig. 7 and Fig. 8 show the marker tracking results w.r.t. a reference line. Among both green lines, the reference line is the one which is drawn between the center of Smith chart (50 Ω point) and the centroid of marker at its initial position (shown in Fig. 4). This reference line holds its position throughout the tracking process. When the filter is tuned manually, the second green line tracks the marker movement on the VNA's Smith chart in real-time.

The output of the YOLOv5-based model was compared to those of the Faster R-CNN and SSD techniques. The specifics of training the YOLOv5 model are previously covered in

[1] https://drive.google.com/file/d/1wzAYKqaHbO9KfMKuJodRXBITJSt6x7-n/view?usp=sharing

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:51:25 UTC from IEEE Xplore. Restrictions apply.

Section II B. The same batch size and image size were considered for the comparative analysis. While training the Faster R-CNN model, the Stochastic Gradient Descent (SGD) optimizer with a momentum value of 0.9 and the weight adjusting parameter of 10 were used. With a learning rate of 0.001, and 1500 epochs, the Faster-CNN model took almost 5 h 40 min for completion of training. The SSD model used RMSProp optimizer and the SSD model training took 4 h 43 min when 1500 epochs, decay factor 0.9, and learning rate 0.001 were defined.

Table II provides the comparative analysis of the results. The outcomes are for the scenario in which the aforementioned pre-processing steps were also implemented while testing the different tracking methodologies.

TABLE II. COMPARITIVE ANALYSIS

Tracking Technique	Entity	Value
With YOLOv5 (Proposed)	mAP	98.2%
	Precision	97.7%
	Recall	97.3%
Faster R-CNN [27]	mAP	95.8%
	Precision	94.7%
	Recall	93.4%
Single-Shot Detector (SSD) [35]	mAP	94.0%
	Precision	93.2%
	Recall	93.1%

As evident from Table II, the mAP, Precision and Recall values of the model trained using YOLOv5 are better as compared to Faster R-CNN and SSD techniques. Using YOLOv5 tracking, the maximum mAP of 98.2% was attained during testing. While testing the trained YOLOv5 model, only one misclassification occurrence is seen in the video link mentioned in footnote 1 (on previous page). By increasing the dataset, the tracking accuracy can further be increased.

To test the robustness of the proposed methodology, the tracking was tested on polar display of the marker without implementing the pre-processing techniques used. The overall performance got reduced by approximately 1.27% when pre-processing step was excluded. As evident from Fig. 9, the confidence level of the marker also went down when compared to Fig. 7 and Fig. 8. The polar display of the VNA shows the concentric reflection circles (whereas the Smith chart display presents superimposed resistance and reactance circles). The tuning parameter tracking results for the polar chart are made publicly available[2]. Even though the dataset was generated using the Smith chart display, the proposed methodology nevertheless produced promising tracking results on a VNA polar plot, demonstrating its robustness.

Fig. 9. Tracking the tuning parameters in polar plot without implementing the preprocessing block presented in Fig 3.

During the evaluation of the methodology on a polar chart, it was observed that the accuracy of marker detection was lower compared to that on the Smith chart. Furthermore, there were occasional misclassifications and brief instances where the marker disappeared from the frame. By incorporating a polar chart image dataset and expanding the number of images in the dataset, the efficiency of tracking-based filter tuning can be enhanced.

IV. CONCLUSION AND FUTURE SCOPE

Accurate marker tracking is pivotal for achieving efficient tuning of the MW filters. This research focused on developing a robust vision-based technique to track the marker position in the Smith chart of the VNA. A labeled dataset was curated to train the deep network. While testing the proposed methodology on a cavity filter with 5 tuning screws, YOLOv5 showed the highest mAP value of 98.2% when compared with other state-of-the-art tracking techniques. The methodology was also tested on the VNA's polar chart without implementing the pre-processing stage and without creating the dataset specifically related to polar chart display of the VNA. The demonstrated accuracy of the results serves as testament to the robustness of the proposed method.

In future, this research will be expanded to encompass automated filter tuning through a marker tracking-based approach. Researchers also aim to curate a more extensive dataset, which will serve to improve the tracking accuracy and boost the marker's confidence level while tracking. Additionally, a specific module will be introduced to commercial VNA's software for tracking the tuning parameter variations.

REFERENCES

[1] R. J. Cameron, "Computer-Aided Diagnosis and Tuning of Microwave Filters," in *Microwave Filters for Communication Systems: Fundamentals, Design and Applications, Second Edition*, 2018, pp. 609–642.

[2] H. T. Hsu, H. Yao, K. A. Zaki, and A. E. Atia, "Computer-Aided Diagnosis and Tuning of Cascaded Coupled Resonators Filters," *IEEE Trans Microw Theory Tech*, vol. 50, no. 4, pp. 1137–1145, 2002.

[3] H. Hsu, H. Yao, K. A. Zaki, and A. E. Atia, "Computer-Aided Diagnosis and Tuning of Cascaded Coupled Resonator Filters," *IEEE Trans Microw Theory Tech*, vol. 50, no. 4, pp. 1137–1145, 2002, doi: 10.1109/22.993417.

[4] G. Pepe, F. Gortz, and H. Chaloupka, "Computer-Aided Tuning and Diagnosis of Microwave Filters using Sequential Parameter Extraction," in *2004 IEEE MTT-S International Microwave Symposium Digest, Fort Worth, TX, USA*, 2004, pp. 1373–1376. doi: 10.1109/MWSYM.2004.1338824.

[5] M. Meng and K. L. Wu, "An Analytical Approach to Computer-Aided Diagnosis and Tuning of Lossy Microwave Coupled Resonator Filters," *IEEE Trans Microw Theory Tech*, vol. 57, no. 12, pp. 3188–3195, 2009, doi: 10.1109/TMTT.2009.2033868.

[6] W. Meng and K. L. Wu, "Analytical Diagnosis and Tuning of Narrowband Multicoupled Resonator Filters," *IEEE Trans Microw Theory Tech*, vol. 54, no. 10, pp. 3765–3771, 2006, doi: 10.1109/TMTT.2006.881623.

[7] S. Wu, C. Liu, W. Cao, and M. Wu, "A New Method for Hybrid Diagnosis and Tuning of Coaxial Cavity Filter," in *2017 36th Chinese Control Conference (CCC 2017), Dalian, China*, 2017, pp. 9692–9696. doi: 10.23919/ChiCC.2017.8028902.

[8] J. Dunsmore, "Tuning Band Pass Filters in the Time Domain," in *1999 IEEE MTT-S International Microwave Symposium Digest,*

[2] https://drive.google.com/file/d/1G8V77oR8zpIWrSALvRlRqjUMv4sJ4gzw/view?usp=sharing

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:51:25 UTC from IEEE Xplore. Restrictions apply.

Anaheim, CA, USA, 1999, pp. 1351–1354. doi: 10.1109/MWSYM.1999.779638.

[9] J. Dunsmore, "Advanced Filter Tuning using Time Domain Transforms," in *29th European Microwave Conference, EuMC 1999, Munich Germany*, 1999, pp. 72–75. doi: 10.1109/EUMA.1999.338412.

[10] J. B. Ness, "A Unified Approach to the Design, Measurement and Tuning of Coupled-Resonator Filters," *IEEE Trans Microw Theory Tech*, vol. 46, no. 4, pp. 343–351, 1998, doi: 10.1109/22.664135.

[11] S. Li and P. D. Laforge, "A Post-Fabrication Tuning Method for a Varactor-Tuned Microstrip Filter using the Implicit Space Mapping Technique," in *IEEE MTT-S International Microwave Symposium, Phoenix, AZ, USA*, Pheoniz, AZ, USA, 2015, pp. 1–4. doi: 10.1109/MWSYM.2015.7167087.

[12] J. C. Melgarejo, J. Ossorio, S. Cogollos, M. Guglielmi, V. E. Boria, and J. W. Bandler, "On Space Mapping Techniques for Microwave Filter Tuning," *IEEE Trans Microw Theory Tech*, vol. 67, no. 12, pp. 4860–4870, 2019, doi: 10.1109/TMTT.2019.2944361.

[13] S. Li, X. Fan, P. D. Laforge, and Q. S. Cheng, "A Post-fabrication Tuning Method using Space Mapping and Surrogate Modeling Techniques," in *2019 49th European Microwave Conference, (EuMC 2019), Paris, France*, Paris, France: European Microwave Association (EuMA), 2019, pp. 392–395. doi: 10.23919/EuMC.2019.8910767.

[14] S. Li, X. Fan, P. D. Laforge, and Q. S. Cheng, "Surrogate Model-Based Space Mapping in Postfabrication Bandpass Filters' Tuning," *IEEE Trans Microw Theory Tech*, vol. 68, no. 6, pp. 2172–2182, 2020, doi: 10.1109/tmtt.2020.2977022.

[15] V. Miraftab and R. R. Mansour, "Computer-Aided Tuning of Microwave Filters using Fuzzy Logic," *IEEE Trans Microw Theory Tech*, vol. 50, no. 12, pp. 2781–2788, 2002, doi: 10.1109/TMTT.2002.805291.

[16] V. Miraftab and R. R. Mansour, "Tuning of Microwave Filters by Extracting Human Experience using Fuzzy Logic," in *2005 IEEE MTT-S International Microwave Symposium Digest, Long Beach, CA, USA*, Long Beach, CA, USA: IEEE, 2005, pp. 1605–1608. doi: 10.1109/MWSYM.2005.1517011.

[17] V. Miraftab and R. R. Mansour, "Automated Microwave Filter Tuning by Extracting Human Experience in terms of Linguistic Rules using Fuzzy Controllers," in *2006 IEEE MTT-S International Microwave Symposium Digest, San Francisco, CA, USA*, 2006, pp. 1439–1442. doi: 10.1109/MWSYM.2006.249541.

[18] V. Miraftab and R. R. Mansour, "Fully Automated RF/Microwave Filter Tuning by Extracting Human Experience using Fuzzy Controllers," *IEEE Transactions on Circuits and Systems - I: Regular Papers*, vol. 55, no. 5, pp. 1357–1367, 2008, doi: 10.1109/TCSI.2008.916614.

[19] J. J. Michalski, "Artificial Neural Networks Approach in Microwave Filter Tuning," in *Progress In Electromagnetics Research M*, 2010, pp. 173–188.

[20] T. Kacmajor, P. Kant, and J. J. Michalski, "Microwave Filter Tuning for Different Center Frequencies based on Artificial Neural Network and Phase Compensation," in *2014 20th International Conference on Microwaves, Radar and Wireless Communications (MIKON 2014), Gydnia, Poland*, 2014, pp. 1–4. doi: 10.1109/MIKON.2014.6899833.

[21] J. J. Sun, S. Sun, X. Yu, Y. P. Chen, and J. Hu, "A Deep Neural Network based Tuning Technique of Lossy Microwave Coupled Resonator Filters," *Microw Opt Technol Lett*, vol. 61, no. 9, pp. 2169–2173, 2019, doi: 10.1002/mop.31866.

[22] S. B. Wu and W. H. Cao, "Tuning Model for Microwave Filter by using Improved Back-Propagation Neural Network based on Gauss Kernel Clustering," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 29, no. 8, pp. 1–10, 2019, doi: 10.1002/mmce.21787.

[23] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *IEEE Trans Neural Netw Learn Syst*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Sep. 2014, pp. 580–587. doi: 10.1109/CVPR.2014.81.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in *European Conference on Computer Vision, Zurich, Switzerland*, 2014, pp. 346–361. [Online]. Available: http://arxiv.org/abs/1406.4729v1.

[26] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Dec. 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, IEEE Computer Society, Jun. 2015, pp. 91–99. doi: 10.1109/TPAMI.2016.2577031.

[28] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," in *30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain*, 2016, pp. 379–387. [Online]. Available: https://github.com/daijifeng001/r-fcn.

[29] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy*, IEEE Computer Society, Feb. 2017. doi: 10.1109/TPAMI.2018.2844175.

[31] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Sep. 2014, pp. 2155–2162. doi: 10.1109/CVPR.2014.276.

[32] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. S. Kweon, "AttentionNet: Aggregating Weak Directions for Accurate Object Detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Dec. 2015, pp. 2659–2667. doi: 10.1109/ICCV.2015.305.

[33] M. Najibi, M. Rastegari, and L. S. Davis, "G-CNN: An Iterative Grid Based Object Detector," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2016, pp. 2369–2377. doi: 10.1109/CVPR.2016.260.

[34] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo Algorithm Developments," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 1066–1073. doi: 10.1016/j.procs.2022.01.135.

[35] W. Liu *et al.*, "SSD - Single Shot MultiBox Detector," in *14th European Conference on Computer Vision, Amsterdam, The Netherlands*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, vol. 9905. Cham: Springer International Publishing, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0.

[36] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional Single Shot Detector," *arXiv e-prints*. Jan. 23, 2017. [Online]. Available: http://arxiv.org/abs/1701.06659

[37] Z. Shen, Z. Liu, J. Li, Y. G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning Deeply Supervised Object Detectors from Scratch," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., Dec. 2017, pp. 1937–1945. doi: 10.1109/ICCV.2017.212.

[38] E. Sekhri, "Labelled Marker Dataset from Smith Chart for Image Processing based Intelligent Tuning of Microwave Cavity Filters," *IEEE Dataport*, May 15, 2023.

[39] Keysight, "N9914A FieldFox Handheld RF Analyzer, 6.5 GHz," *Official Documentation*, May 15, 2023.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:51:25 UTC from IEEE Xplore. Restrictions apply.

Soft Computing
https://doi.org/10.1007/s00500-023-09390-4

APPLICATION OF SOFT COMPUTING

A recurrent neural network-based identification of complex nonlinear dynamical systems: a novel structure, stability analysis and a comparative study

R. Shobana[1] · Rajesh Kumar[2] · Bhavnesh Jaint[1]

Accepted: 17 October 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023, corrected publication 2023

Abstract
For the purpose of identifying nonlinear dynamic systems, a compound recurrent feed-forward neural network based on the combination of feed-forward neural network (FFNN) and locally recurrent neural network is proposed. FFNNs are known to approximate any function, but since they lack memory in their structures, they are unable to reliably forecast the outputs of complex dynamical systems, which are characterized by severe non-linearities. On the other hand, recurrent neural networks (RNNs) lack simplicity (in their structure) but are able to capture the temporal relationship existing between the input–output data. In this study, an attempt is made to merge the FFNN and RNN to provide a straightforward and efficient model for capturing the complex dynamics of any nonlinear systems. Back-propagation (BP) algorithm is used to derive the weight update equations and the Lyapunov-stability principles are applied to test the stability of the proposed framework. The proposed model performance is evaluated and compared with state-of-the-art neural models by applying it for the identification of the unknown dynamics of three complex nonlinear systems. The results of the simulation demonstrate that, in comparison to other neural models, the suggested structure has provided greater prediction accuracy, better performance in the scenario of disturbance signals, and better response in the case of parameter variation.

Keywords Recurrent neural network · Identification · Lyapunov-stability analysis · Nonlinear systems

1 Introduction

The use of neural networks for dynamic system identification and control has attracted a lot of attention in recent years. If there is insufficient knowledge about the system being modeled, identification is required. In contrast to mathematical modeling, which bases the behavior of the system on the physical principles, one simply needs observed input–output data and the order of the system for identification. Neural networks can be used to readily obtain an empirical model, rather than a mathematical model, to characterize the system once input–output data have been seen and the system's order has been established (Coban 2013). The limits of conventional techniques to nonlinear, resilient, and adaptive control of uncertain complex systems have been increasingly addressed by the application of intelligent and cognitive control mechanisms. Although uncertain complex systems are undoubtedly difficult to regulate, such intelligent controllers have achieved significant theoretical advancements. For example, while NN and fuzzy systems are essentially intelligent architectures inspired by the way the human mind and brain work, they are also architectures with a shown capacity for universal approximation, and they have frequently been utilized to create controllers with theoretically verifiable stability (Baghbani et al. 2018). According to Mohammed et al. (2018) and Mostafa et al. (2019), the neural network (NN) is a well-known supervised machine learning technology that can classify difficult and nonlinear

✉ Rajesh Kumar
rajeshmahindru23@gmail.com;
rajeshmahindru23@nitkkr.ac.in

R. Shobana
r.shobana@galgotiacollege.edu;
shobanaramasubramanian@gmail.com;
shobana.r_2k20phdee509@dtu.ac.in

Bhavnesh Jaint
bhavneshjaint@dtu.ac.in; bhavneshmk@gmail.com

[1] Department of Electrical Engineering, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India

[2] Department of Electrical Engineering, National Institute of Technology, Kurukshetra 136119, India

Published online: 17 November 2023

situations. It is a computational approach that is used to provide answers to some classification or prediction issues based on a collection of parameters that are expressed in mathematical operations (Pathiravasam et al. 2020; Kumar Chandar 2021). There are many different types of NN, such as feed forward neural networks, multi-layer perceptron, recurrent neural networks (RNN), such as the Elman and Jordan recurrent neural networks, as well as modular neural networks, radial basis function neural networks, convolutional neural networks, and self-organizing neural networks. The RNN model is utilized in numerous applications, such as forecasting financial data or power consumption, evaluating water quality, modeling nonlinear systems, adaptive control, and forecasting medical data (Yu et al. 2019; Nawi et al. 2019; Bai et al. 2019; Chen et al. 2020).

Dynamic systems serve as the fundamental structure for modeling and control of a vast variety of complex systems with research significance (Kumpati and Kannan 1990). Dynamic systems, such as heat transfer, distillation column, robotic manipulator, and Mackey glass series prediction to mention a few, have unknown non-linearity that arise at random and are challenging to model and control using linear model structures, such as ARMAX and OE models (Schoukens and Ljung 2019). There are many articles in the literature that examine the social component of supply chain resilience and sustainability (Abbasi and Choukolaei 2023; Abbasi and Erdebilli 2023; Abbasi et al. 2021, 2022, 2023a, b, c). Contrarily, the social dimension is complex and has a variety of uncertain components that cannot be adequately characterized by the traditional Boolean logic of totally true or false. Simulation of nonlinear models is one of the commonly used approaches for studying the complex nature of dynamic systems. With the increasing necessity to generate accurate models, complex models of dynamic systems with various possible configurations are being developed (Quaranta et al. 2020). To overcome this computational complexity, it is now necessary to design an accurate model that is fast and efficient in approximating the full-order dynamic system over an entire range of parameters. The advancement of technology in the field of ANN has created a subject of interest among researchers. ANN is widely been used to solve a variety of applications including image recognition, medical diagnosis, control and identification, forecasting, and speech recognition (Noël and Kerschen 2017; Quaranta et al. 2020; Basheer and Hajmeer 2000; Kroll and Schulte 2014). ANN is distinguished from other state-of-art approaches for its flexibility and self-learning ability based on experience (Basheer and Hajmeer 2000). ANN is categorized as static and dynamic models. The static type of ANN known as Feed Forward Neural Networks (FFNN) only traverses receiving signals in the forward direction. The output layer's influence does not alter the input layer's characteristics. FFNN is further classified into Multi-Layer Perceptron (MLP) and Radial Basis Function Network (RBFN). Elman Neural Network (ENN), the Hopfield neural network (HNN), Jordan Neural Network (JNN), Diagonal Recurrent Neural Network (DRNN), LSTM, and GRU networks form the dynamic models of ANN. RNN structures traverse the receiving signals in both forward and feedback direction (Ge et al. 2009). MLP is a single or several hidden layer network with weight connections between the input and output without dynamic mapping (Savran 2007). An alternate approach to MLP is RBFN, which has only one hidden layer. The output of the hidden layer is calculated by computing the radial distance between the inputs and the weights. ENN (Elman 1990), JNN (Jordan 1986), and DRNN (Kumar et al. 2017) are the extensions of MLP with the addition of a context layer that acts as memory neurons. They possess dynamic mapping by nature (Laddach et al. 2022). ENN and JNN belong to the Fully Recurrent Neural Networks (FRNN) and DRNN belongs to the group of Partial RNN (PRNN). Hopfield structures are fully recurrent structures that depend on the initial conditions due to the absence of external input (Hopfield 1982). Higher order LSTM and GRU are widely used in most of the current applications. They are known for capturing long-time dependencies and do not suffer from optimization issues as RNN (Hochreiter and Schmidhuber 1997). MLP offers better processing capabilities as compared to other structures in literature but lacks proper identification of dynamic systems due to the absence of dynamic mapping.

1.1 Related work

In Perrusquía and Yu (2021), the author has proposed a temporal convolutional network for the identification of dynamic systems. TCN is verified against MLP and LSTM. TCN and LSTM were found to give better results for large data sets and non-white noise. In Psichogios and Ungar (1992), the authors have proposed a hybrid Jordan–Elman network for a single-input single-output system. Online training and control of the CSTR plant are carried out. Extended Kalman Filter (EKF) is used as an optimization algorithm. In this paper (Şen et al. 2020), a modified Elman–Jordan structure with GA as an optimization algorithm is used. Optimization algorithms, such as EKF, and GA though perform better than BP, yet suffer long training times due to their confined search space. With the addition of white noise into the system, the optimization fails to predict accurate models. A novel Dynamic Neural Network (DNN) is proposed in this paper (Kumar et al. 2017). Though DNNs can be used with online control methods, they suffer from mapping capability due to their fixed structures. In Mohajerin and Waslander (2019), Hybrid NN based on global clustering and local learning is proposed. The clustering algorithm is used for updating model weights. Though this structure performs well, a good number of input densities is required for cluster pairing and faster conver-

gence. In Hernández et al. (2020), the authors have proposed a modified dynamic Hopfield network. The dynamic network was found to perform multi-step prediction using the past information of the system. Following the single ANN architecture, hybrid architecture resulting from the combination of the above structures is also implemented in the literature. In Alkhasawneh (2019), the authors have proposed a cascaded FFNN with ENN called HECFNN for disease prediction The results were verified on six different data sets and hybrid models are found to outperform single models efficiently. In Wang and Lin (1998), the author has used FNN, RBFNN, Runge–Kutta neural networks, and ANFIS mechanisms for the identification of nonlinear systems. Runge–Kutta ANN has shown better performance than feed-forward structures and ANFIS. In this work (Huang et al. 2021), the authors have proposed a novel hybrid deep learning model for 1 h-ahead solar forecasting. A hybrid WPD–CNN–LSTM–MLP model is designed and the proposed method is found to give better forecasting results compared to standard MLP and RNN models. The authors in Alkhasawneh and Tay (2018) have proposed a cascaded FFNN with ENN to predict six categories of diseases. The results demonstrate the higher accuracy of the proposed method over other standard ENN models. In Kalinli and Sagiroglu (2006), Elman with Nonlinear ARX is designed for system identification. The results once again show better performance with hybrid models over single models. Compound FNN (CFNN) is combined with ENN for time-series prediction of linear and nonlinear systems. The ability of the proposed model is tested on six data sets and found to perform with greater accuracy. Identification of twin rotor multi-input multi-output system is done on modified MLP and Elman structure in Toha et al. (2008). Many-to-one RNN for recommendation system is proposed by the authors in Dadoun and Troncy (2008). In this Hong et al. (2020), a total of six algorithms like Decision tree algorithm (DT), MLP, random forest algorithm (RF), gradient boosting algorithm (GB), RNN–LSTM, and CNN–LSTM were tested to predict the dam inflow. MLP has shown the best results over others. Further, a hybrid model based on ensemble methods was combined with MLP for the prediction of inflow. The author has extended the conclusion that simple machine learning algorithms along with ensemble methods produce accurate results. In Wang et al. (2023), a time-delay recursive neural network is used to develop the suggested controller. The proposed control method can be easily generalized to the actual systems, which exhibit hysteresis behavior, in contrast to those current DAC approaches established under the generic Lipschitz condition. A Hopfield neural network (HNN) estimator is then suggested as a means of online parametrization of the proposed controller. In the meantime, a modular model based on the HNN estimator is created to represent the piezo-actuated stage in this study. It consists of a linear sub-model, a hysteresis sub-model, and lumped uncertainty. In Legaard et al. (2023), the authors provided a survey of the various approaches for building neural network models of dynamical systems. In addition to providing a general overview, we evaluate the pertinent literature and list the most significant challenges this modeling paradigm faces as a result of numerical simulations. We offer a discussion on promising research areas based on the examined literature and highlighted difficulties. In de Carvalho Junior et al. (2023), the major component of the model reference control method for the rotary inverted pendulum is a recurrent para-consistent neural network. The rotating inverted pendulum is a perfect tool for applying and testing the recurrent para-consistent neural network due to its nonlinearity, two-degree-of-freedom motion, and under-actuated system. Three of these neural models are used by the developed para-consistent neural model reference controller: two of them are used to represent the arm and pendulum angles, and the third one is used to operate the system while following a reference trajectory. In Yang et al. (2023), the authors developed a stock price prediction model through neural network to enhance the stock price prediction effect based on the enhanced Particle Swarm Optimization (PSO) algorithm. To improve the global search ability of the algorithm in the early stage of evolution and the local search ability in the later stage of evolution, the adaptive adjustment of inertial weight is proposed, and the algorithm is improved by combining with neural network. This approach is based on the idea of avoiding particles falling into the same local solution as much as possible and always keeping the particles with a certain diversity. This research also builds a neural network-based stock price prediction system on the basis of the enhanced algorithm. In Villegas et al. (2023), for the purpose of predicting COVID-19 patient death, the authors used deep learning algorithms. Two datasets containing clinical data were used. Two Spanish hospitals had received 2307 and 3870 COVID-19-infected patients, respectively. First, they created a time-line of events, compiling all the clinical data for each patient and evaluating several data representation techniques. The sequences were then utilized to train an RNN model with an attention mechanism to investigate interpretability. They carried out a thorough cross-validation and hyper-parameter search, and then, they ensemble the resulting RNNs to increase sensitivity. In Zhao et al. (2023), the authors created a reduced-order machine learning model for distributed model predictive control of nonlinear processes utilizing feature selection techniques. A subset of input characteristics that significantly affects the prediction of system output is initially chosen using filter, wrapper, and embedded feature selection approaches. The creation of reduced-order RNN models utilizing only the chosen input features after integrating the feature selection techniques to capture the system dynamics. To stabilize the nonlinear system at steady state, the reduced-order RNN models are then included into

sequential and iterative distributed model predictive controls. In Hu et al. (2023), the main goal of the proposed study is to construct a Lyapunov-based economic model predictive control technique that makes use of RNNs with an online update to maximize the economic advantages of switched non-linear systems according to a predetermined switching schedule. To increase model prediction accuracy, we first create an initial offline-learning RNN using operational data from the past. We then update RNNs with data from the present. For RNNs updated online using independent and identically distributed (i.i.d.) and non-i.i.d. data samples, the generalized error bounds are calculated, accordingly. Then, for switched non-linear systems accounting for the RNN generalized error constraint, probabilistic closed-loop stability and economic optimality are concurrently attained by introducing online updating RNNs into Lyapunov-based economic model predictive control. In Han et al. (2023a), an iterative learning model predictive control (FNN-ILMPC) for complicated nonlinear systems is based on a fuzzy neural network. A data-driven model is first created using a dynamic linearization technique that solely uses input and output data. An FNN is utilized to analyze the disturbance in the established model, since it has an unidentified disturbance term that could affect the control performance. This captures the uncertainty of the system. The development of an FNN-ILMPC approach to reduce the impact of disruptions is then done on the basis of the data-driven model discussed above. The developed controller is then shown to be capable of ensuring the stability of the closed-loop system while gradually reducing both modeling error and tracking error. Finally, the experimental findings support the superiority and efficacy of the created controller. In Han et al. (2023b), for the uncertain nonlinear systems, the data-driven robust optimum control approach is suggested. The proposed technique has three advantages: To capture the relationship between the approximation errors and the control variables, a data-driven assessment technique is first developed. After that, nonlinear systems' control performance indices can be determined inside uncertain disturbances. Second, a co-evolution technique is used to construct a multi-objective resilient optimization algorithm. The control performance can then be enhanced by obtaining reliable optimal control laws. Third, a theoretical discussion of data-driven robust optimum control's robust boundedness is presented. Analytical assurance of the control systems' stability is then possible. Last but not least, two multiple-input multiple-output second-order nonlinear systems are used to demonstrate the efficacy of data-driven robust optimum control. In Luo et al. (2021), a numerous large data-related application, including the terminal interaction pattern analysis system under research, typically uses a weighted directed networks is analyzed. It consists of extensive dynamic interactions between a great number of nodes. A dynamically weighted directed network that results is high dimensional and unfinished when the number of involved nodes rises sharply, since it is difficult to see all of their interactions at each time slot. In spite of its shortcomings, it provides extensive knowledge about the diverse behavioral patterns of the involved nodes. A unique alternating direction multipliers-based non-negative latent factorization of tensors model is proposed in this research. In Luo et al. (2023), the authors put forth a unique method for nonlinear canonical polyadic decomposition on a high-dimensional, incomplete tensor called the neural latent factorization of tensors model. Three interesting ideas are used to implement it: using the density-oriented modeling principle to create rank-one tensor series with high computational efficiency and low storage costs; treating each rank-one tensor as a hidden neuron to create an effective neural network structure; and creating an adaptive backward propagation learning strategy for effective model training. In Luo et al. (2020), to prevent accuracy loss due to premature convergence without adding to the computational load, the authors carefully examined the evolution of a particle swarm optimization (PSO) algorithm and then proposed to incorporate more dynamic information into it. This creatively led to the development of a novel position-transitional PSO algorithm.

Based on the above literature, we find the following shortcomings :

1. Feedforward models (MLP and RBFN) cannot be used as a standalone structure, since they lack the memory to record previous observations. They have to be supplied with a large number of inputs and are easily affected by external noise (Kumpati and Kannan 1990). To make MLP dynamic, they must be supplied with the system's order in advance or combined with any of the suggested ANN models above.
2. Although fully recurrent structures, such as Elman, Jordan, and DRNN, are successful in the identification of many applications, they always require modification of the original structure for better prediction accuracy. Most of the original structure causes slow convergence and is therefore unsuitable for online identification and time-series applications. The majority of the modified structure necessitates the use of a complex optimization algorithm to update their parameters, resulting in high computational complexity.
3. LSTM and GRU are very efficient at identifying complex sequential data than feedforward and standard RNN models. The major limitations of these structures are that they are more complicated and require large datasets to learn efficiently. They do not work well with highly noisy data.

4. Even with massive models and methodologies cited in the literature, the best benchmark neural architecture for the identification of dynamic systems is not available.

This has motivated us to attempt to improve the architecture of ANN by designing a novel hybrid neural network for the identification of dynamic systems. A hybrid neural network called CRFNN is proposed in this work. The main motivation and outcome of this work are as follows:

1. An effective structure for dynamic system identification is proposed. The hybrid structure takes advantage of both the feed-forward and feedback structure (fast processing and the presence of memory neurons). With the addition of dynamic LRNN, it also overcomes the shortcomings of static MLP.
2. The proposed model is independent of the order of the plant and takes minimum inputs for prediction (present value of external input and one-time delayed value of the plant).
3. Gradient descent-based BP algorithm is used to develop a weight update equation and the convergence is proved using Lyapunov stability principles.
4. The performance of the proposed structure is evaluated by comparing it with two identification approaches on the benchmark problems. The hybrid structure is found to give better accuracy and prediction of dynamic models as compared to LRNN and FFNN single models.
5. The proposed structure is also analyzed for robustness and parameter variation for various plant complexity. The structure is found to have less mean square error and better accuracy than single models. The performance of the structure is found perform equally well as other suggested methods in the literature.
6. The proposed structure can be identified both online and offline, and hence can be extended for adaptive control of nonlinear dynamic systems.

The rest of the paper is organized as follows: The introduction section discusses the shortcomings of single models and the need for hybrid models. The novel works carried out by different researchers are also elaborated. Section 2 briefs the problem statement. Section 3 discusses the identification structure, learning algorithm, and design of the proposed structure. The convergence of the update weight equations is also proved using Lyapunov stability analysis in this section. Section 4 discusses the simulation results obtained. Two examples of complex plant equations and one example of a benchmark problem are selected and the performance of the hybrid structure is discussed. In Sect. 5 conclusion based on the simulation results is discussed.

2 Problem statement

Let the nonlinear plant has $r(k)$ inputs and $y_{\mathrm{p}}(k)$ outputs. The discrete difference equation of the nonlinear plant is given by

$$y_d(k) = f[y_{\mathrm{p}}(k), y_{\mathrm{p}}(k-1), y_{\mathrm{p}}(k-2), \ldots, y_{\mathrm{p}}(k-n), r(k), r(k-1), r(k-2), \ldots, r(k-m)], \quad (1)$$

where f is the unknown differentiable nonlinear function. m and n are the orders of the plant and $n \leq m$. Here, the output depends on both present as well as past output and the external input of the plant. If CRFNN is considered an identifier, the difference equation of the identifier will be as below

$$y_{\mathrm{crfnn}}(k) = \hat{f}[y_{\mathrm{p}}(k-1), r(k)]. \quad (2)$$

The output of CRFNN depends on the present external input and one delayed output of the plant. The main objective of this work is to make $\hat{f} \simeq f$, so that the identified model becomes the exact representation of the plant. To achieve this, a series–parallel type of identification is selected and the weights are updated continuously using the BP algorithm. With training progress, the weights converge to desired values and the model starts following the desired plant trajectory with error decreasing to a very minimal value or zero. That is

$$\lim_{k\to\infty} |y_{\mathrm{p}}(k) - y_{\mathrm{crfnn}}(k)| \leq \in, \quad (3)$$

where $\in \to 0$.

3 Structural comparison of FFNN, LRNN, and CRFNN

See Figs. 1, 2, and 3.

3.1 Feed forward neural networks

The structure of MLP is as shown in Fig. 1. It consists of input nodes, output nodes, and weights connecting input and output nodes. The input layer has four weight connections between input and hidden layer. This structure takes present and one delayed value of external input denoted as $r(k)$ and $r(k-1)$, respectively, and present value and one delayed value of plant output denoted as $y_{\mathrm{p}}(k)$ and $y_{\mathrm{p}}(k-1)$, respectively. The input vector $X(k) = [r(k), y(k), y(k-1), r(k-1)]$. The weight vector associated between input and the hidden layer $w_{if}(k) = [w_{if_1}, w_{if_2}, \ldots, w_{if_n}]$. The hidden layer receives three weight connections from input layer and sends one weight connection to the output layer. The hidden layer vector $F(k) = [F_1(k), F_2(k), \ldots, F_n(k)]$. The weight vector

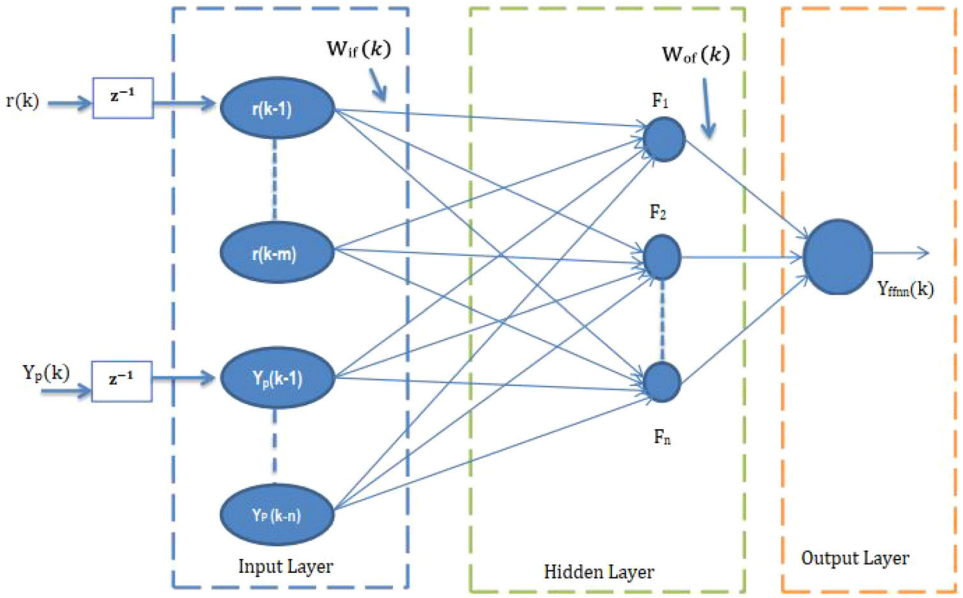


Fig. 1 FFNN structure

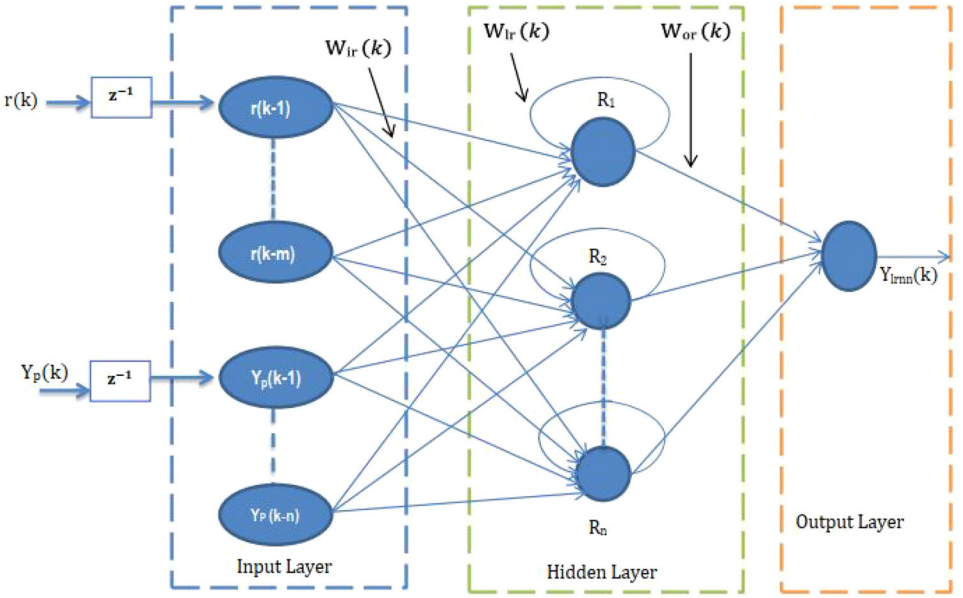


Fig. 2 LRNN structure

associated between hidden and the output layer $w_{of}(k) = [w_{of_1}, w_{of_2}, \ldots, w_{of_m}]$. The output layer computes the final output of the network by receiving signals from hidden layer. The output of FFNN is denoted as y_{ffnn}. One hidden layer is sufficient for most of the time-series prediction problem. The output of the feedforward MLP is given by

$$y_{\text{ffnn}}(k) = f_1\left(\sum_{i=1}^{m} F_i(k) w_{if_i}(k) + b_{ff_i}(k) w_{of_i}(k)\right), \quad (4)$$

where $b_{ff_i}(k)$ denotes the output bias vector and $w_{of}(k)$ denotes the associated output bias weight vector. f_1 denotes the linear activation function acting at output node. The output of the induced field is given by

$$F_i(k) = g_1\left(\sum_{i=1}^{n} X(k-i) w_{if_i}(k) + b_{fx_i}(k) w_{b_i}(k)\right), \quad (5)$$

where $b_{fx_i}(k)$ denotes the input bias vector and $w_{if_i}(k)$ denotes the associated input bias weight vector. g_1 denotes the non-linear activation function acting at hidden nodes.

3.2 Local recurrent neural networks

The structure of LRNN is as shown in Fig. 2. The LRNN is the subset of partial recurrent neural networks. The struc-

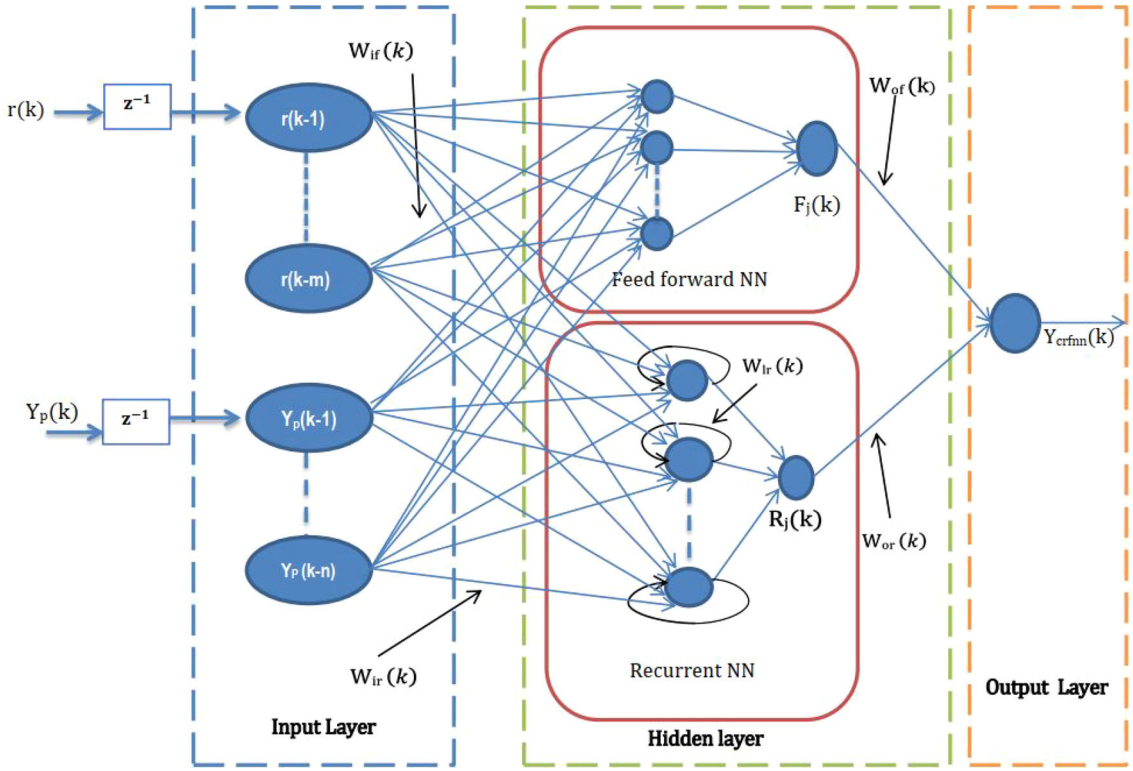


Fig. 3 Proposed CRFNN structure

ture is an extension of FFNN with dynamic memory to retain long-interval data. It consists of input layer, output layer, and local weights that act as memory neurons. The local weights sends the output of the hidden neuron as an input to same hidden neuron. The input layer has three weight connections between input and hidden layer. This structure takes present value of external input $r(k)$, present value and one delayed value of plant output denoted as $y_{\mathrm{p}}(k)$ and $y_{\mathrm{p}}(k-1)$, respectively. The input vector $X(k) = [r(k), y(k), y(k-1)]$. The weight connections of input layer are given by $w_{ir}(k) = [w_{ir_1}(k), w_{ir_2}(k), \ldots, w_{ir_n}(k)]$. The hidden layer receives three weight connections from input layer and from self hidden local neurons and sends one weight connection to the output layer. The hidden layer vector is denoted by $R(k) = [R_1(k), R_2(k), \ldots, R_m(k)]$. The weight connection between of the output layer $w_{or}(k) = [w_{or_1}(k), w_{or_2}(k), \ldots, w_{or_m}(k)]$. The local weight vector also containing the adjustable weight elements is denoted by $w_{lr}(k) = [w_{lr_1}(k), w_{lr_2}(k), \ldots, w_{lr_n}(k)]$. The output layer computes the final output of the network by receiving signals from hidden layer. The output of RNN is denoted as $y_{\mathrm{rnn}}(k)$. One hidden layer is considered in this work. The output of the LRNN is given by

$$y_{\mathrm{rnn}}(k) = f_2\left(\sum_{i=1}^{m} R_i(k) w_{ir_i}(k) + b_{or_i}(k) w_{or_i}(k)\right), \tag{6}$$

where $b_{or_i}(k)$ denotes the output bias vector and $w_{or}(k)$ denotes the associated output bias weight vector. f_2 denotes the linear activation function acting at output node. The output of the induced field is given by

$$R_i(k) = g_2\Bigg(\sum_{j=1}^{n} X(k-j) w_{ir_j}(k) + L(k) w_{lr}(k) + b_{rr_j}(k) w_{r_j}(k)\Bigg), \tag{7}$$

where $b_{rr_j}(k)$ denotes the input bias vector and $w_{r_j}(k)$ denotes the associated input bias weight vector. $L(k)$ denotes the input signal vector from local hidden neurons and $w_{lr}(k)$ denotes the associated local weight vector. g_2 denotes the non-linear activation function acting at hidden nodes. When the local weights of LRNN are made zero, the LRNN structure resembles a simple FFNN.

3.3 Hybrid CRFNN structure

To combine the merits and demerits of both the static and dynamic network, a hybrid FFNN–LRNN model is proposed in this work and it is addressed as CRFNN. Figure 3 shows the structure of CRFNN. The structure has input layer, two hidden layers, and one output layer. The input layer sends two weight connections to the hidden layer. The hidden layer

receives three weight connections from input layer and local self hidden layer and computes the output of FFNN and the LRNN. The hidden layer sends two weight connection to the output layer. The output layer computes the final network output.

1. *Input layer:* The input layer comprises two inputs, namely the present input $r(k)$ and the delayed output of the plant $y_{\mathrm{p}}(k-1)$. The input vector $X(k)$ is $X(k) = [r(k-1), y_{\mathrm{p}}(k-1)]$. These inputs are passed to both the feed-forward and recurrent paths of the structure. The input signals are passed to FNN portion of the hidden layer (indicated by green lines in the figure) through the feed forward weight connections $w_{if}(k)$ and are passed to RNN portion of the hidden layer (indicated by orange lines in the figure) through the recurrent weight links denoted as $w_{ir(k)}$.
2. *Hidden layer:* The hidden layer output is computed into two paths. In the FNN path, the output of induced field is due to the received input signals, and in the RNN path, the output of the induced field is due to the received input signals from input layer and the local self hidden neurons. Both the hidden layer outputs are acted upon by non-linear activation function and further sent to output layer.
3. *Output layer:* The output layer computes the final hybrid network output that has the characteristics of both static and recurrent networks.

The induced field of hidden layer is computed as below

$$F_j(k) = f_1\left(\sum_{j=1}^{n} X(k-j)w_{if_j}(k) + b_{fx_j}(k)w_{b_j}(k)\right) \tag{8}$$

$$R_j(k) = g_1\Bigg(\sum_{j=1}^{m} X(k-j)w_{if_j}(k) + L(k)w_{lr}(k) + b_{fx_j}(k)w_{b_j}(k)\Bigg), \tag{9}$$

where f_1 and g_1 indicates the non-linear tangent hyperbolic activation function at the hidden node. The output of hybrid CRFNN is given by

$$y_{\mathrm{crfnn}}(k) = f\Bigg(\sum_{i=1}^{m} F_j(k)w_{of}(k) + b_{oh}(k)w_{bh}(k) + R_j(k)w_{or}(k)\Bigg), \tag{10}$$

where $w_{of}(k)$ and $w_{or}(k)$ is the weight connection of FFNN and RNN to the output layer. $w_{bh}(k)$ is the output layer weight vector and b_{oh} denotes the associated output bias vector. f is the linear activation function used in the output layer.

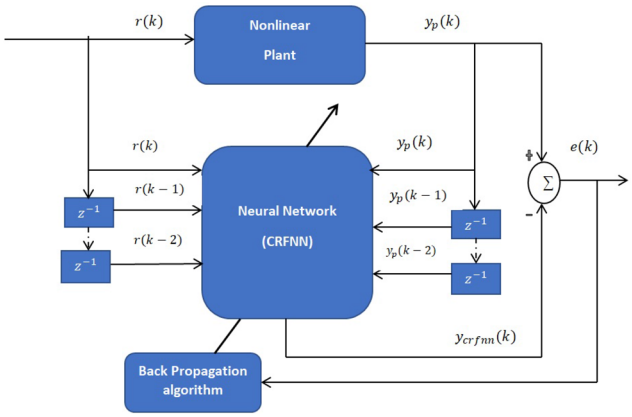


Fig. 4 Series–parallel identification configuration based on the proposed model

3.4 Learning and identification scheme for CRFNN

The system being dynamic, output is dependent on both present and past values of inputs and outputs. The identification scheme can be of parallel type or serial-parallel type. Here, series–parallel identification scheme is used to maintain the overall stability of the structure. The identification scheme used is shown in Fig. 4. Next step will be used to update the weights continuously. A gradient-based back propagation algorithm is used. The update equations are designed to follow the stability principles of Lyapunov theorem. Hence, they also give fast convergence. Now, first step will be to set up a cost function. Mean Square Error (MSE) is taken as the cost function. The MSE is defined as

$$E(k) = \frac{1}{2}[y_{\mathrm{p}}(k) - y_{\mathrm{crfnn}}(k)]^2, \tag{11}$$

where $e(k) = y_{\mathrm{p}}(k) - y_{\mathrm{crfnn}}(k)$ is identification error. The MSE is computed until the identification error reaches zero.

3.5 Update equation between output layer and hidden layer

The error is back-propagated from the output layer $y_{\mathrm{crfnn}}(k)$ to the hidden layer of FFNN $F_j(k)$ and RNN $R_j(k)$ by updating the weights of the output layer associated with FFNN $w_{of}(k)$ and RNN $w_{or}(k)$ as below

$$\frac{\partial E(k)}{\partial w_{of}(k)} = \frac{\partial E(k)}{\partial y_{\mathrm{crfnn}}(k)} \times \frac{\partial y_{\mathrm{crfnn}}(k)}{\partial w_{of}(k)} \tag{12}$$

$$\frac{\partial E(k)}{\partial w_{or}(k)} = \frac{\partial E(k)}{\partial y_{\mathrm{crfnn}}(k)} \times \frac{\partial y_{\mathrm{crfnn}}(k)}{\partial w_{or}(k)}. \tag{13}$$

This equation thus becomes

$$\frac{\partial E(k)}{\partial w_{of}(k)} = -e(k) \times F_j(k) \tag{14}$$

$$\frac{\partial E(k)}{\partial w_{or}(k)} = -e(k) \times R_j(k), \tag{15}$$

where $F_j(k)$ and $R_j(k)$ indicate the induced fields of CRFNN.

3.6 Update equation between the hidden layer and input layer

The error when back propagated further from to hidden to input layer updates the weights such as $w_{lr}(k)$, $w_{if}(k)$, and $w_{ir}(k)$. The weight update equation of hidden layer local weights $w_{lr}(k)$ are calculated as below

$$\frac{\partial E(k)}{\partial w_{lr}(k)} = \frac{\partial E(k)}{\partial y_{\text{crfnn}}(k)} \times \frac{\partial y_{\text{crfnn}}(k)}{\partial R_j(k)} \times \frac{\partial R_j(k)}{\partial w_{lr}(k)}. \tag{16}$$

Further, the weights $w_{if}(k)$ and $w_{ir}(k)$ are updated as follows:

$$\frac{\partial E(k)}{\partial w_{ir}(k)} = \frac{\partial E(k)}{\partial y_{\text{crfnn}}(k)} \times \frac{\partial y_{\text{crfnn}}(k)}{\partial R_j(k)} \times \frac{\partial R_j(k)}{\partial w_{ir}(k)} \tag{17}$$

$$\frac{\partial E(k)}{\partial w_{if}(k)} = \frac{\partial E(k)}{\partial y_{\text{crfnn}}(k)} \times \frac{\partial y_{\text{crfnn}}(k)}{\partial F_j(k)} \times \frac{\partial F_j(k)}{\partial w_{if}(k)}. \tag{18}$$

The weights are calculated using the update equations from Eqs. (12)–(15) and the stochastic gradient formula is used to derive the new weights. For example, to calculate new local weight $w_{lr}(k)(new)$, the formula is as below

$$w_{lr}(k)(new) = w_{lr}(old) - \eta e(k) \frac{\partial E(k)}{\partial w_{lr}(k)}. \tag{19}$$

Here, η denotes the learning rate and the range considered is between 0 and 1. Similarly, other tunable weights are updated using the above formula.

3.7 Stability analysis

These proposed update weight equations guarantee stability in the sense of Lyapunov. According to the Lyapunov stability theorem, when the Lyapunov-based error reaches a minimum and is positive, the system remains stable irrespective of any condition. Mathematically, it is given by

$$E(x) > 0 \text{ for } x > 0 \text{ and } E(x) = 0 \text{ for } x = 0, \tag{20}$$

Here, $E(x)$ is the Lyapunov function. The rate of change of Lyapunov error is given by

$$\dot{E} = \left(\frac{\partial E(k)}{\partial w_{of}(k)} + \frac{\partial E(k)}{\partial w_{or}(k)} + \frac{\partial E(k)}{\partial w_{lr}(k)} + \frac{\partial E(k)}{\partial w_{ir}(k)} + \frac{\partial E(k)}{\partial w_{if}(k)} \right). \tag{21}$$

When above equation is substituted with value $(x - y)^2$ for $\frac{\mathrm{d}w_{of}(k)}{\mathrm{d}t}$, $\frac{\mathrm{d}w_{or}(k)}{\mathrm{d}t}$ $\frac{\mathrm{d}w_{lr}(k)}{\mathrm{d}t}$, $\frac{\mathrm{d}w_{if}(k)}{\mathrm{d}t}$, and $\frac{\mathrm{d}w_{ir}(k)}{\mathrm{d}t}$, it becomes

$$\dot{E} = (x - y)^2 \left(\frac{\partial E(k)}{\partial w_{of}(k)} + \frac{\partial E(k)}{\partial w_{or}(k)} + \frac{\partial E(k)}{\partial w_{lr}(k)} + \frac{\partial E(k)}{\partial w_{if}(k)} + \frac{\partial E(k)}{\partial w_{ir}(k)} \right). \tag{22}$$

For various ranges of x and y, the Lyapunov-based error $\dot{E(k)} \leq 0$. This ensures that the update equations satisfy BIBO stability, and hence, the structure under test also remains BIBO stable. In 'Algorithm 1': the learning procedure is described as a pseudo code:

Algorithm 1 Learning procedure

```
procedure
    α, weights ← minimum value .
start:
    i ← no of iterations
    k ← no of samples
loop starts:
    Compute the hidden layer output
    Compute the network output.
    Calculate MSE, MAE.
    Update trainable weights .
    if cost function (k) = minimum then
        AMSE(i) = Average (MSE)
        AMAE(i) = Average (MAE)
        goto start.
        close;
    else
        goto loop starts.
```

4 Simulation results and discussion

In this section, the performance of the Compound Recurrent Feed-forward Neural network is evaluated. Two examples of real-time plant equation that is highly nonlinear and one nonlinear benchmark problem are considered. The proposed structure is supplied with external input and past output of the plant as inputs. The proposed structure is compared against neural models, such as Elman neural network (Elman) Gao et al. (1996), Jordan neural network (Jordan) Jordan (1986), Local recurrent neural network (LRNN) Kumar et al.

(2017), and Feed-forward neural network (FFNN). Average Mean Square Error (AMSE), Average Mean Absolute Error (AMAE), and Relative Mean Square (RMSE) are selected as evaluation index to measure the efficiency of the proposed structure. Various degrees of complexity of nonlinear plant equation is supplied and all the neural models considered have a single hidden layer. A fixed learning rate of 0.0001 is considered for identification.

4.1 Example 1: A nonlinear plant with order 3

To check the efficiency of the proposed structure, a real-time nonlinear plant equation is considered in Kumpati and Kannan (1990).

$$y_{\mathrm{p}}(k) = \frac{y_{\mathrm{p}}(k-1)}{1+y_{\mathrm{p}}^2(k-2)} + r^3(k-2), \tag{23}$$

where $y_{\mathrm{p}}(k)$ is the plant equation. The plant's output depends on both the present and past input–output values. The plant equation is of order 3 and follows the following identification structure:

$$y_{\mathrm{p}}(k) = f[y_{\mathrm{p}}(k-1), y_{\mathrm{p}}(k-2), r(k-2)]. \tag{24}$$

Here, f is the nonlinear function that maps the inputs and outputs. The proposed structure is applied with an external input $r(k) = \sin(\frac{2\pi k}{100})$ as well as time-delayed plant outputs. When CRFNN, FFNN, LRNN, Elman, and Jordan are chosen as identifiers, the difference equations of the identifiers are as below

$$y_{\mathrm{crfnn}}(k) = \hat{f}_1[y_{\mathrm{p}}(k-1), r(k-1)] \tag{25}$$

$$y_{\mathrm{lrnn}}(k) = \hat{f}_2[y_{\mathrm{p}}(k-1), y_{\mathrm{p}}(k-2), r(k-1)] \tag{26}$$

$$y_{\mathrm{ffnn}}(k) = \hat{f}_3[y_{\mathrm{p}}(k-1), y_{\mathrm{p}}(k-2), r(k), r(k-1)] \tag{27}$$

$$y_{\mathrm{Elman}}(k) = \hat{f}_4[y_{\mathrm{p}}(k-1), y_{\mathrm{p}}(k-2), r(k-1)] \tag{28}$$

$$y_{\mathrm{Jordan}}(k) = \hat{f}_5[y_{\mathrm{p}}(k-1), y_{\mathrm{p}}(k-2), r(k-1)], \tag{29}$$

where the symbols $y_{\mathrm{crfnn}}(k)$, $y_{\mathrm{lrnn}}(k)$, $y_{\mathrm{ffnn}}(k)$, $y_{\mathrm{Elman}}(k)$, and $y_{\mathrm{Jordan}}(k)$ denote the output of CRFNN, LRNN, FFNN, Elman, and Jordan, respectively. The inputs mentioned above are applied to the models and the output is generated. Where $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$, $\hat{f}_4$, $\hat{f}_5$ are the nonlinear functions of the respective identifier. When $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$, $\hat{f}_4$, $\hat{f}_5 \simeq f$, the network is said to follow the plant model, and the identified model is found to be accurate. About 900 samples are provided to the network for training and 400 samples are used for validation. The training is conducted in offline mode for about 900 time-epochs.

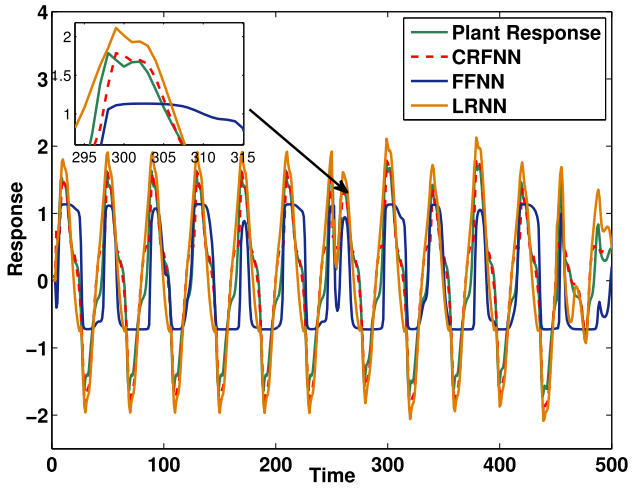


Fig. 5 Response from CRFNN, LRNN, and FFNN models [Example 1]

4.2 Parameter variation in testing phase [Example 1]

During the testing phase, a varied input pattern is supplied to the network to check the effect of parameter variation on the structure. The new input signal $r(k)$ is given as below

$$r(k) = \begin{cases} \frac{\sin(\pi k)}{40}, & \text{for } 0 < k \leq 250 \\ \frac{0.09\sin(\pi k)}{45} \\ -\frac{cos(2\pi k)}{40}, & \text{for } 250 < k \leq 450 \\ \frac{0.3\sin(2\pi k)}{15} \\ +\frac{0.1\sin(2\pi k)}{320} \\ +\frac{0.6\sin(2\pi k}{40}, & \text{for } 450 < k \leq 00. \end{cases} \tag{30}$$

Figure 5 shows the response obtained from CRFNN and other selected neural structures. This concludes that the proposed identifier performs superior to other selected neural models. The error is also found to decrease to a very minimum value with very less computational time. The performance of the identifiers is evaluated against major performance parameters, such as Average Mean Square Error (AMSE), Average Mean Absolute Error (AMAE), and Root Mean Square Error (RMSE). Figure 6 shows the MSE curve obtained from CRFNN and other selected neural structures. Figure 7 shows the MAE curve obtained from CRFNN and other selected neural structures. From the figures, it can be seen that the proposed CRFNN model is capable of extracting the dynamics of the system.

4.3 Random sine wave noise injection test [Example 1]

Now, a random sine wave was added to the network as external noise between the time interval $250 < k < 450$. A sine

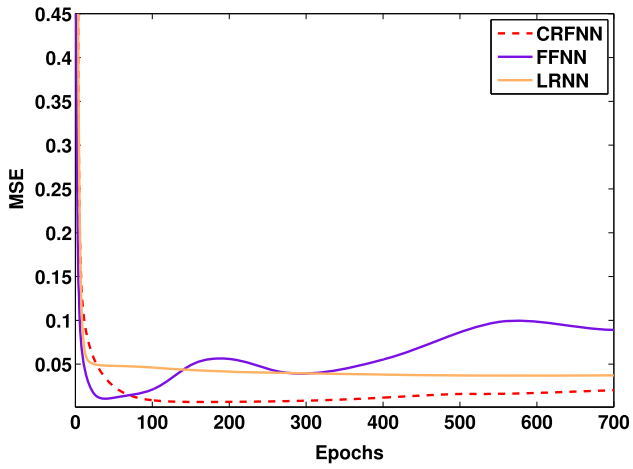


Fig. 6 MSE curves obtained from CRFNN, LRNN, and FFNN models [Example 1]

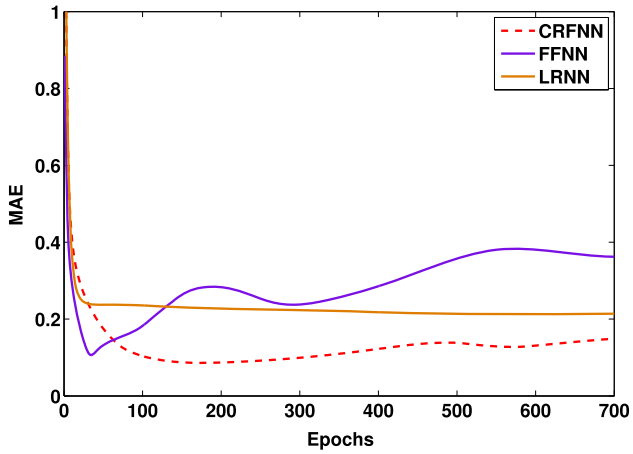


Fig. 7 MAE curves obtained from CRFNN, LRNN, and FFNN models [Example 1]

wave with a value of $\sin(\frac{2\pi k}{15})$ was added. The network initially fluctuated during this interval and deviated from the plant response. However, the network still managed to track back the plant response in a very short time. The proposed network is found efficient in learning the input–output patterns and predicting the similar unknown patterns supplied to them. The response of the proposed CRFNN under the effect of random noise is shown in Fig. 8. Table 1 gives the comparison of proposed CRFNN with Elman, Jordan, FFNN, and LRNN. From the table, the proposed structure is found to have better performance indices.

4.4 Example 2: A nonlinear plant with order 4

The proposed method is further tested for identification ability by applying another nonlinear plant equation of order 3. The plant equation is given below as in Kumpati and Kannan (1990)

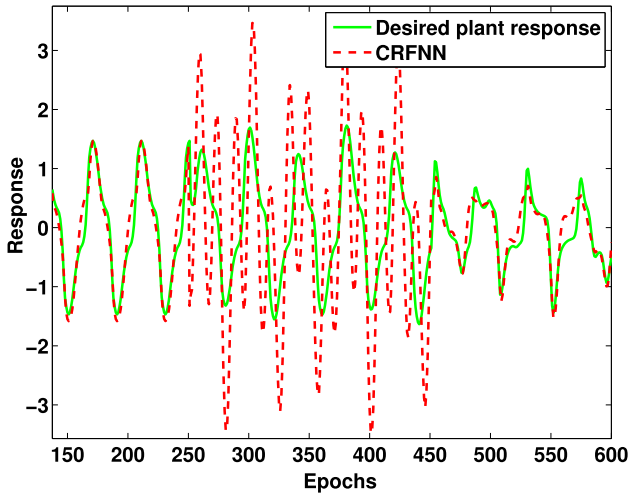


Fig. 8 Effect of random sine noise [Example 1]

$$y_p(k) = 0.72y_p(k-1) + 0.025y_p(k-2)r(k-1) + 0.001r^2(k-2) + 0.2r(k-3). \quad (31)$$

The plant equation takes the series–parallel identification form as below

$$y_p(k) = g[y_p(k-1), y_p(k-2), (k-1), r(k-2), r(k-3)], \quad (32)$$

where g is the nonlinear function. To predict $y_p(k)$, both time-delayed values of external input and that of the plant are used. When CRFNN, LRNN, FFNN, Elman, and Jordan are selected as neural identifiers to map the nonlinear function g, they are defined to take the following identification model forms:

$$y_{cfrnn}(k) = \hat{g}_1[y_p(k-2), r(k)] \quad (33)$$

$$y_{lrnn}(k) = \hat{g}_2[y_p(k-2), r(k-1), r(k)] \quad (34)$$

$$y_{ffnn}(k) = \hat{g}_3[y_p(k-2), y_p(k-1), r(k-2), r(k-3)] \quad (35)$$

$$y_{Elman}(k) = \hat{g}_4[y_p(k-2), r(k-1), r(k)] \quad (36)$$

$$y_{Jordan}(k) = \hat{g}_5[y_p(k-2), r(k-1), r(k)], \quad (37)$$

where $\hat{g}_1, \hat{g}_2, \hat{g}_3, \hat{g}_4, \hat{g}_5$ are the nonlinear functions to be identified. The models are tested by supplying about 900 samples in batch mode of identification. From the results, it can be seen that the proposed method shows the best efficiency compared to other selected neural models.

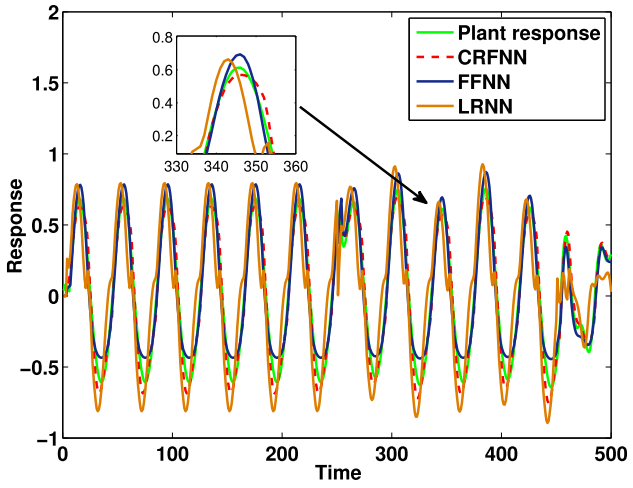


Fig. 9 Response from CRFNN, LRNN, and FFNN models [Example 2]

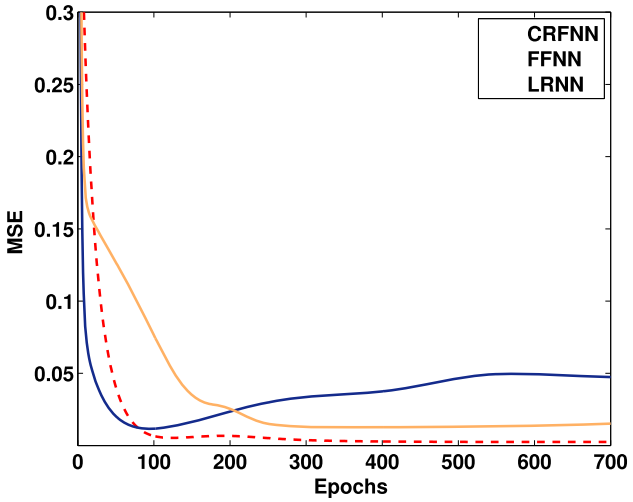


Fig. 10 MSE curves obtained from CRFNN, LRNN, and FFNN models [Example 2]

4.5 Parameter variation in testing phase [Example 2]

During validation, a multi-varied input $r(k)$ is supplied to the network as below

$$r(k) = \begin{cases} \frac{\sin(\pi k)}{40}, & \text{for } 0 < k \leq 250 \\ \frac{0.09\sin(\pi k)}{45} - \frac{cos(2\pi k)}{40}, & \text{for } 250 < k \leq 450 \\ \frac{0.3\sin(2\pi k)}{15} + \frac{0.1\sin(2\pi k)}{320} + \frac{0.6\sin(2\pi k}{40}, & \text{for } 450 < k \leq 900. \end{cases} \quad (38)$$

Figure 9 shows the response obtained from all the selected neural structures. The error is found to decrease with time to a very minimum value. The performances of FFNN are also found to good as the proposed structure for this example, yet they have a large computation time and require neurons compared to the proposed structure. Figure 10 shows the MSE curve obtained from selected neural structures. Figure 11 shows the MAE curve obtained from selected neural structures. From the figures, it can be concluded that the proposed CRFNN model identifies better dynamics of the system.

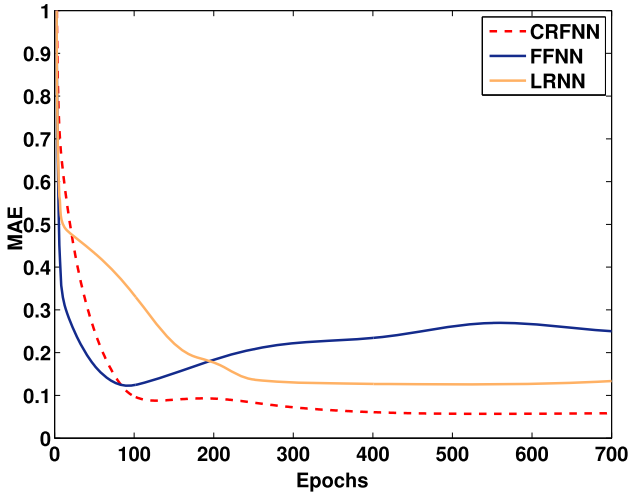


Fig. 11 MAE curves obtained from CRFNN, LRNN, and FFNN models [Example 2]

4.6 Random sine wave noise injection test [Example 2]

A random sine wave as noise is supplied to the network between $250 < k < 450$ time interval. When a sine wave of value, $\sin(\frac{2\pi k}{15})$, was added to the network, the structure was initially found to fluctuate and deviate from the plant's desired trajectory. However, the structure still recovered and was found to track the desired plant trajectory. This was due to learning ability caused by the dynamic self-recurrent links in their structures. Figure 12 shows the effect of random noise on the network. Table 2 gives the comparison of proposed CRFNN with other selected identifiers. The proposed model is found to give better performance indices over Elman, Jordan, LRNN, and FFNN.

4.7 Example 3: Mackey–Glass time-series identification

Further, the proposed method is tested on the benchmark nonlinear time–series prediction problem. The well-known Mackey–Glass identification problem equation is considered in this section. The time-series prediction is given below as

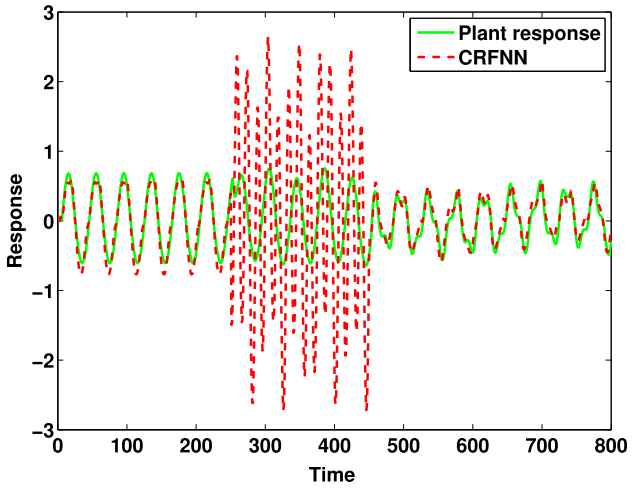


Fig. 12 Effect of random sine noise [Example 2]

from Kumar et al. (2017)

$$\frac{\mathrm{d}y_{\mathrm{p}}(t)}{\mathrm{d}t} = -\beta \times y_{\mathrm{p}}(t) + \frac{\alpha \times y_{\mathrm{p}}(t-\tau)}{1 + y_{\mathrm{p}}^{10}(t-\tau)}. \tag{39}$$

The series is applied with parameter values such as $\alpha = 0.2$ and $\beta = 0.1$. Symbol t denotes the time-series sequence of the prediction. When $\tau \geq 17$, the time-series prediction is found to have a chaotic behavior. Hence, the value of the sampling rate is selected as, $\tau = 17$. The differential equation is given by

$$y_{\mathrm{p}}(k) = -\beta \times y_{\mathrm{p}}(k) + \frac{\alpha \times y_{\mathrm{p}}(k-\tau)}{1 + y_{\mathrm{p}}^{10}(k-\tau)}. \tag{40}$$

Out of the 900 samples considered, 500 values were taken for training, and the remaining 400 values were taken for validation. The proposed identifier takes the series–parallel model form as $y_{\mathrm{crfnn}}(k) = [y_{\mathrm{p}}(k), y_{\mathrm{p}}(k-17)]$. The problem is applied to all the selected neural structures. Figure 13 shows the response obtained for the time-series prediction problem. The response proves that the proposed structure can perform superior to other selected network structures. Figure 14 shows the MSE curve obtained. Figure 15 shows the MAE curves obtained from CRFNN, LRNN, and FFNN models.

4.8 Random sine wave noise injection test [Example 3]

The problem was also introduced to a sudden disturbance of sine wave with a value, $\sin(\frac{2\pi k}{15})$ between the time-interval $250 < k < 450$ to test its learning ability. Figure 16 shows the effect of random noise on the network. The network is found to work superior and more efficiently compared to other selected neural structures. The same can be concluded from the results of Table 3. The number of neurons required to follow the desired trajectory is only 4 for the proposed structure. The other selected neural structures require 6 neurons

Table 1 Performance comparison of CRFNN with other selected identifiers [Example 1]

S. no.	Parameters	CRFNN	FFNN	LRNN	Elman	Jordan
1	No. of input parameters	02	04	03	03	03
2	No. of hidden neurons	04	06	04	04	04
3	No. of tunable weights (excluding bias weights)	10	4	6	7	7
4	Computation time	30 s	76 s	50 s	32 s	26 s
5	AMSE	**0.0139**	0.0336	0.0366	0.0230	0.0389
6	AMAE	**0.0940**	0.1994	0.1867	0.1831	0.1641
7	RMSE	**0.0107**	0.2930	0.1203	0.0915	0.0239

Table 2 Performance comparison of CRFNN with other selected identifiers [Example 2]

S. no.	Parameters	CRFNN	FFNN	LRNN	Elman	Jordan
1	No. of input parameters	02	04	03	03	03
2	No. of hidden neurons	04	06	04	04	04
3	No. of tunable weights (excluding bias weights)	10	4	6	7	7
4	Computation time	41 s	67 s	43 s	46 s	37 s
5	AMSE	**0.0137**	0.0459	0.0324	0.0174	0.0297
6	AMAE	**0.1096**	0.2045	0.1968	0.1268	0.1727
7	RMSE	**0.0446**	0.0851	0.1221	0.999	0.1112

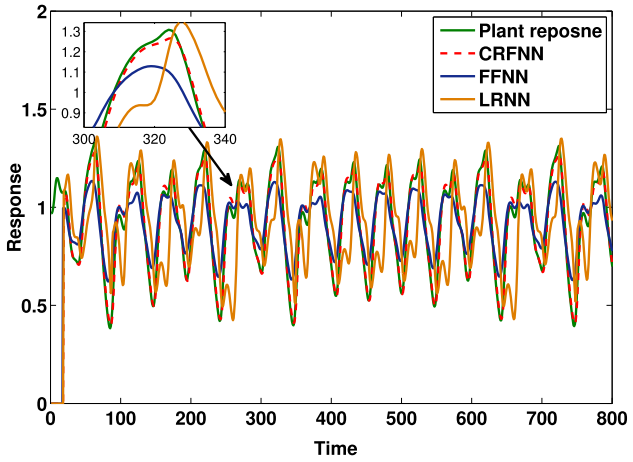


Fig. 13 Response from CRFNN, LRNN, and FFNN models [Example 3]

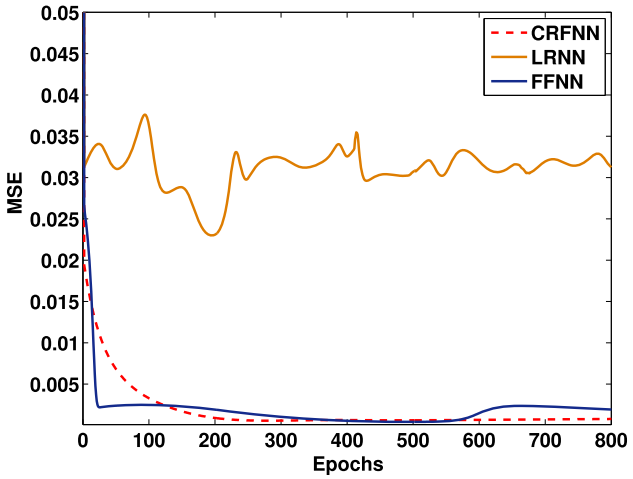


Fig. 14 MSE curves obtained from CRFNN, LRNN, and FFNN models [Example 3]

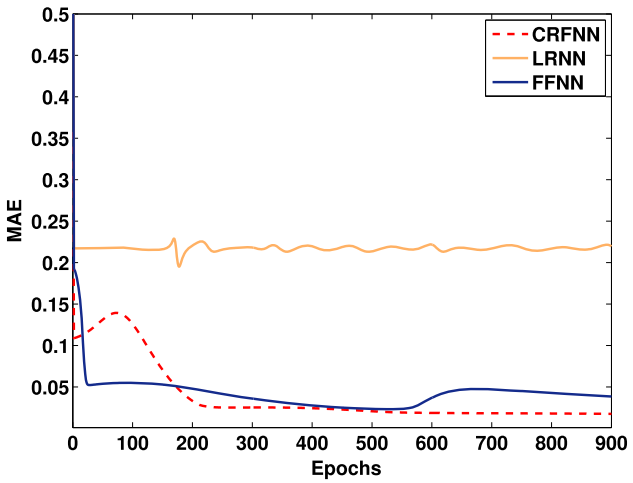


Fig. 15 MAE curves obtained from CRFNN, LRNN, and FFNN models [Example 3]

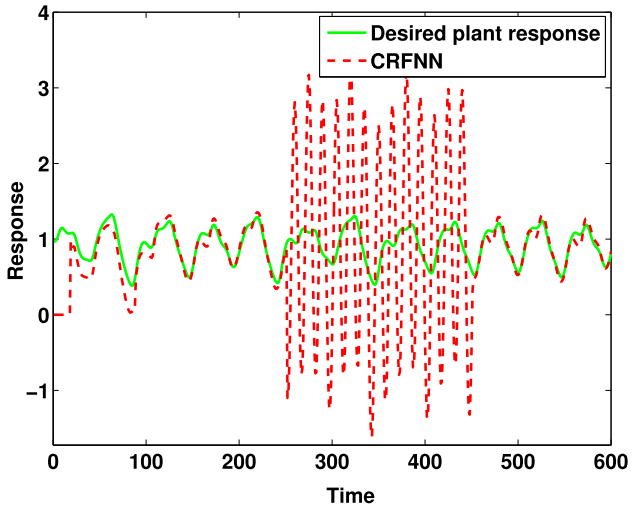


Fig. 16 Effect of random sine noise [Example 3]

to achieve the same. The overall performance parameters are also achieved minimum for the proposed Compound Recurrent Feed forward Neural Network structure.

4.9 Discussion

In this work, we have proposed a novel recurrent neural structure CRFNN for the identification of complex nonlinear dynamic systems. The performance is evaluated on three nonlinear real-time plant equations of varying degrees and complexities and is compared with Feed-forward structure (FFNN), Locally recurrent neural structure (LRNN), and the fully connected recurrent neural structures (Elman and Jordan). The results show that the proposed structure is able to effectively identify the changing dynamics of the plant. The AMSE and AMAE values obtained are far less than other selected neural structures considered in this work. The structure also required only a lesser number of inputs as compared to other structures in this work to efficiently train the tunable parameter and reach the optimum value. This is also an indication that the proposed structure is independent of the order of the plant. From the results of a random noise injection test, it is evident that the proposed structure is able to recover quickly and adjust itself to the changing dynamics of the system. These characteristics make the structure an efficient one for nonlinear identification.

5 Conclusion

In this paper, a Compound Recurrent Feed-forward Neural Network (CRFNN) is proposed for the identification of the complex nonlinear dynamical systems. The proposed structure is the hybridization of the LRNN and a single-layer

Table 3 Performance comparison of CRFNN with other selected identifiers [Example 3]

S. no.	Parameters	CRFNN	FFNN	LRNN	Elman	Jordan
1	No. of input parameters	02	04	03	03	03
2	No. of hidden neurons	04	06	04	04	04
3	No. of tunable weights (excluding bias weights)	10	4	6	7	7
4	Computation time	47 s	67 s	43 s	49 s	41 s
5	AMSE	**0.0065**	0.0324	0.0281	0.0072	0.0156
6	AMAE	**0.0755**	0.0852	0.1271	0.0852	0.1271
7	RMSE	**0.0525**	0.1778	0.1821	0.0653	0.0396

FFNN model that are combined to develop the proposed model. With the help of three benchmark non-linear problems, the structure's performance is evaluated and compared with other well-known neural models. The proposed structure is found to perform super,ior since it provided the least of the values of the error-based indicators, such as MSE, MAE, and RMSE, require lesser number of the number of input parameters, require lesser number of trainable weights. Additionally, the structure is discovered to be resistant to perturbations and parameter changes applied to the system.

5.1 Limitations and recommendations for future research

To make the proposed model simpler, the number of neurons is currently fixed equal to 4; however, this option has a significant impact on the model's overall accuracy. We may need more neurons in some systems (to be identified) to accurately define the dynamics. The development of a strategy to optimize the number of neurons in the model (depending on the system to be identified) is thus one topic of future research. To enhance the overall performance of the learning algorithm, another option may be to create new optimization techniques inspired by nature or integrate the already available optimization techniques with the BP method. The starting settings of the model's parameter can also have an impact on how well the model performs as predicted. One can anticipate the model to tune rapidly and give the desired accuracy (in a short amount of training time) if the values of these parameters are initialized correctly. The creation of an adaptive learning rate scheme is another area of research that will be the focus of our efforts in the future. This is because the rate at which a model evolves depends on its learning rate value, and evolving a model iteratively could help iterate. the model more quickly

Author Contributions All the authors have contributed equally to the study, conception, design, material preparation, and analysis. The whole draft of the manuscript was written jointly by all the authors.

Funding The authors declare that no funds, grants, or other supports were received during the preparation of this manuscript.

Data availability No data set was used in the present study.

Declarations

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

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent This research does not involve any human participants or animals.

References

Abbasi S, Choukolaei HA (2023) A systematic review of green supply chain network design literature focusing on carbon policy. Decis Anal J. https://doi.org/10.1016/j.dajour.2023.100189

Abbasi S, Erdebilli B (2023) Green closed-loop supply chain networks' response to various carbon policies during covid-19. Sustainability 15(4):3677

Abbasi S, Daneshmand-Mehr M, Ghane Kanafi A (2021) The sustainable supply chain of co2 emissions during the coronavirus disease (covid-19) pandemic. J Indus Eng Int 17(4):83–108

Abbasi S, Khalili HA, Daneshmand-Mehr M, Hajiaghaei-Keshteli M (2022) Performance measurement of the sustainable supply chain during the covid-19 pandemic: a real-life case study. Found Comput Decis Sci 47(4):327–358

Abbasi S, Daneshmand-Mehr M, Ghane Kanafi A (2023) Green closed-loop supply chain network design during the coronavirus (covid-19) pandemic: a case study in the Iranian automotive industry. Environ Model Assess 28(1):69–103

Abbasi S, Daneshmand-Mehr M, Ghane K (2023b) Designing a tri-objective, sustainable, closed-loop, and multi-echelon supply chain during the covid-19 and lockdowns. Found Comput Decis Sci. https://doi.org/10.2478/fcds-2023-0011

Abbasi S, Sıcakyüz Ç, Erdebilli B (2023c) Designing the home healthcare supply chain during a health crisis. J Eng Res. https://doi.org/10.1016/j.jer.2023.100098

Alkhasawneh MS (2019) Hybrid cascade forward neural network with Elman neural network for disease prediction. Arab J Sci Eng 44(11):9209–9220

Alkhasawneh MS, Tay LT (2018) A hybrid intelligent system integrating the cascade forward neural network with elman neural network. Arab J Sci Eng 43:6737–6749

Baghbani F, Akbarzadeh-T M-R, Sistani M-BN (2018) Stable robust adaptive radial basis emotional neurocontrol for a class of uncertain nonlinear systems. Neurocomputing 309:11–26

Bai W, Zhou Q, Li T, Li H (2019) Adaptive reinforcement learning neural network control for uncertain nonlinear system with input saturation. IEEE Trans Cybern 50(8):3433–3443

Basheer IA, Hajmeer M (2000) Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods 43(1):3–31

Chen Y, Song L, Liu Y, Yang L, Li D (2020) A review of the artificial neural network models for water quality prediction. Appl Sci 10(17):5776

Coban R (2013) A context layered locally recurrent neural network for dynamic system identification. Eng Appl Artif Intell 26(1):241–250

Dadoun A, Troncy R (2008) Many-to-one recurrent neural network for session-based recommendation. arXiv preprint arXiv:2008.11136

de Carvalho Junior A, Angelico BA, Justo JF, de Oliveira AM, da Silva Filho JI (2023) Model reference control by recurrent neural network built with paraconsistent neurons for trajectory tracking of a rotary inverted pendulum. Appl Soft Comput 133:109927

Elman JL (1990) Finding structure in time. Cognitive science 14(2):179–211

Gao X, Gao X-M, Ovaska S (1996) A modified Elman neural network model with application to dynamical systems identification. In: IEEE international conference on systems, man and cybernetics, information intelligence and systems (Cat. No. 96CH35929), Vol 2. IEEE, pp 1376–1381

Ge H-W, Du W-L, Qian F, Liang Y-C (2009) Identification and control of nonlinear systems by a time-delay recurrent neural network. Neurocomputing 72(13–15):2857–2864

Han H-G, Wang C-Y, Sun H-Y, Yang H-Y, Qiao J-F (2003a) Iterative learning model predictive control with fuzzy neural network for nonlinear systems. IEEE Trans Fuzzy Syst 31(9):3220–3234

Han H, Zhang J, Yang H, Hou Y, Qiao J (2023b) Data-driven robust optimal control for nonlinear system with uncertain disturbances. Inf Sci 621:248–264

Hernández G, Zamora E, Sossa H, Téllez G, Furlán F (2020) Hybrid neural networks for big data classification. Neurocomputing 390:327–340

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hong J, Lee S, Bae JH, Lee J, Park WJ, Lee D, Kim J, Lim KJ (2020) Development and evaluation of the combined machine learning models for the prediction of dam inflow. Water 12(10):2927

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Nat Acad Sci 79(8):2554–2558

Hu C, Chen S, Wu Z (2023) Economic model predictive control of nonlinear systems using online learning of neural networks. Processes 11(2):342

Huang X, Li Q, Tai Y, Chen Z, Zhang J, Shi J, Gao B, Liu W (2021) Hybrid deep neural model for hourly solar irradiance forecasting. Renew Energy 171:1041–1060

Jordan M (1986) Attractor dynamics and parallelism in a connectionist sequential machine. Eighth Annu Conf Cogn Sci Soc 1986:513–546

Kalinli A, Sagiroglu S (2006) Elman network with embedded memory for system identification. J Inf Sci Eng 22(6):1555–1568

Kroll A, Schulte H (2014) Benchmark problems for nonlinear system identification and control using soft computing methods: Need and overview. Appl Soft Comput 25:496–513

Kumar Chandar S (2021) Grey wolf optimization-Elman neural network model for stock price prediction. Soft Comput 25:649–658

Kumar R, Srivastava S, Gupta J (2017) Diagonal recurrent neural network based adaptive control of nonlinear dynamical systems using lyapunov stability criterion. ISA Trans 67:407–427

Kumpati SN, Kannan P et al (1990) Identification and control of dynamical systems using neural networks. IEEE Trans Neural Netw 1(1):4–27

Laddach K, Łangowski R, Rutkowski TA, Puchalski B (2022) An automatic selection of optimal recurrent neural network architecture for processes dynamics modelling purposes. Appl Soft Comput 116:108375

Legaard C, Schranz T, Schweiger G, Drgoňa J, Falay B, Gomes C, Iosifidis A, Abkar M, Larsen P (2023) Constructing neural network based models for simulating dynamical systems. ACM Comput Surv 55(11):1–34

Luo X, Yuan Y, Chen S, Zeng N, Wang Z (2020) Position-transitional particle swarm optimization-incorporated latent factor analysis. IEEE Trans Knowl Data Eng 34(8):3958–3970

Luo X, Wu H, Wang Z, Wang J, Meng D (2021) A novel approach to large-scale dynamically weighted directed network representation. IEEE Trans Pattern Anal Mach Intell 44(12):9756–9773

Luo X, Wu H, Li Z (2023) Neulft: a novel approach to nonlinear canonical polyadic decomposition on high-dimensional incomplete tensors. IEEE Trans Knowl Data Eng 35(6):6148–6166

Mohajerin N, Waslander SL (2019) Multistep prediction of dynamic systems with recurrent neural networks. IEEE Trans Neural Netw Learn Syst 30(11):3370–3383

Mohammed MA, Al-Khateeb B, Rashid AN, Ibrahim DA, Abd Ghani MK, Mostafa SA (2018) Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. Comput Electr Eng 70:871–882

Mostafa SA, Mustapha A, Mohammed MA, Hamed RI, Arunkumar N, Abd Ghani MK, Jaber MM, Khaleefah SH (2019) Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. Cogn Syst Res 54:90–99

Nawi NM, Khan A, Rehman M, Naseem R, Uddin J (2019) Studying the effect of optimizing weights in neural networks with meta-heuristic techniques. In: Proceedings of the international conference on data engineering 2015 (DaEng-2015), Springer, pp 323–330

Noël J-P, Kerschen G (2017) Nonlinear system identification in structural dynamics: 10 more years of progress. Mech Syst Signal Process 83:2–35

Pathiravasam C, Arunagirinathan P, Jayawardene I, Venayagamoorthy Y, Wang GK (2020) Spatio-temporal distributed solar irradiance and temperature forecasting. In: 2020 international joint conference on neural networks (IJCNN). IEEE, pp 1–8

Perrusquía A, Yu W (2021) Identification and optimal control of nonlinear systems using recurrent neural networks and reinforcement learning: An overview. Neurocomputing 438:145–154

Psichogios DC, Ungar LH (1992) A hybrid neural network-first principles approach to process modeling. AIChE J 38(10):1499–1511

Quaranta G, Lacarbonara W, Masri SF (2020) A review on computational intelligence for identification of nonlinear dynamical systems. Nonlinear Dyn 99(2):1709–1761

Savran A (2007) Multifeedback-layer neural network. IEEE Trans Neural Netw 18(2):373–384

Schoukens J, Ljung L (2019) Nonlinear system identification: a user-oriented road map. IEEE Control Syst Magn 39(6):28–99

Şen GD, Günel GÖ, Güzelkaya M (2020) Extended kalman filter based modified elman-jordan neural network for control and identification of nonlinear systems. In: Innovations in intelligent systems and applications conference (ASYU). IEEE, pp 1–6

Toha SF, Tokhi MO, Mlp and elman recurrent neural network modelling for the trms, in, (2008) 7th IEEE international conference on cybernetic intelligent systems. IEEE 2008:1–6

Villegas M, Gonzalez-Agirre A, Gutiérrez-Fandiño A, Armengol-Estapé J, Carrino CP, Pérez-Fernández D, Soares F, Serrano P, Pedrera M, García N et al (2023) Predicting the evolution of covid-19 mortality risk: a recurrent neural network approach. Comput Methods Prog Biomed Update 3:100089

Wang Y-J, Lin C-T (1998) Runge–Kutta neural network for identification of dynamical systems in high accuracy. IEEE Trans Neural Netw 9(2):294–307

Wang Y, Zhou M, Shen C, Cao W, Huang X (2023) Time delay recursive neural network-based direct adaptive control for a piezo-actuated stage. Sci China Technol Sci 66(5):1397–1407

Yang F, Chen J, Liu Y (2023) Improved and optimized recurrent neural network based on PSO and its application in stock price prediction. Soft Comput 27(6):3461–3476

Yu Q, Hou Z, Bu X, Yu Q (2019) Rbfnn-based data-driven predictive iterative learning control for nonaffine nonlinear systems. IEEE Trans Neural Netw Learn Syst 31(4):1170–1182

Zhao T, Zheng Y, Wu Z (2023) Feature selection-based machine learning modeling for distributed model predictive control of nonlinear processes. Comput Chem Eng 169:108074

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/s11042-023-17529-6

A survey on intelligent human action recognition techniques

Rahul Kumar[1] **· Shailender Kumar**[1]

Received: 10 October 2022 / Revised: 29 May 2023 / Accepted: 15 October 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Human Action Recognition is an essential research area in computer vision due to its automated nature of video monitoring. Human Action Recognition has several applications, including robotics, video monitoring, health care, elderly monitoring, crowd behavior and the detection of aberrant activity. This study seeks to offer the reader an up-to-date overview of intelligent human activity recognition literature and current advancements in this area. This work discusses the recent state-of-the-art research for activity recognition techniques and challenges associated with identifying human activity and discusses publicly available datasets. This work consists of an in-depth survey of numerous works published from 2010 to 2022 focusing on intelligent techniques. This article describes all steps of human action recognition along with their techniques. This study comes with the Datasets for Human Action Recognition, Handcrafted-Feature technique, Machine Learning (ML), Deep-Learning (DL), Hybrid Deep Learning and limitation of this area. This study offers a comparative analysis between ML and DL approaches to show their effectiveness in action recognition. This study examines some unexplored areas in human action recognition that can be unearthed to create a more resilient system in the presence of issues. Previous research has demonstrated that deep learning surpasses standard machine learning for recognizing human activities. This study also emphasizes the most pressing issues and research direction. All relevant datasets are described in detail. Furthermore, our opinions and suggestions for future research have been shared. Compared to past surveys, this study offers a more systematic description of Human Action Recognition methods regarding comparability, problems, and the most recent evaluation technique.

Keywords Human Activity Recognition · Deep Learning · Machine Learning · Pretrained Models · Handcrafted Feature

✉ Rahul Kumar
rahulkumar_phdco2k19@dtu.ac.in

[1] Department of Computer Science & Engineering, Delhi Technological University, Delhi 42, India

1 Introduction

Because of the enormous demand from numerous applications, like analysis of video, video searching, and health monitoring systems, human activity detection has attracted a lot of interest in recent years in the domain of vision-based approaches [1]. There are many different kinds of human activity. We can theoretically categorize human actions based on their complexity level into four tiers: human body gestures, actions, human-to-human interactions, and activities performed by a set of persons. Gestures are the atomic components describing a human's substantial motion and are the most basic movements of a human's body part. "Extending & waving an arm in the air" and "finger raising" are delicate gestures. Single human actions, such as "jogging," "hand movement," and "thrashing," may be made up of a variety of gestures that are chronologically classified—interactions between human activities that require two or more persons and objects. For example, a "two-person conversation" is a human–human contact, but "the person who steals a bag from someone else" seems to be a two-person object-human interaction. "Playing some game by any person" and "some tasks performed by a set of persons' is an examples of group activities [2].

The Action recognition system uses video sequences to understand the action and utilize the knowledge to recognize the move. Space–time information is used in vision-based methods to identify human activity by specific feature extraction in the video stream. The principal intent of human action recognition approaches is to classify the activities of humans in the sequence of videos, such as running, walking, eating, playing and kicking.

The complete step-by-step process for human activity recognition (HAR).

- Start of HAR process from object detection. First, we must identify and spill human objects from the background picture. This phase is known as object segmentation.
- The second phase in activity recognition is feature extraction, where the critical qualities of image frames are collected and expressed as features methodically. In HAR, feature extraction plays a vital role because strong features enhance the results and weaker features downgrade them. Various features are collected in the feature extraction step, so we use the dimensionality reduction technique to remove redundant features. There is a multiple dimensionality reduction technique.
- Following selecting appropriate characteristics from a video, activity detection and classification methods are considered for human activity recognition.

HAR can be classified into two classes sensor-based and vision-based. Sensor-based data is collected through various sensors like wearable and environmental sensors. The former method to study images and videos was attained with the help of optical sensors. Wearable sensors are different from optical sensors based on the data type. Optical sensors generate 2D images, 3-D images and videos compared to wearable sensors producing only single dimension data. Vision-based data extracted from video surveillance and online videos. Machine Learning is helpful in the area of Human Action Recognition to classify action accurately. Extracted features fed into Machine Learning model (KNN, SVM, many more) to classify action performed in input sequences. In the following section, we explained various Machine Learning algorithms used in detail by the researcher. The Deep Learning model can extract features automatically from input sequences and classify actions more precisely. This review gives an in-depth understanding of Deep Learning models (CNN, LSTM, and Others) and explores the use of each technique in Computer Vision.

Figure 1 shows the steps which are used in action recognition. For Action Recognition (AR), video sequences are used in the first step, and after that, the pre-processing step takes away the background information and the objects from frames are extracted. We found shape, silhouette, motion information, and others in the feature extraction step. Such knowledge is used to label activities inside the video sequence. Feature vector and feature descriptor represent extracted feature—classification algorithm. After that, we have to train the model basis on the extracted feature from sequences of images to classify each action effectively.

Categorization for human activities is as follows:

- **Gesture:** Movements of body parts like hand, legs, stretching an arm. Any verbal communication is not needed.
- **Interaction:** Interaction between humans or humans and objects. Two actors are performing activities.
- **Action:** Activity performed by a person like walking, jogging, punching, kicking.
- **Group Activity:** Group Activity is when a set of persons perform activities like playing football, protesting and other sports. Group activity is crucial in HAR due to sports action detection, crowd behavior and surveillance.

Intelligent techniques have improved in the domain of HAR, like ML and DL methods and introduced the finest method for extracting deep features from the data to recognize the activity accurately. This survey needed to cover all advancements in intelligent techniques in HAR. Numerous works related to computer vision have been discussed and observed how Machine Learning (ML) or Deep Learning (DL) models are employed for activity detection. The primary purpose of this learning is to help the computer vision research community adequately explain the HAR methods. This survey outlines significant contributions to computer vision and explains state-of-the-art methods with their scope. The survey briefly describes each phase of the HAR process: handcrafted features technique, ML methods, DL methods, hybrid DL methods, pre-trained DL methods, vision dataset, HAR

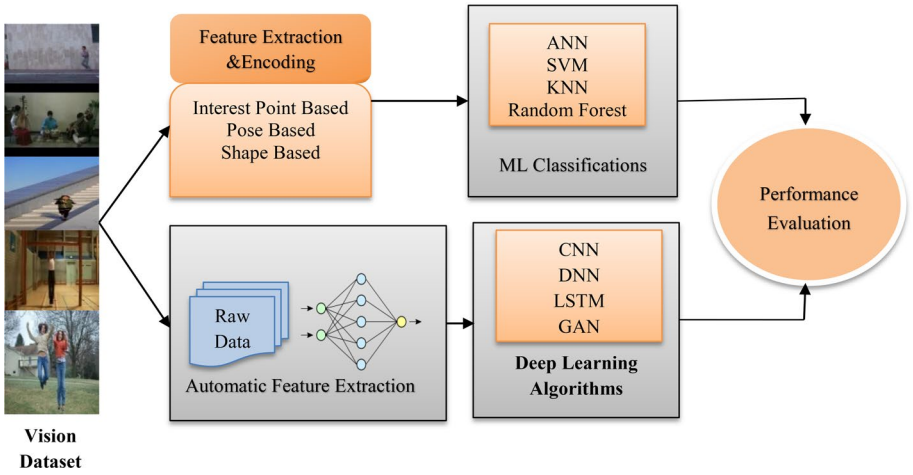


Fig. 1 The flow of activity recognition by employing handcrafted and automated feature extraction approach

challenges, limitations, and future direction. Various research challenges in HAR require efficient techniques to address feature extraction methods, future action prediction, overlapping activity classification and group activity detection. In a subsequent section, this study discusses various approaches to resolve these challenges with the help of the ML approach, DL approach, and hybrid approach suggested.

No prior surveys have discussed the scope of advancement and in-depth complexity analysis of each approach in HAR from the last decade. This comprehensive survey discusses all areas of HAR with handcrafted feature extraction techniques (global feature, local feature, semantic feature), ML-based method, DL-based method, Transfer learning-based method, Automatic feature selection method, Hybrid method and Vision Transformer method for recognizing action. This paper contributes to the summary of publicly available datasets, the progress of HAR approach advancement, computational challenges, limitations and future direction.

This article summarizes as follows: Section 2 of this study briefly describes the Human Action Recognition process with framework & its application. Section 3 summarizes survey studies on the HAR tasks that have been published. Section 4 discusses taxonomy for activity recognition using handcrafted feature techniques. It provides a brief knowledge about approach-based activity recognition and how they recognize activity from video sequences and compare various methods. Section 5 discusses available datasets for vision-based activity recognition and provides limitations of each data set. Section 6 discusses challenges in activity recognition. Section 7 discusses the thoroughness of DL techniques for AR and summarizes various authors' contributions to Human Activity Recognition. Section 8 gives a brief idea about the limitations and research gaps in the field of HAR. Section 8 summarizes the whole study's contribution and future scope.

2 Human activity recognition process & applications

Human Activity Recognition phases: data collection, pre-processing, model training, and performance evaluation are the four primary phases of the HAR process. A HAR device is chosen in the first phase according to the desired application. For instance, the camera serves as the HAR device for data collection in a surveillance application involving numerous people. A monitoring sensor is used for data collection in terms of daily activity. Consider that captured data might occasionally include noise or other problems, such as crowded backgrounds or poor picture quality, which makes post-processing intelligent-based systems difficult. Thus, the feature extraction method is crucial for a better recognition algorithm. Figure 2 shows all the steps which we follow in activity recognition.

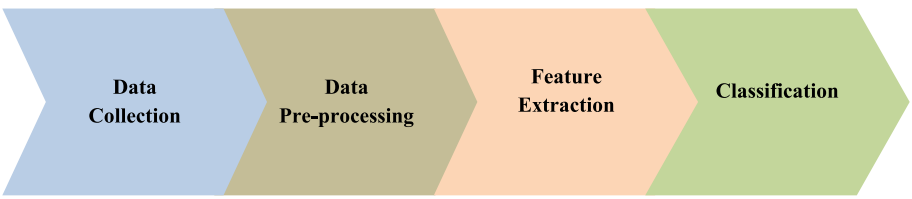


Fig. 2 Human activity recognition process

2.1 Data pre-processing

In phase 2, data pre-processing is performed. This phase is crucial for building models, especially when working on vision data. The model's accuracy mainly depends on pre-processing data and how we parsed the data into the model. Data cleaning, smoothing, noise removal and grouping are the main components of data pre-processing. Pre-processing is frequently used to carry out procedures that simplify and enhance the correctness of the applied algorithm. The dataset is augmented after pre-processing technique that adds perturbed versions of images. Common affine transformations include scaling, rotation, and others. By doing this, you may expand the size of your action dataset and expose the neural network to more diverse versions of the pictures.

2.2 Feature extraction

Action recognition requires feature extraction, which is an essential component. It entails the removal of body positions and movement patterns from pictures and recordings of human activity. However, it is a complex process due to changes in size, insufficient illumination, improper views, and backdrop clutter. Numerous classes of feature extraction are:

- Interest Points: Space–time interest points (STIP), Scale-invariant feature transformation (SIFT).
- Shape-Based: Histogram of Oriented Gradients (HOG), Silhouette.
- Motion Based: Motion History Images (MHI), Motion Binary Images (MBI), Histogram of Optical Flow (HOF).
- Pose Based:2D,3D
- Trajectory Based: Harris Corner Detector, Dense Trajectory.

2.3 Classification

Phase 3 training process involves dataset using ML or DL methods. There are ML-based strategies that can be used with hand-crafted features. The DL framework may be used for automated feature extraction. Conventional ML classifiers include Support Vector Machines, Nearest Neighbor (NN) and Random Forest (RF). In contrast, DL classifiers have Convolution Neural Networks (CNN), Generative Adversarial Networks (GAN), Long-Short Term Memory (LSTM) and hybrid models. Transfer learning models are also useful and provide knowledge reusability. In further sections, these models have been described. Action learning and recognition involves learning new models trained on extracted features, identifying the segments based on action class, and evaluating them using classifiers. Following training, the fourth step is the most challenging since the model is applied to actual data whose behavioral changes depend on physical criteria such as age, physique, and technique of task execution.

In ML-based, we have to feed accurate features to a traditional ML-based classifier to learn action. Extracted features have been used to train the model and learn action from these features. DL models can extract features from the input data and understand activity accordingly.Input data to recognize activity is different, for example, vision data (still images, visuals),sensor data (depth data) and hybrid data. In vision-based action

recognition, we focused on body parts, gestures, interaction, image features, interest points, etc. There are various available input action datasets to recognize action, which we explored in Section 3.

2.3.1 HAR applications

HAR has quite a vast reach of applications, including visual surveillance stream, elderly monitoring, education, crowd behavior, retrieval of videos, and detection of abnormal activity. Action Recognition is crucial in detecting and tracking patients' daily activity in the medical and healthcare field. Because elderly peoples are vulnerable to infection, their wellness has been a significant issue. Close monitoring using automated surveillance devices is necessary for elderly individuals to detect falls or irrational behavior.

Recognizing human activities in the visual stream is essential in teaching and learning. In educational establishments, evaluating human activities from the video may enable behavioral recognition and automatic tracking of presence within the lecture. Train stations, airports, and public areas might use abnormal behavior recognition to maintain safety. Due to the abundance of security cameras, it is difficult to identify such situations.

Understanding sign language for the differently abled person is a gesture-based application of HAR. Continuous monitoring of activity data of patients and children is the application of behavior recognition. Workout analysis with the help of pose estimation.

3 Datasets

Datasets are crucial for assessing methods for a purpose intended. Numerous datasets are available to evaluate the performance of the HAR models. Occlusions; illumination variation, perspective variation, annotation, and fusion of modalities are the most common issues in these datasets. Dataset background can vary from dynamic to static for the clutter background. Overlapping issues arise in complex activities because they contain a collection of activities (playing football has running and kicking actions). Inter- and intra-class variation occurs in the dataset due to the unpredictable nature of people (some run with small steps while others take long steps). This survey gives a comprehensive summary of several HAR datasets with their number of classes, number of videos, characteristics, etc. The classification of datasets is shown in Fig. 3. According to the source of the input data, datasets can be classified into three categories.

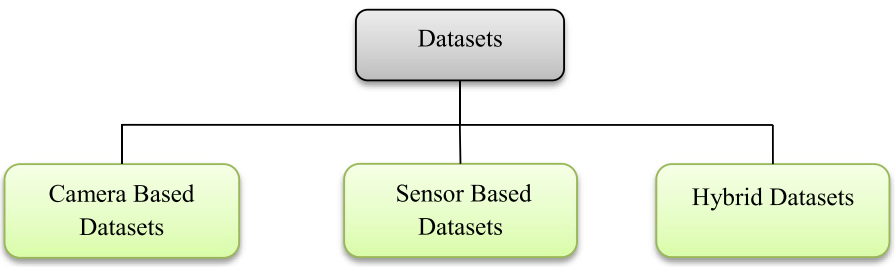


Fig. 3 Classification of datasets

4 Camera based datasets

Human traits are repeated in various contexts in the KTH [3] dataset. Human activities are divided into six categories: walking, running, running slowly, punching, waving one's hands, and cheering Fig. 4. Each action class comprises 100 sequences with 25 actors doing the acts in four distinct settings: outside, outside with various costumes, inside with varied lighting, and outside with different scales. However, since they were collected in a controlled context, the data did not accurately represent real-world activities, which is a significant drawback. Each researcher has carried out an action on the video clip in these data sets. Each movement is done essentially identically from a static point of view against a motionless background dataset, which is much more complicated than various lights and clothes but still unrealistic. Video captured in KTH with the help of the static camera. This dataset is more typical due to the dynamic background compared to the Weizmann dataset, which contains various challenges like view invariance, illumination changes, and fewer actions.

The Weizmann dataset [4] includes 90 movies of 10 movements ("walking," "running," "jumping," "galloping sideways," "bending," "one hand waving," "two hands waving," "jumping in place," "jumping jack," and "skipping") in Fig. 5. A fixed camera and a plain backdrop were used to capture 90 video sequences at a low resolution of 180 to 144 50 fps. The fundamental issue with this dataset is that it has a poor resolution and lacks variation in terms of cluttering, group interaction, camera movement, and lighting. Weizmann is a single-view dataset; an individual actor performs all activity with low-resolution frames. This dataset is not suitable for recognizing complex actions.

Diverse sports such as golf swing, horseback riding kick, lifting, running, swing bench, skating, and walking are included in the UCF Sports [5] dataset. Figure 6 shows sequences of images collected from the UCF Sports dataset. There are 150 720×480 video clips in total. Camera movement and backdrop clutter detract from the dataset's appeal. It's mostly for training, while the remainder is utilized for testing. Camera speed, object look and

Fig. 4 Different types of scenarios and actions from KTH [3] Dataset

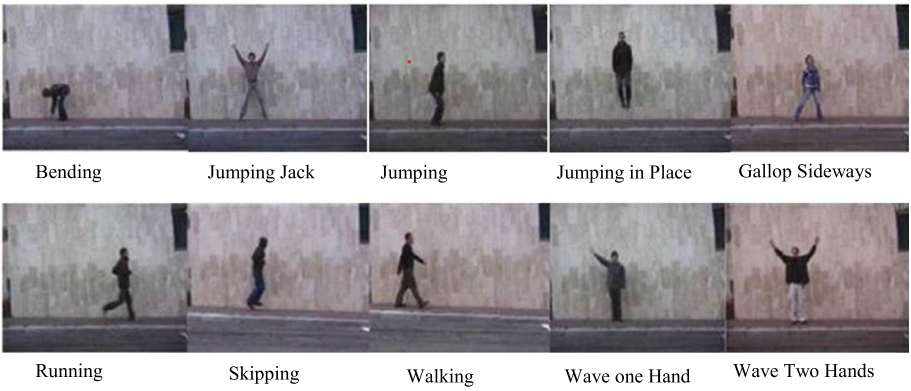


Fig. 5 Graphic representation of Vision Sequences from Weizmann Dataset [4]

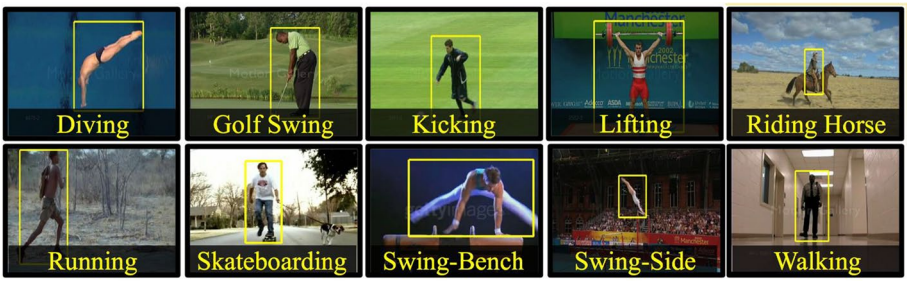


Fig. 6 Illustration of UCF Sports dataset sequences of action [5]

posture, object size, unstructured backdrop and other characteristics make dealing with the UCF sports data collecting tough. The low-resolution sequence of frames and detailed description of sequences make the more complicated to UCF Sports dataset. These videos are collected from the web, which has unrealistic conditions. Sports action requires a more realistic environment for the categorization of action.

HMDB is an extensive database for human action identification. H. Kuehne et al. [6] proposed 51 different action categories and included 101 clips for each action. Figure 7 represents action images of the HMDB dataset. At least two people verified the consistency of the video clip. More metadata enables more precise test data selection, training, and assessment of validation systems. HMDB is an extensive collection of action datasets to detect daily activity. It provides human–human interaction as well as human-object interaction. This dataset is more typical as compared to UCF Sports.

The UCF 50 [7] is a dataset for identifying actions in 50 distinct categories, including real YouTube videos. Figure 8 represents the action sequence of the UCF 50 dataset. It was developed in 2012 by the University of Central Florida's Center for Research in Computer Vision. UCF 11 was updated with an increased number of actions named UCF 50. All videos come with a realistic environment.

This dataset, known as UCF-101 [8] has 13,320 movies from 101 activity classes. It is one of the most comprehensive video databases currently accessible. It comprises 13

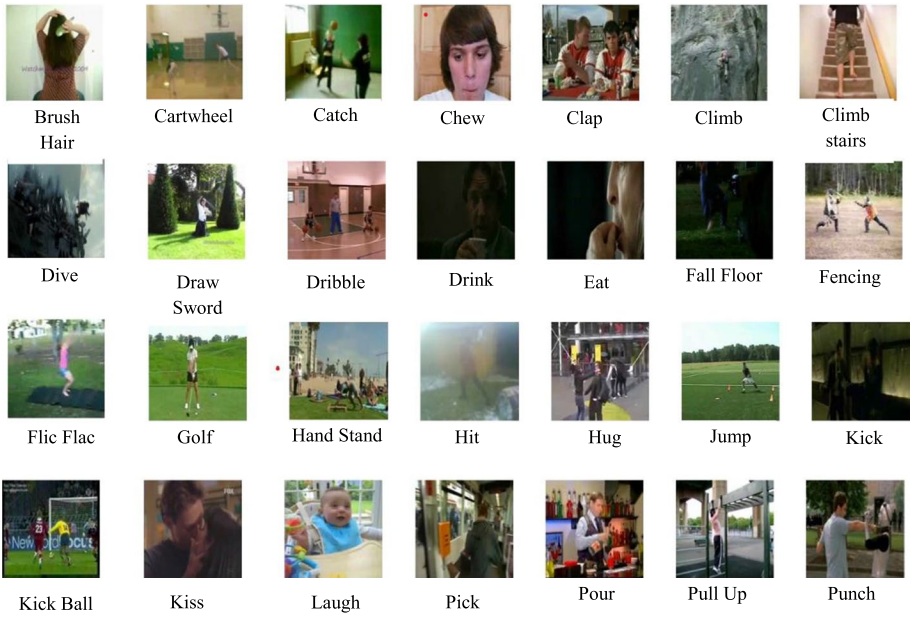


Fig. 7 Multiple action images are taken from HMDB 51 dataset [6]

Fig. 8 UCF 50 dataset frames [7]

320 motion clips from 101 realistic motion categories that were culled from the YouTube platform Fig. 9. The IM daily activity dataset [9] covers everyday routine actions captured in real-world settings such as laboratories, classrooms, and hallways, among other places. Among the 15 tasks performed by the 15 persons in the video are hands waving, bending, boxing, clapping, cleaning and eating. Also included are phone conversations and reading articles. The dataset consists of both RGB and depth images that show the silhouettes of people, as well.

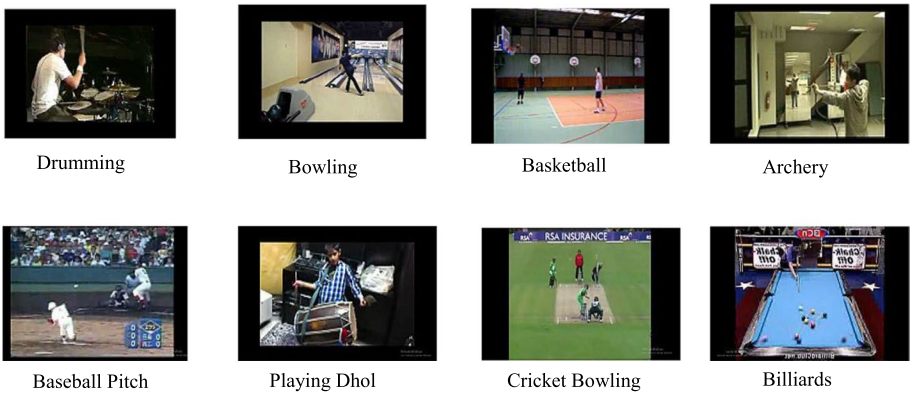


Fig. 9 UCF 101 dataset frames [8]

4.1 Sensor based datasets

The activities in the **IXMAS** dataset [9] were taken from five different perspectives, with 11 people performing 14 action types in Fig. 10. The camera angle is fixed, the background is unchanging, and the lighting is consistent. The dataset includes a collection of silhouettes and voxel representations of the underlying data. The data was gathered in indoor and controlled environments for this dataset. The camera setup is fixed, and action is performed with a change in direction. This dataset does not provide accurate detection of activity in real-life situations.

The WISDM [10] dataset used smart phone and watch data to recognize action. The data collection uses an accelerometer and gyroscope sensor. This dataset holds various activities, like sitting, jogging, and others. This dataset holds more than 5,000 activities performed by 29 subjects.

UTD MHAD [11] contains 27 different actions with a resolution of color image 640 *480 and 320*240 resolution of the depth image. One Kinect camera and one wearable inertial sensor were used to capture the action indoors. This dataset is multimodal, multi-view, and sample frame in Fig. 11.

4.2 Hybrid datasets

Heilbron et al. [12] from the University of Del Norte in Colombia created the Activity Net dataset. This massive collection aims to develop a semantic taxonomy of movies displaying human activities and cover a wide variety of topics for each of these activities. The

Fig. 10 The multi-view images from IXMAS Dataset [9]

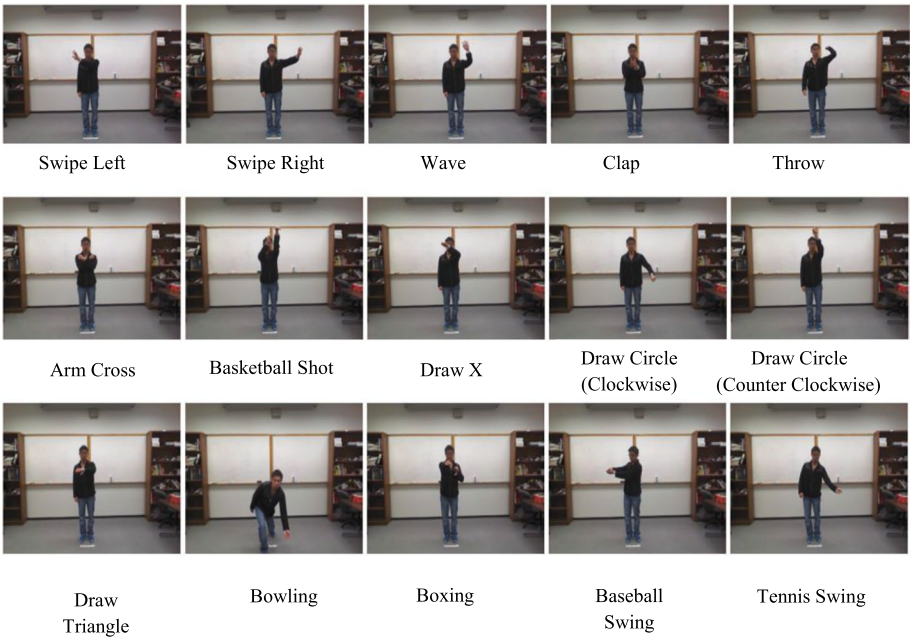


Fig. 11 UTD MHAD dataset frames [11]

Activity Net dataset is divided into three sections based on the kind of application they are intended for classifying untrimmed films, trimmed movies, and activity detection on all untrimmed videos. Figure 12 represents the action sequence images of this dataset.

The Northwestern UCLA [13] dataset contains RGB videos with accelerometer and gyroscope data, which hold various views. 1494 total sequences performed by 10 individuals have 10 different actions of variable length. For a cross-view (CV) environment, the author suggested two views for learning and one view for evaluation. UWA3D [14] holds RGB sequences,,3D joints, and depth data recorded with Kinect V1. It also stores 1070 actions performed by 10 individuals in 30 different human activities. These actions are reordered with 4 different viewpoints, like the front, left, right, and top view. Due to the variation in views and self-occlusion, this dataset is very complex.

Action samples from 60 classes and 40 people are included in the NTU-RGBD [15] action recognition dataset, including skeleton and depth data from 80 viewpoints using RGB movies. It provides depth maps, 3Djoint information, RGB frames, and IR sequences. The categorization was based on the quantity of daily acts, health-related activities, and joint actions. NTU RGB + D has many videos for daily life assistance, but these videos are

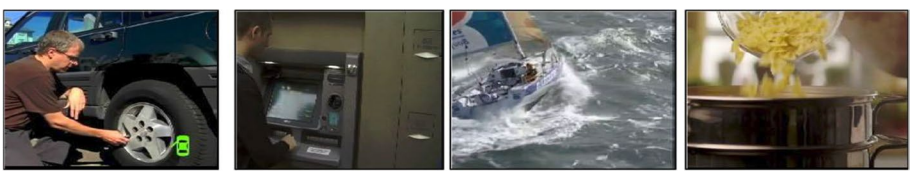


Fig. 12 Activity Net dataset video sequence [12]

captured in restricted environments and with strict guidelines. Therefore, the real-world use of this dataset is complex. The resolution and number of activities in each dataset are summarized in Table 1.

5 Challenges & issues in human activity recognition

Using these systems for surveillance, senior care, and medical monitoring, along with rising implementation costs, creates new social difficulties, such as acceptability by society, privacy concerns and other consequences of house installation, as well as large-scale implementations. Some challenges are like a list of human endeavors, action detection and prediction in congested surroundings, including simultaneously identifying several activities conducted by the same person. Many obstacles emerge in Human Action Recognition owing to poor delivery of actions such as punching. It can be regarded as a fighting activity or a friendly welcome depending on the intensity, punch speed, and other atom-level actions, such as a grin. Many networks must be examined to recognize these kinds of behaviors. With the action recognition process, there may be differences in actions achieved by the same themes and even distinct class actions may look identical. For example, running might be perceived as fast running. The model should be able to include the diversity of each category.

5.1 Variation in inter and intra class

The performance of any system is influenced by the wide range of activities it performs. Recognition of human activity should be capable of distinguishing various activities. Although the same activity may differ from subject to topic, intraclass differences occur. Interclass disparities widen as performance speed and strength improve. Alternatively, various activities, such as talking or reading, may transmit comparable forms. This scenario is known as the interclass replica.

5.2 Dynamic and clutter background

With a busy and dynamic background, it is tough to recognize the action. Video quality is the essential aspect of obtaining high system accuracy. The dynamic or crowded backdrop is a type of disruption in the video flow that occurs when the true activity of interest encounters ambiguous details [24]. Some models have been introduced that remove background information [25]. A separate background behind the Gaussian Mixture model [17] and a random forest model [26] was examined. Shadow-based representation can deal with background variations. However, because the shadow selection removes essential information from the video, an impact on the recognition system's performance will arise. Iosifidis et al. [27] removed clutter using depth data.

5.3 Occlusion

The activity performed should be visible in the video sequence. This is an issue since a considerable proportion of persons of people in the area of view of the camera. During video recording, self-obstructions or occlusions produced by other objects in the camera's

Table 1 Datasets summary for activity recognition

Dataset	Year	Type	No. of Video (Resolution)	No of Action	Applied-Domain
KTH [3]	2000	SV	600 (160 × 120)	6	Human Activity Recognition in Real Outdoor Condition
Weizmann- [4]	2006	SV	90 (180 × 144)	10	Human Activity Recognition
IXMAS- [9]	2006	MV	1650 (390 × 291)	13	Multiview Invariant Action Recognition
UCF Sports [5]	2008	SV	150 (720 × 480)	10	Sports Action Recognition
HMDB51 [6]	2011	SV	6766 (320 × 240)	51	Human Object Interaction, Human–Human Interaction
UCF50 [7]	2012	SV	6681 (320× 240)	50	Human Sports Activity Recognition
UCF 101 [8]	2012	SV	13,320 (320 × 240)	101	Human Sports Activity Recognition
IM-daily activity [16]	2014	MV	45 (640 × 480)	15	Daily Activity Recognition in real scenarios
UTD-MHAD [11]	2015	MV	861 (512 × 424)	27	Human Activity Recognition with a multimodal approach
Activity Net [12]	2015	SV	27,801 (1280 × 720)	203	Human Behaviour Understanding
NTU RGB-D [15]	2016	MV	56,880 (512 × 424)/(1920 × 1024)	60	Human Activity Recognition
NTU RGB D 120 [17]	2016	MV	114,000 (512 × 424)/(1920 × 1024)	120	Human Activity Recognition
Kinetics-400 [18]	2017	SV	300,000	400	Human Activity Recognition
AVA [19]	2018	SV	57,600	80	Daily Activity
Epic Kitchen [20]	2018	SV	432	149	Daily Activity in real scenarios
Kinetics-600 [21]	2018	SV	500,000	600	Human Activity Recognition
Kinetics-700 [22]	2019	SV	65,000	700	Human Activity Recognition
Moments in Time [23]	2020	SV	1,000,000	339	Event base Activity

SV Single View, *MV* Multi-View

field of focus are both conceivable. Because all body components executing the movement are not visible in the video sequence, this causes a significant dispute in the scientific community. Occluded portions significantly influence action recognition, whether self-occluded or owing to objects. Since the characteristics from occluded body parts are frequently lost, the system may decide that the action is incorrect, even though the system is right, because the characteristics from the occluded parts have a minor influence on the output. To cope with the change in camera view location, the most typical technique is to train the classifier with several camera views [28].

5.4 Viewpoint variation

Many methods presume that actions are carried out from a fixed vantage point. Depending on the camera angle, the position and posture of individuals change significantly. In addition, the movement patterns in each view may seem different, making recognizing the action necessary. Ji et al. [29] used a history picture and an energy image of motion to understand the underlying human movement in the images. The 3D body model shows the strongest consistency in all directions. The human body's 3D modeling exhibits the highest resiliency of all the techniques. According to one interpretation, all actions are characterized by features from the human body. Using the human body as a reference point makes identifying relative from absolute acts easier. However, owing to the many factors that must be analyzed, identifying correct 3-D poses from 2-D pictures or video frames is challenging, and viewpoint projection may result in confusing postures.

Machine Learning algorithms are not more capable of handling when data has some variation (unbalance data). A slow learning rate is another challenge compared to an ML-based approach when the large dataset has low recognition efficiency. Most of the work was conducted on the supervised algorithm in the ML base method.

The DL base method needs large-scale data, but we have limited data for the initial training phase in some scenarios. The automatic feature selection method requires more robustness.

6 Comparative analysis of prior surveys

Prior surveys in the domain of HAR in videos provide insight into the evaluation technique and cover new techniques with their accuracy. This study concentrated on zero-shot action learning (ZSL) in visuals, not still images-based AR. This study helps the researcher gain knowledge on feature extraction, semantic features of frames, datasets and action learning techniques. Estevam et al. [30] summarized only a few prior studies on ZSL and limited state-of-the-art methods discussed. The preceding survey expressed new research challenges and research gaps in the region of vision-based activity recognition. Pareek and Thakkar [31] discuss various handcrafted methods with conventional ML methods and describe the dataset. This study does not cover any pre-processing technique. Dang et al. [32] summarize both vision-based and sensor-based studies. This comprehensive survey describes the Convolution Neural Network (CNN) and Recurrent Neural Network (RNN) based methods. Beddiar et al. [33] review numerous studies on input data, ML method (supervised, unsupervised method), datasets, limitations and challenges of the HAR system. This study does not compare technique basis on their effectiveness. Zhang et al. [34] focused on both action detection and recognition. The action representation method was

also elaborated and extended. Herath et al. [35] summarized numerous handcrafted methods (Optical Flow, MHI and other motion descriptors) apart from these approaches, multiple streams of the network were compared in this review study. This section focused on prior survey studies. Based on studies, we choose attributes like pre-processing, dataset, feature extraction, and classification methods. This study compares prior literature in the HAR field. Most prior surveys did not cover all the aspects of HAR with comparative analysis.

Table 2 illustrates the prior surveys on HAR. Yes or no denotes the presence or absence of the associated attribute of HAR in the survey.

7 Handcrafted feature action recognition representation

Handcrafted feature-based approaches rely on manual feature extraction from the input data. Single-layer and hierarchical techniques for HAR can be categorized. The single-layer process identifies activity from acquired videos and does not deal with sub-activity. Conversely, the hierarchical system identifies activity using sub-events of the video sequence to represent high-level human activity. Figure 13 illustrates the Human Activity Recognition approach classification.

7.1 Single layer approaches

Human activity may be detected using a single-layered approach based on image ordering. Because of the character of single-layered processes, they are appropriate for identifying human emotions and movements with a sequential characteristic [41]. The space–time and sequential methods are two of the most common ways to classify them. Authors employ the temporal dimension differently in the sequential process than in the space–time approach. The space–time approach handles time as a spatial dimension and pull-out feature from 3-D volumetric videos. The sequential approach examines human activity while monitoring the timeline [39].

7.2 Space–time based approaches

Space–time techniques consist of four parts: an interest point, a description of features, a classifier, and a vocabulary builder. There are two kinds of detectors for space–time interest points: sparse and dense. V-FAST, Hessian detector, and dense sampling are all examples of dense detectors that cover the whole video material for the identification of interest locations. The Harris 3D cuboid detector (to track space–time features for activity recognition) and the space–time-implicit shape model employ a small subset of these contents for their sparse detectors. Local and global feature descriptors are the two varieties of feature descriptors. Local texture, color, and posture information are the basis for several descriptors, including the cuboid descriptor, improved speed-up robust features, and N-jets. Changes in illumination, phase shifts, and video speed are all examples of global descriptors. Vocabulary builders and aggregation techniques make use of the bag-of-words (feature extraction, feature pre-processing, codebook generation, encoding, pooling and normalization) or state-space model. There are three subcategories of space–time **(i) Space–time volume, (ii) Space–time trajectories (iii)**

Table 2 Overview of prior surveys in the area of human activity recognition

Author	Year	Description of Survey	Attributes						
			D	Pre	FE	C	Acc	ML	DL
Neha Gupta et al. [2]	2022	This survey enlists HAR devices, brief notes on datasets, strengths, and limitations, and describes various AI models (ML/DL) for activity recognition	Yes	Yes	Yes	Yes	No	Yes	Yes
Valter Estevam et al. [30]	2021	Discussed visual and semantic feature extraction for zero-shot action recognition in videos	Yes	No	No	Yes	Yes	No	Yes
Singh et al. [36]	2021	Discussed unimodal and multimodal briefly. Machine Learning and deep learning approaches with pros and cons	Yes	No	Yes	Yes	Yes	Yes	Yes
Pareek et al. [31]	2020	Discussed traditional machine learning and deep learning approaches for video activity recognition. Described the feature extraction technique & dimensionality reduction technique	Yes	No	Yes	Yes	Yes	Yes	Yes
L. Minh Dang et al. [32]	2020	Pros & Cons are associated with a method of vision-based Human Action Recognition. This survey does not discuss the handcrafted feature base method	Yes	Yes	Yes	Yes	Yes	No	Yes
Djamila Romaissa Beddiar et al. [33]	2020	Two machine learning techniques are defined here: standard and deep learning approaches. Did not explore the accuracy of each methodology	Yes	No	Yes	Yes	No	Yes	Yes
Zhang et al. [34]	2019	Described action feature representation techniques involving different data types and hand-drawn action feature representation Investigate deep learning methods for RGB, RGB-D, Skelton data and hand-drawn action features. Search deep learning methods only for the specific data type	Yes	No	No	Yes	Yes	No	Yes
Charmi Jobanputra et al. [37]	2019	This survey thoroughly examined several tools and approaches that may be employed in Human Activity Recognition, including several machine learning algorithms and neural network methodologies	No	No	Yes	No	Yes	Yes	Yes
Yu Kong et al. [38]	2018	This survey focused on feature representation and state-of-the-art video activity recognition methods Describe action classifiers and their effectiveness for activity recognition	Yes	No	Yes	Yes	No	Yes	Yes
Samita Herath et al. [35]	2017	Discussed the deep learning architecture for action recognition in videos. Define representation-based solution on video, which used handcrafted features in action recognition	No	No	Yes	Yes	No	No	Yes

Table 2 (continued)

Author	Year	Description of Survey	Attributes						
			D	Pre	FE	C	Acc	ML	DL
Guangchum et al. [39]	2015	Discussed hierarchical and non-hierarchical approaches for action recognition like space–time volume and space–time feature It does not cover pre-processing and feature extraction methodology used for action recognition	Yes	No	No	No	Yes	Yes	No
Shian Ru Ke et al. [1]	2013	Discussed various steps, which include activity recognition	No	No	Yes	Yes	No	Yes	Yes
Sarvesh Vishwakarma et al. [40]	2012	This survey entails the approaches which use object tracking. This survey is not incorporate feature extraction and accuracy of methods	Yes	Yes	No	No	No	Yes	No

Acronym for Table 2: *D* Dataset, *Pre* Pre-processing, *FE* Feature Extraction, *C* Classifier, *Acc* Accuracy, *ML* Machine Learning, *DL* Deep Learning.

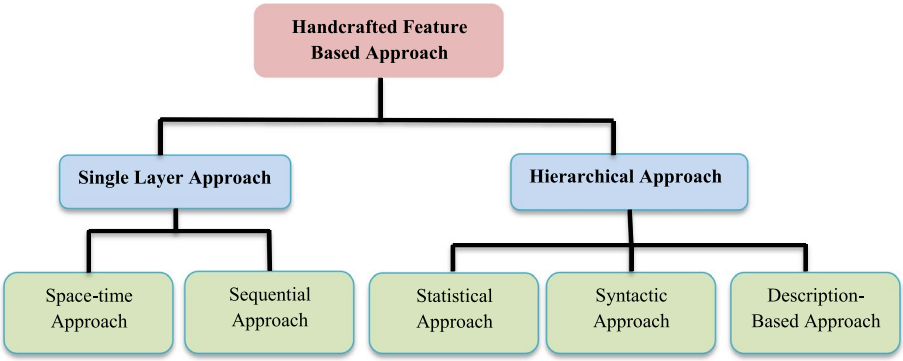


Fig. 13 Human Action Recognition Hand-crafted Feature-Based: Taxonomy

Space–time feature. For a reason, they pull sequential relationships into a review; sequential techniques produce better results than their space–time counterpart [42].

Space–time volumes (STVs) are 3D Spatiotemporal cuboids representing features in the space–time domain. At the core of STV-based techniques for activity identification is an estimate of proximity between two volumes. Rather than relying on space–time volumes, Bobick and Davis suggested a template-based action recognition system. They described activities using two-dimensional binary Motion-Energy-Image (MEI) and MHI templates, then used a basic template matching approach to recognize them [42]. We would use a template or feature as the 3D volume in the space–time volume. This method suffers from redundant information and noise. Qian et al. [43] combined local and global features for recognizing human activity. Binary motion energy image (MEI) was used as a global feature. Still, they used contour coding on motion energy images to overcome the limitation of motion energy images because the human blob is undetected by motion energy images. Object bounded box used as a local feature. The position of a person's joints changes as they engage in a specific activity, depending on the movement.

The **space–time trajectories** that track joint locations in the body must be able to distinguish among various forms of activity. Campbell and Bobick released [38] an article in 1995 in which they trace the trajectories of joint locations to determine nine atomic motions of a ballet dancer. In a three-dimensional XYT plane. Combining 13 joint trajectories in 4 Dimensional XYZT plane is used in the matching technique, which Sheik et al. proposed [43]. Despite variations in perspective, anthropometry, and execution pace, this approach has detected basic activities such as sitting, standing and dancing. The Kinect sensor captures depth videos, which may be used to recognize human actions [44]. This method displays the human body's dynamic skeletal form on Kendall's shape manifold as trajectories. The usual Euclidean norm is used in trajectory trajectories transported-square root vector fields to maintain computing efficiency. Such approaches based on space–time characteristics for Human Activity Recognition, space–time volumes, or space–time trajectories can be used to extract attributes. Often, these traits are local in nature and comprise action-specific traits. Depending on the volume and trajectory of space–time, these characteristics may be segregated into two classes: sparse and dense. Wang et al. [45] employed dense trajectories to get these results. Wang sampled and tracked dense regions for each frame using the optical flow

field's displacement information. Histogram of Oriented Gradient (HOG), Histogram of Optical Flow, and MBH (Motion-Boundary histogram) local descriptors were constructed in the vicinity of interest locations.

Local aspects of a 3D volume on a space–time scale are referred to as **Space–time features**. Recognizing human activity using the space–time features method provides a decent approximation because this approach assumes a 3-dimensional space–time volume as a 3-dimensional rigid object; it executes an object matching technique to represent each action in a 3-dimensional volume. These features solved the object-matching issue.

To localise the compact representation of the event, a generalization of Harris and Forstner's interest point detector was presented by Laptev et al. [46]. To circumvent the constraints of the Dollar detector [47], clouds of space–time interest spots were established by Bregonzio et al. [48]. Features are selected automatically after accumulating cloud interest points across various time ranges. Thi et al. [49] used a Harris detector [46] for space–time interest points by utilizing a Bayesian classifier; labels of [-1; 1] were assigned, indicating whether it belonged to the group of interest action. The author used a PCA-SVM classifier to recognize the action. Utilizing Harris corner detectors in each frame, Sadek et al. [50] generated local features with fuzzy log-polar using temporal self-similarities. SVM is used for the classification of actions based on global features. Figure 14 represents space–time interest point detection on different frames.

Another basic approach for feature point detection and descriptor use is optical flow. More complicated feature descriptors, including SIFT descriptors or HOG descriptors, which are histograms, are used in recognition approaches. We can no longer explain histograms' temporal development using linear dynamical systems, and we can't categorize them using a metric for linear dynamical systems since they dwell in a non-Euclidean space. Chaudhary et al. [51] suggested a novel method based on the Histogram of Oriented Optical Flow (HOOF) and achieved 94.4% accuracy on the Weizmann dataset. For the collection of precise geometrical interest points, R-transform is presented by Yuan et al. [52]. Fusion of histograms is employed in action recognition by Sahoo et al. [53] and is known as the "bag of optical flow". After generating features, a Sobel edge filter was applied [54], followed by median filtering [55] to

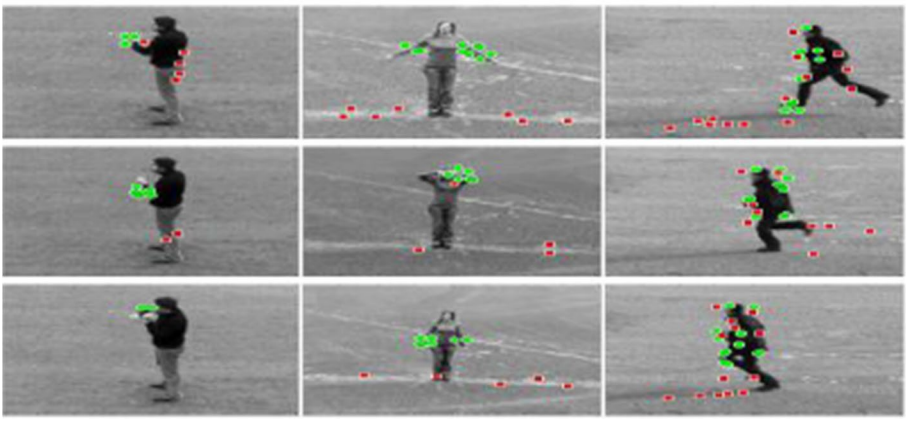


Fig. 14 Detection of interest point in frames using Dollar et al. [47]

attenuate background noise. On the KTH dataset, a bag of the Histogram of Optical Flow (HOF) was blended with extracted HOG and then sent to the SVM Classifier, which obtained 96.70 percent accuracy.

Le et al. [56] developed a hierarchical adaptation of the independent subspace analysis (ISA) technique to grasp invariant spatiotemporal properties from unlabelled visual data. For example, small data patches were combined to form a vector that was then utilized as an input to the layer above, which in turn learned features. Both layer characteristics were combined into a single set of local features for categorization. The limitation of independent subspace analysis for lengthy inputs was overcome by two-layer stacked convolution independent subspace analysis and did well on typical datasets.

7.3 Sequential approaches

Sequentially oriented techniques vary from space-oriented approaches because they capture the temporal link between observations. The sequential approach uses feature analysis to recognize human activity in video. These approaches construct the feature vector with the help of feature extraction from the sequence of images, which explains the person status per image frame. Sequential methods evaluate the sequence after feature vectors have been retrieved to determine how probable the feature vectors were created by the individual doing the activity. Activity has occurred if the likelihood between activity class and sequence is high enough. Two sub-categories of sequential methods are exemplar-based recognition and state model-based analysis. Using a template sequence observation or a sample sequence, exemplary approaches explain human behavior.

Exemplar-based techniques concentrate on determining whether the latest visual stream may be compared to a template or sample action observation sequence**.** In the exemplar-based approach, a template sequence was used with a compared feature vector extracted directly from the sequence of videos. Dynamic Time Warping (DTW) was presented by Darrell et al. [57] to effectively identify the gesture and movement in the execution of an action. The most commonly used technique in exemplar-based human activity recognition was DTW. When Jiang defined [58] an action as a succession of postures, they used a geometrical model of a local human component, which they subsequently identified using a sequence-matching technique.

State model techniques come up with a state model for each action, and each activity is depicted by a collection of hidden states, which are created using state model methods. For exemplar-based human action recognition, dynamic temporal warping has been frequently used. The state-based model produces observational sequences connected with the appropriate activity occurrence. For state-based techniques, standard hidden Markov models have often been employed. The Hidden Markov model could not recognize the complex activity, like when more than two people interact due to large parameters. Oliver et al. [59] suggested some variation on the hidden Markov model for recognizing complex activities, known as the Couple Hidden Markov Model (CHMM). The results of CHMM are better compared to HMM from a learning point of view. Discriminative semi-Markov models and a Viterbi-like dynamic programming methodology were devised by Shi et al. [60] to effectively handle the inference problem of simultaneously detecting distinct actions.

7.4 Hierarchical approaches

A hierarchical approach to human activity identification aims to identify high-level activities based on lower-level activities that are part of a more extensive set of actions (sub-activities). Sub-activities can be adjusted comparatively simply to recognize the earliest and following; authors use them to acknowledge another activity, like higher-level activities. Higher-level activity recognition builds with sub-activities like dancing containing hand and leg movements. Aside from that, hierarchical models provide an easy-to-use interface for integrating past data and grasping how actions are organized. There are several benefits to using hierarchical approaches, such as the ability to mimic various human activities, interactions between people and human objects, and the ability to simulate collective actions. Taxonomy allows us to separate the hierarchical method into three distinct areas. Statistical, syntactic, and description-based methods are all viable options.

7.5 Statistical approaches

Hidden-Markov Models (HMMs) and Deep Belief Networks (DBNs) are employed in statistical approaches to identify sequential actions. A state-based model is the sequence of operations in response to external stimuli. The Layered Hidden Markov model was suggested by Oliver [61] as a series of Hidden Markov models (extension of HMM). In LHMM, both layers (upper HMM or bottom HMM) are connected with inferential results because of the multiple layers of state-based models used to recognize action in the statistical model. Various researchers use LHMM to identify complex activity. Zhang [62] recognized group action and [63] proposed behavior recognition, but did not perform well on variation in behavior activity. Shi presented a propagation network to handle concurrent and sequential activities efficiently [64]. Simultaneous and sequential phases of sub-activities and fuzzy vector quantization were introduced by Iosi et al. [65] to identify and recognize persons from different views. Xu et al. [66] described a hierarchical space–time model (HSTM) to resolve the issue by customizing spatial and temporal constraints altogether. Hidden conditional random field (HCRF) has two layers where the lower-layer HCRF speaks about spatial links in one and all frames while learning new discriminative characterizations, and the upper-layer HCRF leverages these high-level characteristics to describe temporal relationships throughout the video series. As a result of the use of various data preparation techniques, such as Random Forest, Support Vector Machine, Naive Bayes, and Multilayer Perceptron (MP), Human Activity Recognition (HAR) may be enhanced.

7.6 Syntactic approaches

Human activity is accompanied by a collection of rules for string production, generating atomic actions in syntactic approaches. Obtain parsing methods from the area of computer languages used for recognition. Context-free grammars (CFGs) and stochastic context-free grammars have previously been used to identify high-level activities (SCFGs). Kitani et al. [67] offered a strategy default grasp grammatical rules from observation. It leads to a restriction for parallel action recognition even though actions are expressed as a string of symbols. Context-free grammar (CFG) has been investigated and suggested based on syntactic techniques in Human Action Recognition. Ivanov et al. [68], Moore et al. [69],

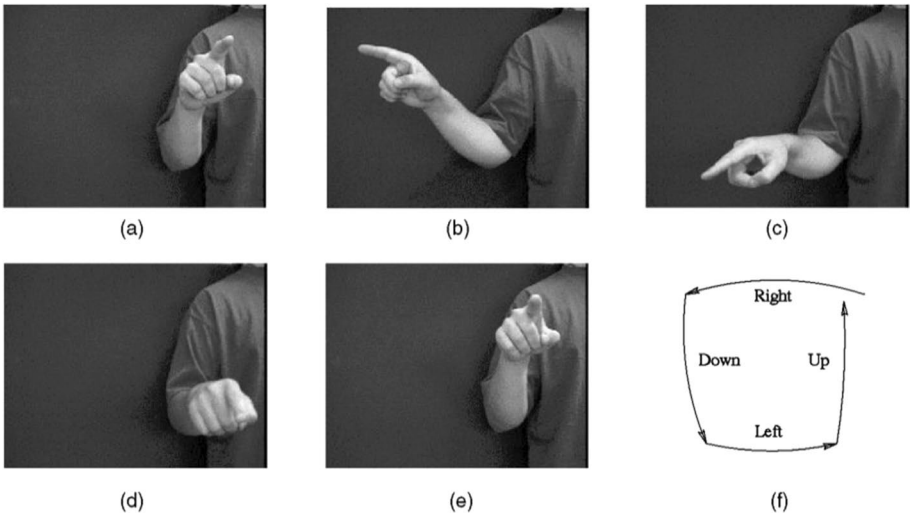


Fig. 15 Structure of activity (**a**)-(**e**) Square showing video sequences, (**f**) Phases of motion are shown [68]

Minnen et al. [70], and Joo et al. [71] presented various stochastic context-free grammars, which are probabilistic extensions of CFGs (SCFGs). For action recognition, Joo uses attribute grammar, an extension of SCFG. Figure 15 expresses the structure of the activity.

7.7 Description based approach

A description-based approach is a recognition process that explicitly records spatiotemporal frames for industrial activities with other data. Description-based methods have the capability to express human activity spatiotemporal structure and are different from syntactic and statistical techniques. The description-based system recognizes sequential and concurrent action. Instead of being confined to sequential activities, description-based approaches may identify sequential and accompanying actions. Siskind et al. [72] suggested a method to recognize simple actions with the help of primitive events acquired through high-level activities. In order to detect sophisticated semantic-level complex events, event logic is utilized to mix the earliest occurrence of involved individuals or objects. Several probabilistic extensions of recognition frameworks are described in [73] to compensate for the low-level components' failures due to the deterministic features of description-based approaches. Ijsselmuiden and Stiefelhagen [74] recommended a framework for action recognition of human activity, which is high-level activities. It was based on temporal logic and employed a variety of input sources. They presented cutting-edge real-time computer vision systems for tracking and identifying people and estimating their visual center of attention and gesture activity.

7.8 Discussion

When numerous persons appear in a scenario, the volume representation is limited to recognizing activities. They are computationally costly to locate activities since they lack view invariancy. A space–time trajectory can better express and identify human movement

from various perspectives. To simulate more complicated, nonperiodic, and viewing-invariant representations of activities, one cannot rely on the spacetime features method, which is robust against noise and illumination fluctuations. By using sequential characteristics, sequential techniques can detect complicated and non-periodic behaviors. While state-based strategies need more training data, exemplar-based approaches are more flexible. A drawback of state-based methods is that they need an enormous amount of training data as the complexity of an activity increases.

In contrast to statistical or grammatical techniques, description-based approaches are better at identifying high-level activities with contemporaneous and sequentially occurring subevents. A noisy activity video may be cleaned up using statistical and syntactic methods. The objective of the activity depiction challenge is to identify characteristics that are insensitive to occlusion, backdrop variation, and perspective modification.

Optical flow-based techniques regularly enhance those based on optical flow like HOF, MBH, dense trajectories, MHI, etc. Each feature-based approach has their limitation as well. For example, space–time value does not perform in view-invariant and is only effective when numerous individuals participate in a single action. Space–time trajectory serves better when video frame points are notable and can fit various angle views in this condition. The Space–time features identify collective actions except for impenetrable activities with steady view representation. Appearance-based techniques rely on the descriptor of motion and action part region for sequence matching and these are generated with the help of shape and motion-allied information. This approach uses silhouette and interest point detectors. After that, this information is fed into the classifier for action identification. Sequential base methods can manage complex activity and dynamic view information. Layered approaches easily perform high-level activity recognition. Exemplar-based approaches are amenable to identifying complex activity as compared to state-based approaches. To handle noisy data, syntactic and statistical approaches are capable. Authors in [75] propose multiresolution feature extraction using Daubechies complex wavelet transform. The extracted features are fed into a multi-class SVM classifier for action recognition. The depth motion image was constructed from the depth data to generate an action multiscale static representation [76]. A multi-granularity descriptor is employed to bring down static information from frames.

Table 3 provides a comparative analysis of various handcrafted-placed methods for AR in video sequences with their limitation.

8 Deep learning-based classification model

A subset of machine learning methods known as Deep Learning is one that uses numerous levels of hierarchical knowledge stages of processing to identify patterns or images. Many improved DL models have been introduced in the latest years, each focusing on a specific issue and applying an appropriate dataset to solve it. Deep learning models are becoming more popular due to their performance. In DL models, it is possible to create high-level traits from low-level ones [90]. Deep Learning produces superior outcomes in identifying human activities compared to machine learning methodologies. Deep Learning techniques are subdivided into generative, discriminative, and hybrid. Generative deep-learning approaches are the most common. In these models, independent or dependent distributions in sensor data are described by graphical models, with graph nodes representing random variables and arcs demonstrating the link between variables in the sensor data that

Table 3 Hand-crafted representations using various datasets for recognizing Human Activity

Author	Approach	Dataset	Results in %(Accuracy)	Observations	Limitation
Nazli Ikizler et al. [77]	Space–time volumes	KTH	89.4	By extracting rectangular patches in an organized pattern over the human silhouette, authors depict every human pose in an activity sequence and identify human motion using Histogram-of-Oriented-Rectangles (HOR)	Dependent on Silhouette extraction and only evaluated on two datasets KTH, Weizmann
Heng Wang et al. [45]	Dense Trajectories	KTH	95	The author suggests using dense trajectories to describe images, sampling dense points from each sequence and trailing them using dense optical flow field displacement data	Camera motion is the main hurdle for dense trajectories. They cannot extract the target feature for recognizing action if the action reorders are not steady
Vili kellokumpu et al. [78]	LBP (Dynamic Texture-HMM)	Weizmann KTH	100 93.8	The author proposes temporal templates to record movement dynamics and then uses textural attributes to describe the motions seen, using image data and silhouettes	This method shows results on KTH, Weizmann, and Online action recognition, but it works efficiently on image data but only performs better when the silhouette is used. Compared with the state-of-the-art approach, the KTH dataset only acquires a correctness of 93.8%
Kliper Gross et al. [79]	Bag of visual words	HMDB-51	29.2	The author suggested employing a suppression mechanism to separate picture edges from motion edges and compensating for overall camera motion by combining a custom-designed registration approach with a common bag-of-words technique	Further improvement with the addition of a learning-based hierarchical layer

Table 3 (continued)

Author	Approach	Dataset	Results in %(Accuracy)	Observations	Limitation
Yu-Gang Jiang et al. [80]	Trajectory	Hollywood2 Olympicsports HMDB-51	59.5 80.6 40.7	The foreground–background distinction generally required to describe object interactions is not necessary for this technique since it acts on top of visual codewords formed from local patch trajectories	Local and global patch information is used for motion, but background subtraction is unnecessary. This method is not appropriate for those applications where background information is needed
C Wang et al. [81]	Pose-Based	UCF-Sports MSR-Action-3D	90 90.22	To get the poses and use data mining approaches to extract spatial–temporal pose structures for action representation, they expand the pose estimation algorithm in this manner	The effectiveness of these methods relies on the estimation of the pose. If occlusion happens, then the result of recognition will be affected
Mihai Zanfir et al. [82]	Pose-Based	MSR-Action-3D	91.7	This technique employed a pose descriptor framework for action recognition that leverages posture and kinematic data stored as differential 3D values	They are focused only on human action detection within a short time frame. The extensive sequence of the video is complex to detect
A.A Chaaraoui et al. [83]	Shape-Based	Weizmann	92.8	The described approach for recognizing human actions uses sequences of multi-view key poses to teach activities, with pose representation based on the contour points of the human silhouette. This approach yields action recognition of Multiview poses. Primary pose extraction based on view with the help of silhouette	Hard to deal with a cross-view milieu using this proposed approach

Table 3 (continued)

Author	Approach	Dataset	Results in %(Accuracy)	Observations	Limitation
S. A. Rahman et al. [84]	Shape-Motion	Weizmann KTH	95.56 94.49	They used areas around the subject, known as negative spaces, and suggested an implicit way of identifying actions in videos. Real-time action recognition improved with low-resolution images. Required less processing time. Pacts with partial occlusion, tiny shadows, and distorted picture sequences	This approach requires more processing power to deal with variations in poses
Imran N. Junejo et al. [85]	Shape-Based	IXMAS	89	Time series are created using the actor silhouettes retrieved from each frame, and the symbolic vector SAX is then effectively transformed into each of these time series	This approach evaluates the performance on a single Multiview video dataset recorded in a controlled environment. This approach is not suited for real-time recorded data
B.B Amor et al. [44]	Trajectory Based	MSR-Action-3D	89	The form evolution of skeletons has been used to provide a comprehensive framework for studying human behavior	Extensive data require in a skeletal form to identify the action
D.K Vishwakarma et al. [86]	Shape-Based	Weizmann	100	This research aims to offer a unique integrated substructure for detecting personal activities utilizing a precise geometric orientation of a human silhouette and a spatial distribution of edge gradient (SDEG) of a human stance in a video series	Hard to handle feature vector dimension to improve the correctness of the introduced method

Table 3 (continued)

Author	Approach	Dataset	Results in %(Accuracy)	Observations	Limitation
Ahmad Jalal et al. [87]	Spatio-Temporal Based	MSR Action-3D	93.3	The online Human Activity Recognition (HAR) method uses many fused characteristics to identify human activity from continuous depth map sequences	Computational complexity handling issues arise in the fusion of multi-modality
		MSR Daily	94.1		
F. Patrona et al. [88]	Joint Based	MSR Action -3D	96.2	Presenting real-time action detection, identification, and interpretation of motion capture data	Large data volume occurs due to recognizing action in real-time. The training phase of the approach needs a lot of computation power
C. Zhang et al. [89]	Hybrid Handcrafted Model	UCF-11	95.40	This work combines DWT into the DT model to describe human activity. Investigating pre-trained dual-stream CNN-RNN models may combine learned and hadcrafted based features to meet spatial–temporal analytical needs	Relie on trimmed video classification. The proposed approach cannot process variable lengths of videos
		UCF-50	95.68		
		UCF-101	95.1		
		HMDB-51	70.8		
		JHMDB-51	72.5		

has been provided to the modeler. In discriminative models, the distribution of label data may be used to discriminate between classification and modeling. In the hybrid model, we combine both generative and discriminative models for Human Activity Recognition. Various deep neural networks like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM(BLSTM), etc., are used to detect Human Activity.

8.1 Convolution neural network

As a member of the family of biologically inspired visual recognition models, CNNs have spawned several other variations. Nonlinear, high-dimensional features may be automatically extracted and fully learned by a CNN. Ji et al. [91] proposed "3D convolution for capturing spatial and temporal features in video analysis". A CNN was used to identify the specific activities in each footage frame. The KTH dataset gave a 90.2 percent accuracy rate for the 3D CNN technique. The most common procedures are convolution, activation of non-linear neurons, and feature pooling. Figure 16 [92]. describes the process of deep convolution neural networks. A CNN is referred to as "deep" if it has many feature extraction phases. This network keeps an interconnected structure and operates on raw data. One of the most common uses of CNN is in voice recognition, activity detection, phrase modeling and other areas. Each layer has different roles. Multiple kernel sizes and strides are used to acquire feature maps in the convolutional layer. As an outcome, the convolutional layer and pooling layer have fewer connections. The primary work of the pooling layer is to minimize the counting of feature mappings and parameters and to strengthen the network's resistance to distortion and translation [93].

As a result of the 3D CNN network's motion stream fusion, 3D posture, 2D appearance and motion stream fusion may all be acquired ". A 15-channel heatmap is constructed since it will be challenging to recover joint colour features in the 3D CNN; thus, convolution is undertaken on each map [94]. Lots of data is required to train the DL method. However, collecting & analysing large amounts of video content takes time and a lot of computer resources, it is well worth it. Images, item identification, speech recognition, and human action recognition may all be classified using traditional 2D and 3D CNN models [95].

The extremely used network of deep learning is CNNAs part of their work on the ImageNet LSVRC-2010 challenge, Krizhevsky et al. [96] trained a huge, deep convolutional neural network to sort the 1.2 million high-resolution photos into 1000 different categories. Mo et al. [98] used Microsoft Kinect data to develop a deep-learning approach for tracking human movement. The multilayer perceptron is utilized for classification, whereas

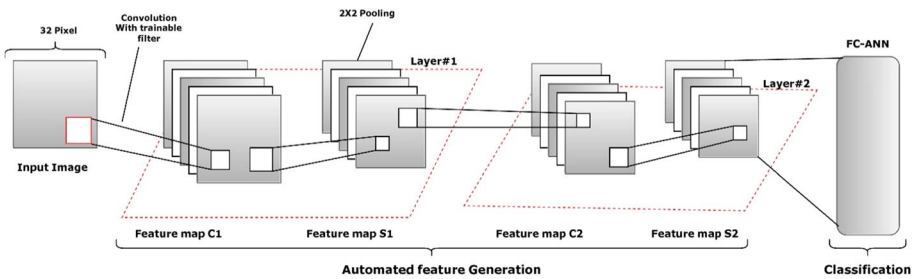


Fig. 16 Convolution deep neural network [92]

ConvNets are used for feature extraction directly. The accuracy rate of this model is 81.8%, and it recognizes twelve activities. Extensive CNN-based techniques on large-scale datasets were given by Karpathy et al. [97]. In this study, various models for combining information across time were tested. Because of this, a model that uses gradual fusion instead of early or late fusion has been shown to perform better. The use of low-power, wearable devices and temporal convolutions to distinguish action classes was employed in [98]. The CNN-based technique provided in [99] predicted future activity and detected the form of the image's prominent regions. Choquet integrals were applied to the decision scores created by utilizing CNN-based classification models for the four feature representations encoded into one greyscale to get the final choice-based confidence score [100]. It's been shown that combining CNN with custom features or representations may boost its performance.

8.2 Autoencoder

Autoencoder uses encoder and decoder units to build the most discriminative features from unlabeled sensor input. To reduce the calculation time, the training autoencoder's hidden units have a lower dimensional bias for activity detection than the inputs or outputs. The autoencoder method uses multiple encoder layers to convert high-dimensional data into low-dimensional feature vectors. Various autoencoders are introduced to enhance the robustness of feature representation, like contractive, sparse [101] and denoising autoencoders [102]. Originally, autoencoders were suggested for dimensionality reduction. Lower-dimensional activity detection may be achieved with less computation time using a training autoencoder with smaller hidden units than the inputs or outputs. It is possible to learn a more powerful generalization using nonlinear encoders and decoders than linear encoders, but PCA can only retain a limited subspace.

Hasan and Roy-Chowdhury [103] developed a system that uses stacked SAEs and active learning to identify streaming visual actions. Wang et al. [104] demonstrated a stack AE for recognizing egocentric activity in films. Light GBM-based Stacked Denoising Auto-Encoders were used in a human action recognition method presented in [105]. It is often used in HAR, AE, and their derivatives to bring down data dimensionality and learn features from the data. They may also employ unlabelled data in addition to their exceptional performance. In Fig. 17, an algorithm is taught to minimize errors in reconstructing the original input from the characteristics. It reduces the error between input and output, which help to comprehend the attribute available in data.

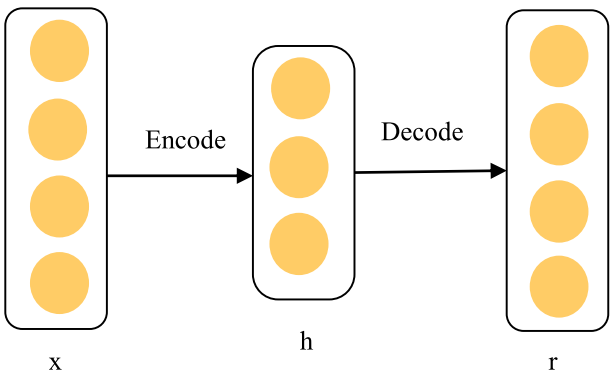


Fig. 17 Autoencoder architecture

8.3 Recurrent neural network

RNN records sequential input using a temporal layer, then utilizes the current cell's hidden unit to understand complex changes. Figure 18 illustrates the graphical representation of the RNN model. The invisible unit cells can alter depending on the network's accessible information, continually updated to represent the network's present state. With RNN, it is possible to determine the current hidden state by computing the future hidden state as a function of the activation of the prior hidden state. [106] suggested a hierarchical RNN model for action recognition, which they used in their solution. In addition, five alternatives deep RNN designs built from previous research have been compared to the proposed network. The MSR Action3D dataset, the Berkeley MHAD dataset, and the HDM05 dataset were all used to evaluate their performance. RNNs are better at solving sequential issues because they have recurrent connections than ordinary neural networks. It's very uncommon for RNN training methods to restrict the amount of recurrent memory stored in the network because of difficulties in disappearing or ballooning gradients. In [107] LSTM RNNs are a prominent method for reducing the impacts of gradient vanishing and explosion. The LSTM, RNN differs from other deep neural networks in that it does not employ hidden units but memory cells with inputs and outputs controlled by gates to store or forget information [108].

To recognize the group and individual actions, stagNet, a semantic RNN, was presented in [109]. The researchers employed structural RNNs to expand their semantic network model, including the temporal dimension between video frames. Using this approach, 90.5 percent of the volleyball dataset was obtained through group activity and 83.5 percent through individual action. To expand LSTM-based learning across concurrent spatial and temporal domains, Liu et al. [110] built an LSTM-based network. Their Spatio-temporal LSTM with Trust Gates achieved 93.3 and 97.0% accuracy without any prior smoothing procedures on the SBU Interaction dataset and the UT Kinect dataset, respectively. When it came to the Berkeley MHAD data set, they attained perfect accuracy.

In contrast, although the traditional LSTM has shown promise in various applications, it may not be easy to understand input structures that are more complex than a series of numbers. To overcome this issue, s-LSTM is a tree-structured LSTM network presented in [113]. The S-LSTM network has input and two forget gates, a cell gate, and an output gate.

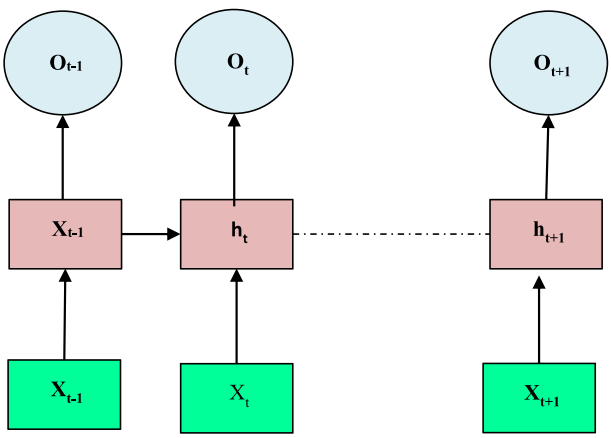


Fig. 18 Recurrent neural network

Gated Recurrent Units are another essential variation of the basic RNN that may adaptively capture sequential dependencies (GRUs) [111]. Most LSTM and RNN-based approaches consider low-level features, and the raw skeleton coordinates are used to train their models.

$$h_t = \sigma(W_h h_{t-1} + W_x x_{t)} \tag{1}$$

$$O_t = \sigma(W_s h_t) \tag{2}$$

where σ stands for a Sigmoid activation function, W_h is the weight matrix between the previous hidden state and the hidden layer,W_x is the weight matrix between the current input and the hidden layer,b_t stands for the bias at time t, h_t stands for the hidden state at time t, x_t stands for the input of the network at time t, and W_s is the weight matrix between the hidden layer and the output layer.

8.4 Generative adversarial network

In a GAN, the generator and discriminator are the two main components. The generator's primary responsibility is to get familiar with the distribution of data so that it can create samples that will fool the discriminator. The discriminator evaluates models to distinguish between actual data (the generator's training data) and bogus data (those generated by the discriminator). However, the original GAN [112] suffers from many issues, including gradient vanishing and low variety in its training.

A self-supervised task was introduced to a GAN to increase its ability to encode visual representations, which is crucial in downstream tasks like human activity detection and identification of moving objects. To generate a model allocation that closely mimics a preset distribution, the generative adversarial network [113] is used. For constructing cganthe model distribution, it has a generator G (z; g) and a discriminator (x;) for identifying it. In this case, Z represents the input noise and x represents the training data, with a Pz (z) probability distribution. The generator and discriminator are implemented in CNNs with the same parameters, g and d. D is programmed to optimize the likelihood of correctly labeling training examples and G data. G is also being taught to reduce the log (1 D(G(z)). However, implementing GAN is difficult due to training instability, mode collapse, and other factors. One of the main difficulties with GANs is discriminator remembering, which limits their capacity to offer efficient results [114]. Zaki et al. [115] proposed that GAN be given a self-supervised task to improve its effectiveness in generating meaningful video representations. In this method, the self-supervised job is to pick a transformation at random and apply it to video frames discriminator is then pushed to predict the right changes that have been implemented.

The findings imply that including additional transformations into different systems increases baseline GAN's performance and outperforms rotation-only self-supervised GAN when providing a film representation suitable for human activity detection, as demonstrated by the results. Li et al. [116] define a conditional generative adversarial network to identify activities inside each source image sequence by first creating an activity mask (cGAN). Color channels of the input pictures are then processed using a VGG-LSTM network to build a mask that may be used for activity identification. Yang et al. [117] utilize Open GAN to recognize activities based on an open set. The open-set problem in the train and test sets entails assembling a dataset with various categories. The open GAN extracts and combines characteristics using dense blocks, which are joined with these layers.

Sub-blocks link two convolution layers with a concatenation layer to create thick blocks. With a dense block, Fig. 19. demonstrates how convolutional and deconvolutional layers are applied. Convolutional and pooling layers are employed in the Dense Net for GAN's discriminator network. The final layer uses the SoftMax classifier and Mean Squared Error loss algorithm.

GANs and their variations have yet to be actively employed in HAR, and they might be extensively employed in HAR because of the difficulty and difficulty of getting labeled data. GANs may drastically minimize the time and effort required to acquire labeled data. They could identify group activities from video sequences using LSTMs and a multi-level GAN architecture [118]. Based on [119], the DCGAN model is trained on a large unlabeled video activity dataset before fine-tuning using the GAN model's labeled dataset to identify specific activities in the video. When labeled data is insufficient or non-existent, GANs, in general, may be used to supplement the information. Figure 20 shows the architecture of the Open GAN Discriminator and how convolution, pooling and dense blocks are placed.

8.5 Hybrid deep learning models

A hybrid model is a mixture of a few deep models. All recent work incorporating machine vision and human action detection has employed the hybrid deep learning approach. The majority of the precision comes from exposing machine vision information due to extracting important features, such as picture segmentation. Hybrid networks may also be used to collect temporal information by layering a recurrent layer on top of CNNs, such as LSTMs [120]. CNN's and LSTMs may be combined to create hybrid networks that can capture spatial motion patterns, temporal orderings, and long-range relationships [121]. In a video sequence, Jaouedi et al. [122] created a unique technique to employ motion tracking via human monitoring and spatial feature extraction. Human motion tracking techniques such as GMM and K.F. are used to display the motion characteristics. Additional characteristics are calculated using only a Recurrent Neural Networks (RNN) model with Gated Recurrent Units (GRU) and are dependent on the visual qualities within each frame in a series of videos, as shown in Fig. 21. The evaluation & extraction of all information within every interval and frame of vision are the significant benefits of this unique technology. UCF Sports,

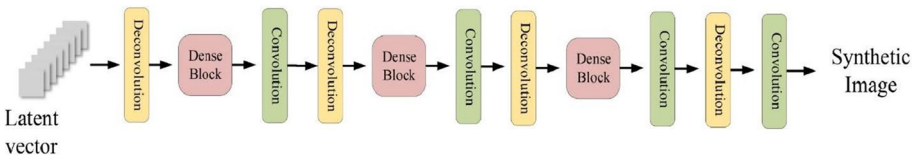


Fig. 19 Open GAN generator [117]

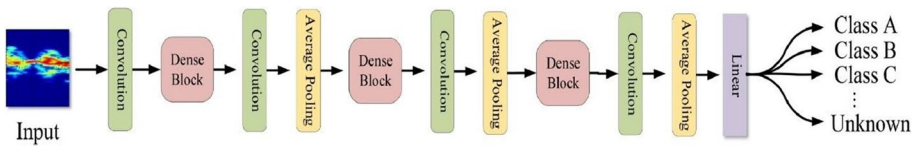


Fig. 20 Open GAN discriminator [117]

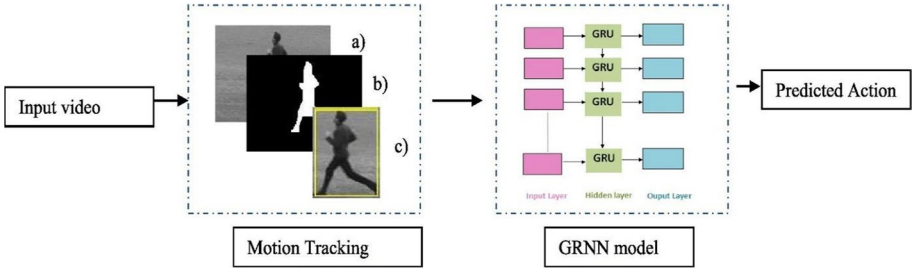


Fig. 21 Human action prediction model using gated recurrent neural network [122]

KTH, and University of Central Florida datasets were used to test their strategy, and they achieved 89.01 percent, 89.30 percent and 96.30 percent, respectively.

DL approaches are used in a hybrid model in [123] to extract features and identify interesting characteristics. In [124], After extracting motion data from video frames, they put them into a Bi-Directional LSTM model to simulate long-term temporal relationships. The relationships between appearance and motion characteristics were captured via a regularised fusion approach. This hybrid method was essential for precisely recognizing human behaviours. A hierarchical CNN and RNN hybrid deep neural network may also take advantage of intra-sensor, inter-sensor, and temporal connections in multimodal sensory data. Experiments have shown that the suggested technique is more successful than existing techniques [125]. Deep neural network models that combine the outputs of a CNN classifier with a binary decoder were proposed by Ijjina et al. [126]. In order to construct an effective classifier fusion model, it is necessary to use the high confidence that classifiers have in making correct predictions and the diversity of models (which may be obtained by adjusting input characteristics and model initialization). This classifier model is beneficial because fusion with maximum value ensures that the proper class label is picked in Fig. 22. Azzawi et al. [127] propose a novel model for classifying the activities of humans using

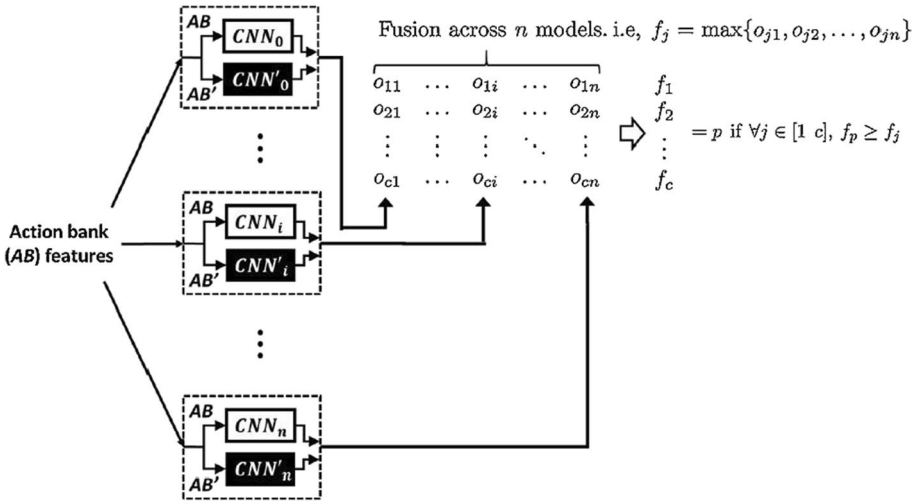


Fig. 22 A hybrid CNN model that can recognize human actions in [126]

video sequences. Segmentation is the most excellent feature for detecting human activities. Feature extraction may have an impact on the calculation's complexity as well as the method's efficiency. Santosh et al. [128] used the skeletal posture of the body to recognize action and fall detection. Deep CNN and LSTM are applied successively. Skelton parameters are precalculated from images and visual information, which use pose estimation or body detection from data. Some methods have focused on developing novel transformer-based architectures for human activity recognition such as [129].

In [130] author used hybrid convolutional network and BI-LSTM to process raw data for feature extraction with kernel of different size to capture local temporal dependency. The author [131] combined 3D human recovery and spatiotemporal bifurcations. It is composed of 3D spatial and 2D temporal branches. The 3D spatial extend uses contiguous human representations to 3D recovery to discover 3D recognition characteristics. The author [132] used Spatio-temporal relations to identify interaction and activity. Scene graphs are used to extract key features from image sequences. Skelton joint tracking is employed to recognize human activity [133]. A deep recurrent neural network is used to evaluate performance [134]. RGB D stream features are extracted to represent action with complex network architecture. This architecture [135] is used to learn single-person action from skeleton information of body joints.

8.6 Performance evaluation matrices for model

Researchers have developed different criteria to measure the effectiveness of HAR models, with accuracy, Precision, Recall, and F1-score being the most common.

i. Accuracy: It defines the model working with the help of correct prediction compared to the total number of instances. The accuracy of the categorization is what this metric performance measures and its formulated as:

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{3}$$

ii. Precision: Precision represents the probability of a monitored instance of activity occurring in the actual World. Conversely, (1—precision) is the likelihood of the recognizer wrongly detecting an observed action. Measure the correct classification and its formulated as:

$$\text{Precision} = \frac{T_P}{T_{P+}F_P} \tag{4}$$

iii. Recall: The approach range to identify positive trials is estimated with the help of the recall metric and its formulated as follows:

$$\text{Recall} = \frac{T_P}{T_{P+}F_N} \tag{5}$$

iv. F-Measure: It provides information regarding the correctness of the evaluation. As a result, the F-measure influences both the precision and robustness of the classifier. Use 1 & 0 to represent model classification evaluation greatest or lowest. Harmonic Mean of Precision and Recall and we can write as.

$$F - \text{Measure} = \frac{2 * (\text{Pr} * \text{Re})}{\text{Pr} + \text{Re}} \tag{6}$$

where Pr: Precision, Re: Recall

T_P = True-Positive estimation (Count of positive samples that were actually predicted).

F_P = False-positive estimation (Count of incorrectly anticipated as a positive class).

T_N = True negative estimation (Accurately estimated count of negative sample from the data).

F_N = False-negative estimation (Count of incorrectly identified as a negative class).

v. Confusion Matrix: The confusion matrix represents the prediction made by the model and the model's performance. It also shows the incorrectly classified action and what is the actual action. The confusion matrix has precise class and predicted class.

8.7 Discussion

The deep learning approach is a sub-type of ML that focuses on interpreting input such as sounds, images, and text by learning distinct levels of representation and abstraction. DL strategies can evaluate raw images/videos and automate feature extraction from raw data, presentation, and stratification procedures. Deep learning techniques have the potential to self-extract distinguishing traits, ushering in a new age in the identification of human behavior. A subtype of machine learning called deep learning, which stresses modeling intricate data relationships by learning numerous layers of representation, is a thing. The hierarchical design of this method allows high-level concepts to be created from related low-level ideas, and several models from this category are utilized in unsupervised learning [136, 137].

According to Juan and colleagues [138], ConvNets (CNN) and Long-Short-Term Memory Recurrent Networks were used in the development of a deep learning-based technique for identifying 3D temporal 3D postures in three dimensions. Dobhal et al. [139] suggested a binary motion image-based DL technique (BMI). Three CNN layers were used to extract the element and categorize the activities and Gaussian Mixture Models (GMM) to recover the binary origin, which was then used to create BMIs. Since BMI is generated from frames, it is difficult to utilize this approach for multi-human action identification.

According to Khelalef et al. [140], human body silhouettes were utilized to calculate binary space–time mappings, which incorporate the video stream's spatial information in a certain time period, and CNN was used to acquire features and classification. Si et al. [141] present an attention graph-based CNN to concentrate on the skeleton's joint position. It aids in the enhancement of crucial node qualities. The AGCN-LSTM network can track differentiating characteristics. According to Majid et al. [142], the LSTM units should be changed to a category of action classes in a video that contains all three types of necessary information. Convolution and correlation operations are also employed to build a new structure that can extract spatial and motion information as well as their temporal dependencies. We put the recommended unit, dubbed Correlational Convolution LSTM, to the test in a deep architecture using convolutional layers. According to studies, the proposed unit enhances LSTM units that use motion information efficiently. For video-based recognition of human action, Cheng Dai et al. introduced A recurrent two-stream space–time attention network [143]. The LSTM stream encapsulates temporal sequence data, whereas the CLSTM stream extracts spatial–temporal data. Second, we created spatial–temporal

attention and used it for the feature-learning process to exploit the deep-learning feature weights.

Simonyan et al. [144] proposed the ConvNet design with two streams, one for geographical networks and the other for temporal networks. Second, we demonstrate that, despite lacking training data, a ConvNet learned on multi-frame intense optical flow may produce outstanding results. Only those areas of the surveillance stream are utilized where humans appear. The initial significant photographs from the video stream are picked using CNN-based human saliency characteristics, according to Ullah et al. [145]. Second, using temporal optical flow characteristics, a convolutional neural network was utilized to show activity in the chosen picture. Additionally, long-term memory for producing extended passages in the temporal optical flow characteristics for action detection has been provided.

Deep Belief Networks are generative probabilistic models that use observational data and labels to create a probability distribution. These networks are constructed by stacked Restricted Boltzmann Machine and excitedly training them, as proposed in [146]. The two parts of DBN are pre-training and fine-tuning. The Bolt Restricted Boltzmann Machine is utilized during pre-training (RBM). After this network has been pre-trained, the fine-tuned algorithm is used to alter the network's weights. RBM can help avoid local optimum errors because it is helpful for unsupervised learning. Using a hybrid of deep belief networks, Uddin et al. [147] summarized an algorithm for recognizing human activities. These authors proposed a modified Weber descriptor to retrieve features from motion using a single network. Another network is used to take out characteristics from images.

Sheeba et al. [148] suggested a new classifier incorporating skeletal characteristics to recognize human actions. The author developed a fuzzy-based deep belief network to classify the activity of humans. DBNs also have several drawbacks, such as the record computational cost (Complexity) of tutoring a DBN and this lack of clarity about improving the network further using the maximum likelihood training approximation. DBNs also have the disadvantage of failing to consider a parent picture's two-dimensional structure, which can significantly impact performance and usability in machine vision and multimedia evaluation. As demonstrated in [149], the Convolutional Deep Belief Network (CDBN) [150] is just a subsequent variation of such DBN that uses convolutional RBMs to use the location data of neighborhood pixels, resulting in the ability to understand a generative model that successfully scales when it comes to high-dimensional images.

Combining homogeneous convolutional neural network classifiers yields a hybrid deep neural network model. Changing the network weights network's initialization and diversifying the input characteristics generates an ensemble of classifiers. Combining recurring, regulated recurrent neural networks over various scales with shearlet-based image segmentation extraction, The shearlet transform is used to improve the resilience and segmentation of training data. Consequently, the shearlet transform and recurrent neural video categorization will attain their full potential. Savadi Hosseini et al. [151] suggested the first and second streams process the video's raw frames and optical flow, respectively. They first segmented the video frames to closely monitor the video contents and extract spatial–temporal information using 3D CNNs. The sequence of local features was given as an input to the GRU network, where the output of the two processing units, called global elements, was aggregated using a weighted averaging operator. The results of the assessments reveal that the two datasets, HMDB51 and UCF101, generate acceptable results.3D CNN may be used to increase the performance of CNN since it has the capacity to leverage spatiotemporal properties. Ensemble learning is another potential area of improvement.

However, DL-based approaches for effective feature extraction necessitate a sizable quantity of training data, whereas handcrafted features have a lower discriminative

potential for HAR. As a result, Wang et al. [152] combine the benefits of handcrafted and DL-based features utilizing better trajectories. Another name for dual-stream ConvNets is the Trajectory- pooled Deep-Convolutional Descriptor (Tpdd).

Numerous Deep Learning method for action recognition is presented with the help of CNN, but these methods cannot detect real-time action with superior accuracy. Some researchers used optimized CNN for feature extraction and Deep Autoencoder for surveillance frames [153]. Action detection from the surveillance data stream is an example of real-time action identification, and some visual sensors are also used for collecting real-time data, like smartphones.

In Table 4, we include numerous techniques for Human Activity Recognition that use DL methodologies and the accuracy of those approaches. In the previous sections, we studied the handcrafted, deep-learning method for feature extraction and compared these methods in Table 5.

9 Limitation in human activity recognition

Although a lot of progress has been made in the HAR field, certain issues still need to be fixed. The current state of the art allows for identifying fundamental and elemental actions by a single subject. The model must also be pre-trained on comparable tasks. There are a lot of activities that are made up of more than a straightforward activity. For example, working out is a composite activity of separate parts like sitting, standing, and running. Compared to recognizing atomic activities, this complex activity is complicated to identify.

Nearly every method shown so far can identify a single subject's actions. Existing work in tracking, gesture recognition, posture recognition, and other fields focuses on identifying the activities of a single individual at a time. But in the actual World, there are many instances in which tasks are done concurrently by several individuals or when multiple people are engaged in a single action (Hugging). The consideration of overlapping activities is another area of focus. For example, a person may walk while also gripping on the phone. About such a specific issue, there have only been a few publications.

Existing solutions are now addressing the issue of activity recognition variability. When one individual does an activity and another performs the same action at a different pace, we refer to the action as the variable. When a different person performs an identical activity, the identification accuracy of many existing systems declines dramatically. In addition, the system's performance decreases when the same person does the same work differently. Modern systems must solve this variability issue, which requires them to be robust. It remains unresolved, so further research is necessary.

Concurrent activity performed by a person is still to be explored. Because researchers are only focused on a single activity performed by humans, a person may, for instance, browse a newspaper while having tea or eat a meal while surfing on a mobile. Not much exploration has been carried out in this domain, and improvement is needed.

Daily activity-based HAR systems must continually identify activities (untrimmed videos). Action recognition from continuous video feeds is referred to as online action recognition. An online action recognition method can be introduced to handle large-scale data in the future.

Efficient feature extraction techniques are needed to focus on due to accurate action recognition from overlapping activity. Real-time action recognition is more focused when we

Table 4 Performance metrics for different deep learning models

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
Wang et al. [152]	2015	2 Stream CNN	HMDB-51 UCF101	65 91.5	NA NA	NA NA	Researchers recommend a powerful video presentation referred to as a trajectory-pooled deep convolutional descriptor (TDD), which combines the benefits of hand-crafted and deep-learned characteristics	Not applicable for longest sequence of videos
Karpathy et al. [97]	2014	ConvNet	Sports-1 M UCF-101	63.9 63.3	NA 68	NA NA	The performance of convolutional neural networks in huge video categorization was investigated. It discovered that CNN architectures can learn strong features from weakly-labeled data that significantly outperform feature-based approaches, and these advantages are remarkably resistant to the connectedness of the designs over time	Prearrange network is not worthy of action recognition

Table 4 (continued)

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
Shi et al. [154]	2017	Three streams CNN	KTH UCF101 HMDB51	96.8 92.2 65.2	NA NA NA	NA NA NA	In this article, dense trajectories are represented in two-dimensional planes, and a CNN-RNN network is then used to develop an effective representation for long-term motion. The sDTD stream is added into a three-stream architecture instead of the common two-stream Conv-Nets to recognize activities from a video sequence	Intricacy occurs during the training phase of the model and some redundant features are created during fusion
Liu et al. [155]	2017	Multi-Stream CNN	"NTU-RGBD (CS) NTU-RGBD (CV) MSRC-12 (CS) Northwestern-UCLA	80.03 87.21 96.62 92.61	NA NA NA NA	NA NA NA NA	This work seeks to recognize skeletal sequences from several perspectives	Intrinsic relationships amongst joint angles are not investigated in these frameworks
Ullah et al. [153]	2019	CNN	UCF-50 UCF-101 YouTube action HMDB51	96.4 94.33 96.21 70.33	93.21 89.32 69.06 95.41	91.24 90.35 62.34 93.87	The author provided a DAE-optimized human framework for actions representation that may be applied in extremely dynamic circumstances	Individual activity can only be monitored and identified by the approach. Overlapping activity should be noticed

Table 4 (continued)

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
Li et al. [156]	2017	Covnets & LSTM	NTU-RGBD (CS) NTU-RGBD (CV)	82.89 90.10	NA NA	NA NA	To acquire as much spatial–temporal data as feasible, LSTM and CNN are used for successful recognition with score fusion	Training time is so high and requires more processing power
Das et al. [157]	2019	3D CNN and LSTM	NTU-RGBD (CS) NTU-RGBD (CV) Northwestern-UCLA	93 95.4 93.1	NA NA NA	NA NA NA	For action categorization, the study presents a classification network consisting of spatiotemporal subnetworks simulating the appearance of human body parts and an RNN attention subnetwork implementing our attention method	Static learning for action recognition with a fixed number of frames
Majid et al. [142]	2019	Correlation Convolution LSTM	UCF01 HMDB51	61.3 92.8	NA NA	NA NA	This research introduced a CNN-based feature extraction method that extracts local dependence and scale invariant acceleration time series features	This approach is not suitable for the most extended sequence motion

Table 4 (continued)

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
Dai et al. [143]	2018	Two Stream Attention-based LSTM Network	UCF11 UCF Sports j-HMDB	96.9 98.6 76.3	NA NA NA	NA NA NA	This study proposed a two-stream spatial–temporal attention network for human action recognition. LSTM encodes temporal sequence information, whereas CLSTM captures spatial–temporal information	This method does not perform better on dark videos or Multiview data
Ijjina et al. [158]	2017	CNN	SBU- K –interaction Weizmann	90.98 100	NA NA	NA NA	A multimodal approach that uses CNN to extract features from RGB and depth videos. For classifying action, an extreme learning machine (ELM) was used	Multi-view is not supported. The limited comparison shows more datasets can be used for comparison
Verma et al. [159]	2020	CNN + LSTM	UTD-MHAD CAD-60 NTU RGB + D120	96.5 NA 77.9	NA 93.2 NA	NA 91.9 NA	The ConvNet only pulls spatial characteristics from a picture, but the LSTM tends to handle a series of inputs	This approach is not evaluated the performance with real-time data

Table 4 (continued)

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
Tanberk et al. [160]	2020	3D CNN Optical Flow + LSTM	MCDS CDS	95 91.3	NA NA	NA NA	The suggested architecture is constructed by integrating a dense optical flow strategy with supplementary video movement data utilizing deep learning techniques	Other DL methods can be used to enhance the set of features
Singh et al. [161]	2020	CNN + BI-LSTM	SBU Interaction MIVIA Interaction MSR Daily Activity MSR Action Pairs	98.70 99.41 98.3 94.37	NA NA NA NA	NA NA NA NA	This model employs two deep-learned architectures: a pre-trained CNN and Bi-LSTM to extract discriminative features from RGB frames and a pre-trained CNN with fully connected layers fed a dynamic motion picture as input	Skelton stream can add to make the system more robust against challenges. This approach does not differentiate the same class of action
Mukherjee et al. [162]	2020	Ensem ConvNet (CNN)	WISDM UniMib Shar Mobi Act	97.2 99.7 95.4	NA NA NA	NA NA NA	Researchers used less resource-intensive architectures as a base model for the ensemble, and the final prediction is more accurate than the base models	Feature selection method not robust

Table 4 (continued)

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
Tasnim et al. [163]	2021	Spatio- temporal image formation + MobileNetV2 Spatio- temporal image formation + DenseNet121 Spatio- temporal image formation + ResNet18	UTD-MHAD MSR Action -3D	98.93 96 99.65 98.75 98.80 97.08	NA NA NA NA NA NA	NA NA NA NA NA NA	Researchers leverage pre-trained deep learning models to assess the utility of the proposed method's spatiotemporal representation	This approach evaluated sensor data and required more computational power
M. Bilal et al. [164]	2021	Transfer Learning Models + DAE + RNN	UCF-101	96.03	95.66	95.65	Deep Learning Pipelines (DLPL), including one pre-trained CNN model followed by a Deep Autoencoder (DAE) and an RNN with LSTM, are suggested in this study for similar or overlapping actions	Multiview, complex action needs more focus to make the approach robust against these challenges
Khan Muhammad et al. [165]	2021	DCNN + BILSTM	UCF-11 UCF-Sports J-HMDB	98.3 99.1 80.2	NA NA NA	NA NA NA	BiLSTM-based attention mechanism with a DCNN that selectively focuses on practical input frame attributes to detect human activities	Limited applicability for large action datasets as well as complex action recognition

Table 4 (continued)

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
Yair et al. [166]	2021	Temporal CNN	KARD CAD-60 MSR daily activity-3D	100 100 95.6	NA NA NA	NA NA NA	This study discussed building a temporal convolutional neural network employing spatiotemporal characteristics to assess and detect people's activity using a brief video as input	This approach is not evaluated performance on complex datasets
A. Ullah et al. [167]	2021	Lightweight CNN + DS-GRU	UCF50 UCF 101 HMDB 51 Hollywood 2 YouTube Action	95.2 95.5 72.3 71.3 97.17	91.29 86.39 64.98 68.21 92.63	89.64 83.54 61.95 66.84 91.54	The proposed method is used to track individuals in the stream throughout. The lightweight CNN method was used to extract features (pyramidical feature); after that, DS-GRU (deep skip connection gated recurrent unit) learned sequential features and classified them with the SoftMax layer. The main motive of this approach is to provide real-time action detection from the video stream and trained only two datasets	

Table 4 (continued)

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
Santosh et al. [168]	2021	Pose Estimation + CNN + GRU	ADLF UP Fall	89.64 96.7	NA 96.9	NA 96.7	Utilize key points in frames to recognize the action. Spatial features are extracted using CNN, and temporal features are learned using GRU	Setup on a single cue of frames does not help classify multiple activities in a single frame
Hritam Basak et al. [169]	2022	Modified CNN + Swarm Optimizer	UTD MHAD HDM05 NTU-RGBD + 60	98.13 90.67 89.98	NA NA NA	NA NA NA	This paper presented D-Swarm-Net, which uses a swarm intelligence-based optimization technique and the CNN framework for the HAR problem and extracts four Spatiotemporal feature vectors from skeletal joint motions. After encoding features into pictures, CNNs retrieve in-depth features. This approach uses Ant-Lion Optimization (ALO) to reduce unnecessary and misleading information	Applicability for limited datasets and feature selection methods may be improved

Table 4 (continued)

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
P U Putra et al. [170]	2022	VGG16+LSTMRes	IXMAS i3DPost	97.27 96.87	NA NA	NA NA	This work aims to characterize human activity based on numerous perspectives using a unique DNN model that employs shared weight and score fusion	Limited recognition does not recognize online action. Model performance was evaluated on two datasets
Adrian et al. [171]	2022	3DFCNN	NWUCLA UWA3D	83.6 66.6	NA NA	NA NA	3-D fully CNN model uses depth data to encode spatial or temporal information without pre-processing computation. The main objective of this model is to learn actions quickly without requiring so much computational time. Identify action within 1.09 s	Not appropriate for intra-class similarities actions
Atiya Usmani et al. [133]	2023	Deep RNN	UTD-MHAD MSR Daily Activity	99.07 91	NA NA	NA NA	Skelton's modality is used for action recognition. Perform well when the visual sequence is short	Not explored other modalities like RGB and depth data

Table 4 (continued)

Author	Year	Model	Dataset	Result in %			Observations	Limitation
				Accuracy	Precision	Recall		
Inzamam et al. [172]	2023	Genetic Algorithms + Transfer Learning Models	UT Kinect-Action MSR Action Florence 3D Youtube-8 M	99.63 99.69 98.63 91.46	NA NA NA NA	NA NA NA NA	A method is suitable for an uncontrolled environmental video sequence	This method, limited to identifying atomic actions, performs poorly when more than two objects are placed in a single frame. Intraclass similarities occur
Nikpour et al. [173]	2023	Spatio -Temporal Hard Attention + Deep Reinforcement Learning	NTU RGBD UT-Kinect SBU	90.8(CS) 96.7(CV) 99.1 88.7	NA	NA	For action recognition, the Authors used spatial and temporal attention. This approach used spatial and temporal agents to discover key frames and joints. Deep reinforcement learning was used to train these agents. Appropriate for small actions	Focusing on keyframes and joints in frames leads to less effective group activity identification

CS Cross Subject, *CV* Cross View

Table 5 Comparison of handcrafted versus Deep Learning-Based Models for HAR [174]

Characteristics	Handcrafted Feature Based Model	Deep Learning Based Models
Feature Extraction & Representation	Pre-processing algorithms or detectors are needed to determine nearly all effective patterns to maximize identification accuracy	Capability to directly extract characteristics from raw data
Data Preparation	Extensive data pre-processing and standardization are needed to gain considerable performance	In order to obtain excellent performance using frameworks that are relied on DL, data pre-processing and normalization are not required
Inter-Class & Intra-Class	Do not have a proper mechanism to handle this type of problem	In order to overcome this challenge, the characteristics of such models, which are hierarchical and translationally invariant, are obtained
Training & Computation Time	Use less memory and CPU time while still achieving an excellent training result with a smaller data set	High computing capability using a Graphical Processing Unit (GPU) to accelerate training necessitates enormous data to prevent over-fitting
	No generalized method for feature selection	Automatically extract different features from raw data like spatial, temporal and scale
	Handcrafted feature selection methods are dataset biased	Deep models are not biased for a specific dataset

get continuous data from surveillance and other sensor devices. DL methods need more computation and processing power, so lightweight methods need exploration.

The multimodal approach needs improvement to characterize overlapping actions with emotional states like (talking in anger, running with pain, and clapping with a smile). Improvement in context-based HAR when data are gathered from various sources. Machine learning needs improvement to recognize large action datasets. Some deep-learning method needs to be explored to change some layers or addition of attention layer in the model to improve recognition accuracy. The lightweight deep learning method needs to be explored to reduce computation time.

10 Conclusion & future scope

This review study covers essential modules of HAR, including handcrafted-based feature extraction, feature representation, available action datasets, and classification of activity with the support of various methods. Recent research in the field of HAR have been summarized in their framework. Numerous handcrafted feature extraction methods are described with their comparative analysis and limitations. This review study has also represented various classification methods such as ML, DL, hybrid, and some pre-trained DL approaches. A considerable amount of data is needed to train, so hybrid DL methods and pre-trained DL methods are represented with improvement. This review study presents a comparative analysis of the latest approaches to HAR and their performance evaluation matrices (precision, recall, f1-score). This review study summarized various available action datasets along with their limitations and resolutions. Most HAR datasets are recorded in a controlled environment, which does not accurately depict the real-life challenges of HAR detection. This review study concentrates on DL methods and the fusion of deep learning techniques for recognizing human action with their accuracy and future scope, which will help researchers identify research gaps in the field of HAR. The hybrid model is gaining attention due to the availability of large amounts of data. Some models are designed specifically for real-time human activity recognition. However, such models have several limitations, and these limitations need to be addressed to improve prediction accuracy in real-time. Deep learning has various features that provide better results for identifying human activity. This article also compares previous literature surveys, citing their contributions and discusses limitations in Human Activity Recognition.

Future research may focus on constructing robust human recognition approaches that can provide better performance during low light, cluttered backgrounds, dynamic light conditions, and noisy data. However, most of the models in the literature are tested in indoor scenarios but to offer a more robust system, the models should be tested in real-time environment conditions or features. Additionally, to get higher performance, it will be necessary to examine new deep learning algorithms (vision transformer-trained DL method, lightweight DL method) to efficiently detect Human Activity Recognition from poorly labeled data in the near future. It is possible to use 3D CNN to increase its performance of CNN because 3D CNN has the power to leverage spatiotemporal characteristics, which is advantageous. In most cases, labeled data collection is both costly and time-consuming. Although deep generative models (e.g., AEs and GANs) may employ unsupervised data, they are not immediately relevant to HAR. In the future, it may be possible to conduct multi-person recognition. Future scope research may be done by predicting what will happen in the future based on what is happening now. Ensemble-based solutions may

be helpful if you're working with too much data. To be sure, additional study is needed to appropriately and effectively comprehend the value of multiple classifier systems. Taking into account activities that overlap is another factor. People may be walking and hand-waving at the exact moment, for example.

Thus, this study enclosed various branches of AR to contribute towards numerous approaches, datasets, limitations, challenges and future scopes in this domain.

Funding NA.

Data availability N.A.

Code availability N.A.

Declarations

Ethics approval N.A.

Consent to participate N.A.

Consent for publication NA.

Conflicts of interest/Competing interests Authors declare that they have no conflict of interest.

References

1. Ke SR, Thuc HLU, Lee YJ, Hwang JN, Yoo JH, Choi KH (2013) A review on video-based human activity recognition. Computers 2(2): 88–131. MDPI AG. https://doi.org/10.3390/computers2020088
2. Gupta N, Gupta SK, Pathak RK et al (2022) Human activity recognition in artificial intelligence framework: a narrative review. Artif Intell Rev 55:4755–4808. https://doi.org/10.1007/s10462-021-10116-x
3. Laptev I, Lindeberg T (2004) Local descriptors for spatio-temporal recognition. In: International workshop on spatial coherence for visual motion analysis
4. Gorelick L, BlankM SE, Irani M, Basri R (2005) Actions as space-time shapes. In: The tenth IEEE international conference on computer vision (ICCV'05)
5. Rodriguez MD, Ahmed J, Shah M (2008) Action of MACH a spatio-temporal maximum average correlation height filter for action recognition. In: 26th IEEE conference on computer vision and pattern recognition, CVPR, pp 1–8
6. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (n.d.) HMDB: a large video database for human motion recognition. In: International conference on computer vision, Barcelona, pp 2556–2563. https://doi.org/10.1109/ICCV.2011.6126543
7. Reddy KK, Shah M (2012) Recognizing 50 human action categories of web videos. Machine Vision and Applications Journal (MVAP)
8. Soomro K, Zamir AR, Mubarak Shah (2012) UCF101: A dataset of 101 human action classes from videos in the wild, CRCV-TR-12-01
9. Weinland D, Boyer E, Ronfard R (2007) Action recognition from arbitrary views using 3D exemplars. In: IEEE11th international conference on computer vision, Rio de Janeiro
10. Kwapisz JR, Weiss GM, Moore SA (2011) Activity recognition using cell phone accelerometers. ACM SIGKDD Explorations Newsl 12(2):74–82. https://doi.org/10.1145/1964897.1964918
11. Chen C, Jafari R, Kehtarnavaz N (2015) UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: Proceedings of IEEE international conference on image processing. Canada
12. Heilbron FC, Escorcia V, Ghanem B, Niebles JC (n.d) ActivityNet: a large-scale video benchmark for human activity understanding. In: IEEE conference on computer vision and pattern recognition (CVPR), Boston M A

13. Wang J, Nie X, Xia Y, Wu Y, Zhu SC (2014) Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 2649–2656. https://doi.org/10.1109/CVPR.2014.339
14. Rahmani H, Mahmood A, Huynh DQ, Mian A (2014) HOPC: histogram of oriented principal components of 3D pointclouds for action recognition. Lect Notes Comput Sci (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8690 LNCS(PART 2):742–757. https://doi.org/10.1007/978-3-319-10605-2_48/COVER
15. Shahroudy A, Liu J, Ng T-T, Wang G (n.d.) NTU RGB+D: a large-scale dataset for 3D human activity analysis. In: IEEE conference on computer vision and pattern recognition (CVPR)
16. Jalal A, Kamal S, Kim D (n.d.) A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. Sensors 14(7):11735–11759
17. Liu J, Shahroudy A, Perez M, Wang G, Duan L-Y, Kot AC (n.d.) NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. In: IEEE transactions on pattern analysis and machine intelligence (TPAMI)
18. Kay W et al (2017) The Kinetics Human Action Video Dataset. [Online]. Available: http://arxiv.org/abs/1705.06950
19. Li A, Thotakuri M, Ross DA, Carreira J, Vostrikov A, Zisserman A (2020) The AVA-Kinetics Localized Human Actions Video Dataset, [Online]. Available: http://arxiv.org/abs/2005.00214
20. Damen D, Doughty H, Farinella GM et al (2022) Rescaling egocentric vision: collection, pipeline and challenges for EPIC-KITCHENS-100. Int J Comput Vis 130:33–55. https://doi.org/10.1007/s11263-021-01531-2
21. Carreira J, Noland E, Banki-Horvath A, Hillier C, Zisserman A (2018) A Short Note about Kinetics-600, [Online]. Available: http://arxiv.org/abs/1808.01340
22. Carreira J, Noland E, Hillier C, Zisserman A (2019) A Short Note on the Kinetics-700 Human Action Dataset, [Online]. Available: http://arxiv.org/abs/1907.06987
23. Monfort M et al (2018) Moments in Time Dataset: one million videos for event understanding, [Online]. Available: http://arxiv.org/abs/1801.03150
24. Niebles JC, Wang H, Fei-Fei L (n.d.) Unsupervised learning of human action categories using spatio-temporal words. Int J Comput Vis 79:299–318
25. Calderara S, Cucchiara R, Prati A (n.d.) Action signature: a novel holistic representation for action recognition. In: Proc. IEEE 5th international conference on advanced video and signal-based surveillance, pp 121–128
26. Kalal Z, Mikolajczyk K, Matas J (2012) Tracking-learning-detection. IEEE Trans Pattern Anal Mach Intell 34(7):1409–1422. https://doi.org/10.1109/TPAMI.2011.239
27. Iosifidis A, Tefas A, Pitas I (2012) Neural representation and learning for multi-view human action recognition. In: The 2012 international joint conference on neural networks (IJCNN), Brisbane, pp 1–6. https://doi.org/10.1109/IJCNN.2012.6252675
28. Lu Y et al (2012) A human action recognition method based on Tchebichef moment invariants and temporal templates. In: 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, 2:76–79
29. Ji X, Liu H (2010) Advances in view-invariant human motion analysis: a review. In: IEEE transactions on systems, man, and cybernetics, Part C (applications and reviews), 40(1):13–24. https://doi.org/10.1109/TSMCC.2009.2027608
30. Estevam V, Pedrini H, Menotti D (2021) Zero-shot action recognition in videos: a survey. Neurocomputing 439:59–175. https://doi.org/10.1016/j.neucom.2021.01.036
31. Pareek P, Thakkar A (n.d.) A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artif Intell Rev 54:2259–2322
32. Dang LM, Min K, Wang H, Piran MJ, Lee CH, Moon HJ (2020) Sensor-based and vision-based human activity recognition: a comprehensive survey. Pattern Recogn 108(107561):31–3203
33. Beddiar DR, Nini B, Sabokrou M et al (2020) Vision-based human activity recognition: a survey. Multimed Tools Appl 79:30509–30555. https://doi.org/10.1007/s11042-020-09004-3
34. Zhang H-B, Zhang Y-X, Zhong B, Lei Q, Yang L, Du J-X, Chen D-S (2019) A comprehensive survey of vision-based human action recognition methods. Sensors 19:1005. https://doi.org/10.3390/s19051005
35. Herath S, Harandi M, Porikli F (2017) Going deeper into action recognition: a survey. Image and Vision Computing 60:4–21. https://doi.org/10.1016/j.imavis.2017.01.010
36. Singh PK, Kundu S, Adhikary T, Sarkar R, Bhattacharjee D (2021) Progress of Human Action Recognition Research in the Last Ten Years: A Comprehensive Survey. Arch Comput Methods Eng 29:4:2309–2349. https://doi.org/10.1007/S11831-021-09681-9
37. Jobanputra H, Bavishi J, Doshi N (2019) Human activity recognition: a survey. Procedia Comput Sci 155:698–703. https://doi.org/10.1016/j.procs.2019.08.100

38. Kong Y, Yun Raymond F (2018) Human action recognition and prediction: a survey. Int J Comput Vis 130:1366–1401
39. Guangchun C, Yiwen W, Abdullah S, Kamesh N, Bill B (2015) Advances in human action recognition: A survey
40. Vishwakarma S, Agrawal A (n.d.) A survey on activity recognition and behavior understanding in video surveillance. Vis Comput 29(10):983–1009
41. Aggarwal JK, Ryoo MS (2011) Human activity analysis. ACM Computing Surveys (CSUR) 43:1–43
42. Bobick AF, Davis JW (n.d.) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267
43. Sheikh Y, Sheikh M, Shah M (n.d.) Exploring the space of a human action. In: Tenth IEEE Int Conf on Computer Vision, pp 144–149
44. Amor BB, Su J, Srivastava A (n.d.) Action recognition using rate-invariant analysis of skeletal shape trajectories. Trans Pattern Anal Mach Intell 38:1–13
45. Wang H, Kläser A, Schmid C, Liu C (n.d.) Action recognition by dense trajectories. CVPR 3169–3176
46. Laptev I, Lindeberg T (n.d.) Space-time interest points. In: Proc. 9th IEEE Int. Conf. On computer vision, pp 432–439
47. Dollar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse Spatio-temporal features. In: IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance
48. Bregonzio M, Gong S, Xiang T (2009) Recognising action as clouds of space-time interest points. In: 2009 IEEE conference on computer vision and pattern recognition, Miami, pp 1948–1955. https://doi.org/10.1109/CVPR.2009.5206779
49. Thi TH, Zhang J, Cheng L, Wang L, Satoh S (n.d.) Human action recognition and localization in video using structured learning of local space-time features. IEEE International Conference on Advanced Video and Signal Based Surveillance, pp 204–211
50. Sadek S, Al-Hamadi A, Michaelis B, Sayed U (n.d.) An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity. EURASIP J Adv Signal Process
51. Chaudhry R, Ravichandran A, Hager G, Vidal R (n.d.) Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: IEEE computer Soc. Conf. Computer. Vis. Pattern recognition work. CVPR work. IEEE, pp 1932–1939
52. Yuan C, Li X, Hu W, Ling H, Maybank S (n.d.) 3D R transform on spatio-temporal interest points for action recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 724–730
53. Sahoo SP, Silambarasi R, Ari S (n.d.) Fusion of histogram-based features for human action recognition. In: 5th international conference on advanced computing & communication systems, pp 1012–1016
54. Gupta S, Mazumdar S, Student M (2013) Sobel edge detection algorithm.
55. Teoh SH, Ibrahim H (n.d) Median filtering frameworks for reducing impulse noise from grayscale digital images: a literature survey. Int J Future Comput Commun 1:323–326
56. Le QV, Zou WY, Yeung SY, Ng AY (n.d.) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 3361–3368
57. Darrell T, Pentland A (n.d.) Space-time gestures. In: Proc. IEEE computer society Conf. On computer vision and pattern recognition, pp 335–340
58. Jiang H, Drew MS, Li ZN (n.d.) Successive convex matching for action detection. In: IEEE computer society Conf. On computer vision and pattern recognition, pp 1646–1653
59. Oliver NM, Rosario B, Pentland AP (n.d.) A Bayesian computer vision system for modelling human interactions. IEEE Trans Pattern Anal Mach Intell 22(8):831–843
60. Shi Q, Cheng L, Wang L, Smola A (n.d.) Human action segmentation and recognition using discriminative semi-Markov models. Int J Comput Vis 93:22–32
61. Oliver N, Horvitz E, Garg A (n.d) Layered representations for human activity recognition. In: Proc. 4th IEEE Int. Conf. On multimodal interfaces, pp 3–8
62. Zhang D, Gatica-Perez D, Bengio S, McCowan I (n.d.) Modelling individual and group actions in meetings with layered hmms. IEEE Trans Multimed 8(3):509–520
63. Nguyen NT, Phung DQ, Venkatesh S, Bui H (n.d.) Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model, IEEE computer society Conf on computer vision and pattern recognition, pp 955–960

64. Shi Y, Huang Y, Minnen D, Bobick A, Essa I (n.d.) Propagation networks for recognition of partially ordered sequential action. In: Proc. of IEEE computer society Conf. On computer vision and pattern recognition, pp 862–869
65. Iosifidis A, Tefas A, Pitas I (n.d.) Action-based person identification using fuzzy representation and discriminant learning. IEEE Trans Inf Forensics Secur 7:530–542
66. Xu W, Miao Z, Zhang X, Tian Y (n.d.) Learning a hierarchical spatio-temporal model for human activity recognition. In: International conference on acoustics, speech and signal processing (ICASSP). IEEE, New Orleans, pp 1607–1611
67. Kitani KM, Sato Y, Sugimoto A (2007) Recovering the basic structure of human activities from a video-based symbol string. In: 2007 IEEE workshop on motion and video computing (WMVC'07), Austin, p 9. https://doi.org/10.1109/WMVC.2007.34
68. Ivanov Y, Bobick A (n.d.) Recognition of visual activities and interactions by stochastic parsing. IEEE Trans Pattern Anal Mach Intell 22:852–872
69. Moore D, Essa I (n.d.) Recognizing multitasked activities from video using stochastic context-free grammar. AAAI National Conference on Artificial Intelligence, pp 770–776
70. Minnen D, Essa I, Starner T (n.d.) Expectation grammars: leveraging high-level expectations for activity recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 626–632
71. Joo SW, Chellappa R (n.d.) Attribute grammar-based event recognition and anomaly detection. IEEE Conference on Computer Vision and Pattern Recognition Workshop, pp 107–114
72. Siskind JM (n.d.) Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. J Artif Intell Res 15:31–90
73. Gupta A, Srinivasan P, Shi J, Davis L (n.d.) Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2012–2019
74. Ijsselmuiden J, Stiefelhagen R (n.d.) Towards high-level human activity recognition through computer vision and temporal logic. In: The 33rd annual German conference on advances in artificial intelligence, pp 426–435
75. Khare M, Jeon M (2022) Multi-resolution approach to human activity recognition in video sequence based on combination of complex wavelet transform, Local Binary Pattern and Zernike moment. Multimed Tools Appl 81(24):34863–34892. https://doi.org/10.1007/S11042-021-11828-6/FIGURES/10
76. Li C, Huang Q, Li X, Wu Q (2021) Human action recognition based on multi-scale feature maps from depth video sequences. Multimed Tools Appl 80(21–23):32111–32130. https://doi.org/10.1007/S11042-021-11193-4/TABLES/8
77. Ikizler N, Duygulu PD (n.d.) Histogram of oriented rectangles: a new pose descriptor for human action recognition. Image Vis Comput 27(10):1515–1526. https://doi.org/10.1016/j.imavis.2009.02.002
78. Kellokumpu V, Zhao G, Pietikäinen M (n.d.) Recognition of human actions using texture descriptors. Mach Vis Appl 22:767–780
79. Kliper-Gross O, Gurovich Y, Hassner T, Wolf L (n.d.) Motion interchange patterns for action recognition in unconstrained videos. In: European conference on computer vision. Springer, Berlin/Heidelberg, pp 256–269
80. Jiang YG, Dai Q, Xue X, Liu W, Ngo CW (n.d.) Trajectory-based modeling of human actions with motion reference points. In: European conference on computer vision. Springer, Berlin/Heidelberg, pp 425–438
81. Wang C, Wang Y, Yuille AL (n.d.) An approach to pose-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Portland, OR, USA, pp 915–922
82. Zanfir M, Leordeanu M, Sminchisescu C (n.d.) The moving pose: an efficient 3d kinematics descriptor for low-latency action recognition and detection. In: Proceedings of the IEEE international conference on computer vision. Sydney, Australia, pp 2752–2759
83. Chaaraoui AA, Climent-Pérez P, Flórez-Revuelta F (n.d.) Silhouette-based human action recognition using sequences of key poses. Pattern Recogn Lett 34:1799–1807
84. Rahman SA, Song I, Leung MK, Lee I, Lee K (n.d.) Fast action recognition using negative space features. Expert Syst Appl 41:574–587
85. Junejo IN, Junejo KN, Al Aghbari Z (n.d) Silhouette-based human action recognition using SAX-shapes. Vis Comput 30:259–269
86. Vishwakarma DK, Kapoor R, Dhiman A (n.d.) A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics. Robot Auton Syst 77:25–38

87. Jalal A, Kim YH, Kim YJ, Kamal S, Kim D (n.d.) Robust human activity recognition from depth video using spatiotemporal multi-fused features. Pattern Recogn 61:295–308
88. Patrona F, Chatzitofis A, Zarpalas D, Daras P (2018) Motion analysis: action detection, recognition and evaluation based on motion capture data. Pattern Recogn 76:612–622
89. Zhang C, Xu Y, Xu Z et al (2022) Hybrid handcrafted and learned feature framework for human action recognition. Appl Intell 52:12771–12787. https://doi.org/10.1007/s10489-021-03068-w
90. Bengio Y (n.d) Learning deep architectures for AI. Found Trends Mach Learn 2:1–127
91. Ji S, Xu W, Yang M, Yu K (n.d.) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
92. Weimer D, Scholz-Reiter B, Shpitalni M (n.d.) Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. CIRP Ann Manuf Technol 65(1):417–420
93. Le QV (n.d.) Building high-level features using large scale unsupervised learning. In: 2013 IEEE Int. Conf. On acoustics, speech and signal processing (ICASSP)
94. Huang Y, Lai S-H, Tai S-H (n.d.) Human action recognition based on temporal pose CNN and multi-dimensional fusion. In: Proceedings of the European conference on computer vision (ECCV)
95. Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18(5):851–869. https://doi.org/10.1093/bib/bbw068
96. Krizhevsky A, Sutskever I, Hinton GE (n.d.) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, Lake Tahoe, Nevada, pp 1097–1105
97. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li FF (n.d.) Large-scale video classification with convolutional neural networks, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit, pp 1725–1732
98. Ravi D, Wong C, Lo B, Yang GZ (n.d.) Deep learning for human action recognition: a resource efficient implementation on low-power devices. In: BSN 2016—13th annual body sensor networks conference, pp 71–76
99. Marjaneh S, Hassan F (2017) Single image action recognition by predicting space-time saliency
100. Banerjee A, Singh PK, Sarkar R (n.d.) Fuzzy integral based CNN classifier fusion for 3D skeleton action recognition. IEEE Trans Circ Syst Video Technol 31(6):2206–2216
101. Ng A (n.d.) Sparse autoencoder. CS294A Lect Note 72:1–19
102. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A (n.d.) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11:3371–3408
103. Hasan M, Roy-Chowdhury AK (n.d.) A continuous learning framework for activity recognition using deep hybrid feature models. IEEE Trans Multimed 17:11
104. Wang X, Gao L, Song J, Zhen X, Sebe N, Shen HT (n.d.) Deep appearance and motion learning for egocentric activity recognition. Neurocomputing 275:438–447
105. Gao X, Luo H, Wang Q, Zhao F, Ye L, Zhang Y (2019) A human activity recognition algorithm based on stacking Denoising autoencoder and LightGBM. Sensors. 19(4):947. https://doi.org/10.3390/s19040947
106. Du Y, Wang W, Wang L (n.d.) Hierarchical recurrent neural network for skeleton-based action recognition. Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 1110–1118
107. Graves A (n.d.) Generating sequences with recurrent neural networks.
108. Salehinejad H, Sankar S, Barfett J, Colak E, Valaee S (n.d.) Recent advances in recurrent neural networks.
109. Qi M, Wang Y, Qin J, Li A, Luo J, Gool L (n.d.) stagNet: an attentive semantic RNN for group action and individual action recognition. IEEE Trans Circ Syst Video Technol 30:1
110. Liu J, Shahroudy A, Xu D, Wang G (n.d.) Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 9907. LNCS, pp 816–833
111. Cho K et al (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing, pp 1724–1734
112. Goodfellow I et al (n.d.) Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp 2672–2680
113. Huang GB, Lee H, Learned-Miller E (n.d.) Learning hierarchical representations for face verifcation with convolutional deep belief networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR' 12), pp 2518–2525

114. Radford A, Metz L, Chintala S (n.d.) Unsupervised representation learning with deep convolutional generative adversarial networks.
115. Zadeh MZ, Babu AR, Jaiswal A, Makedon F (n.d.) Self-supervised human activity recognition by augmenting generative adversarial networks, p 11755
116. Li R, Pan J, Li Z, Tang J (n.d.) Single image Dehazing via conditional generative adversarial network.
117. Yang Y, Hou C, Lang Y, Guan D, Huang D, Xu J (n.d.) Open-set human activity recognition based on micro-Doppler signatures. Pattern Recogn 85:60–69
118. Gammulle H, Denman S, Sridharan S, Fookes C (2019) Multi-level sequence GAN for group activity recognition. In: Jawahar C, Li H, Mori G, Schindler K (eds) Computer vision – ACCV 2018. Lecture notes in computer science(), vol 11361. Springer, Cham. https://doi.org/10.1007/978-3-030-20887-5_21
119. Ahsan U, Sun C, Essa I (n.d.) DiscrimNet: semi-supervised action recognition from videos using generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. Women in computer vision (WiCV'17)
120. Donahue J et al (n.d.) Long-term recurrent convolutional networks for visual recognition and description. CVPR
121. Kar A, Rai N, Sikka K, Sharma G (n.d.) Adascan: adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. CVPR
122. Jaouedi N, Boujnah N, Bouhlel MS (n.d.) A new hybrid deep learning model for human action recognition. J King Saud Univ - Comput Inf Sci 32
123. Gowda SN (2017) Human activity recognition using combinatorial deep belief networks. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 1589–1594. https://doi.org/10.1109/CVPRW.2017.203
124. Wu Z, Wang X, Jiang Y-G, Ye H, Xue X (2015) Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: Proceedings of the 23rd ACM international conference on multimedia (MM '15). Association for Computing Machinery, New York, pp 461–470. https://doi.org/10.1145/2733373.2806222
125. Lv M, Xu W, Chen T (n.d.) A hybrid deep convolutional and recurrent neural network for complex activity recognition using multimodal sensors. Neurocomputing 362
126. Ijjina EP, Mohan CK (n.d.) Hybrid deep neural network model for human action recognition. Appl. Soft Comput 46:936–952
127. Al-Azzawi NA (n.d.) Human action recognition based on hybrid deep learning model and Shearlet transform. In: 2020 12th international conference on information technology and electrical engineering (ICITEE, Yogyakarta), pp 152–155
128. Yadav SK, Tiwari K, Pandey HM, Akbar SA (2022) Skeleton-based human activity recognition using ConvLSTM and guided feature learning. Soft comput 26(2):877–890. https://doi.org/10.1007/S00500-021-06238-7/FIGURES/11
129. Wensel J, Ullah H, Member S, Munir A, Member S (2022) ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos. Accessed: May 11, 2023. [Online]. Available: https://arxiv.org/abs/2208.07929v2
130. Challa SK, Kumar A, Semwal VB (2022) A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data. Vis Comput 38(12):4095–4109. https://doi.org/10.1007/S00371-021-02283-3/TABLES/7
131. Jiang N, Quan W, Geng Q, Shi Z, Xu P (2023) Exploiting 3D human recovery for action recognition with Spatio-temporal bifurcation fusion. In: ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096404
132. Merlo E, Lagomarsino M, Lamon E, Ajoudani A (2023) Automatic Interaction and Activity Recognition from Videos of Human Manual Demonstrations with Application to Anomaly Detection. Accessed: May 12, 2023. [Online]. Available: http://arxiv.org/abs/2304.09789
133. Usmani A, Siddiqui N, Islam S (2023) Skeleton joint trajectories based human activity recognition using deep RNN. Multimed Tools Applic 2023:1–25. https://doi.org/10.1007/S11042-023-15024-6
134. Yin M, He S, Soomro TA, Yuan H (2023) Efficient skeleton-based action recognition via multi-stream depthwise separable convolutional neural network. Expert Syst Appl 226:120080. https://doi.org/10.1016/J.ESWA.2023.120080
135. Barkoky A, Charkari NM (2022) Complex Network-based features extraction in RGB-D human action recognition. J Vis Commun Image Represent 82:103371. https://doi.org/10.1016/J.JVCIR.2021.103371
136. Deng L (n.d.) A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Trans Signal Inf Process 3:2

137. Dosovitskiy A, Fischer P, Springenberg JT (n.d.) Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Trans Pattern Anal Mach Intell 38(9):1734–1747
138. Núñez JC, Cabido R, Pantrigo JJ, Montemayor AS, Vélez JF (2018) Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recogn 76
139. Dobhal T, Shitole V, Thomas G, Navada G (n.d.) Human activity recognition using binary motion image and deep learning. Procedia Comput Sci 58:178–185
140. Khelalef A, Ababsa F, Benoudjit N (2019) An efficient human activity recognition technique based on deep learning. Pattern Recognit Image Anal 29:702–715
141. Si C, Chen W, Wang W, Wang L, Tan T (n.d.) An attention enhanced graph convolutional LSTM network for skeleton-based action recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1227–1236
142. Majd M, Safabakhsh R (2020) Correlational convolutional LSTM for human action recognition. Neurocomputing 396:224–229. https://doi.org/10.1016/j.neucom.2018.10.095
143. Dai C, Liu X, Lai J (n.d.) Human action recognition using two-stream attention-based LSTM networks. Appl Soft Comput
144. Simonyan K, Zisserman A (n.d.) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp. 568–576
145. Ullah A, Muhammad K, Ser JD, Baik SW, Albuquerque VHC (n.d.) Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM. IEEE Trans Ind Electr 66(12):9692–9702
146. Hinton GE, Osindero S, Teh Y-W (n.d.) A fast-learning algorithm for deep belief nets. Neural Comput 18:1527–1554
147. Uddin MZ (n.d.) Facial expression recognition utilizing local direction-based robust features and deep belief network. IEEE Access 5:4525–4536
148. Sheeba PT, SSM, Rani SD (n.d.) Fuzzy Based Deep Belief Network for Activity Recognition. In: Proceedings of International Conference on Recent Trends in Computing, Communication & Networking Technologies (ICRTCCNT)
149. Lee H, Grosse R, Ranganath R, Ng AY (n.d.) Unsupervised learning of hierarchical representations with convolutional deep belief networks. Commun ACM 54(10):95–103
150. Li X et al (n.d.) Region-based Activity Recognition Using Conditional GAN. In: Proceedings of the 25th ACM international conference on Multimedia Association for Computing Machinery, New York, NY, USA, pp. 1059–1067
151. Savadi Hosseini M, Ghaderi F (n.d.) A Hybrid Deep Learning Architecture Using 3D CNNs and GRUs for Human Action Recognition. Int J Eng 33(5):959–965
152. Wang L, Qiao Y (n.d.) Tang X Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4305–4314
153. Ullah A, Muhammad K, Haq IU (n.d.) Baik SW Action recognition using optimized deep Autoencoder and CNN for surveillance data streams of non-stationary environments. Future Gener Comput 96:386–397
154. Shi Y, Tian Y, Wang Y, Huang T (n.d.) Sequential deep trajectory descriptor for action recognition with three-stream cnn. IEEE Trans Multimed 19(7):1510–1520
155. Liu M, Liu H, Chen C (n.d.) Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognit 68:346–362
156. Li C, Wang P, Wang S, Hou Y, Li W (n.d.) Skeleton-based action recognition using LSTM and CNN. In: IEEE international conference on multimedia and expo workshops (ICMEW). IEEE, pp. 585–590
157. Das S, Chaudhary A, Bremond F, Thonnat M (n.d.) Where to focus on for human action recognition? In: IEEE winter conference on applications of computer vision (WACV). IEEE, pp. 71–80
158. Ijjina EP, Chalavadi KM (n.d.) Human action recognition in RGB-D videos using motion sequence information and deep learning. Recognition 72:pp. 31–3203, 504–516
159. Verma P, Sah A, Srivastava R (n.d.) Deep learning-based multi-modal approach using RGB and skeleton sequences for human activity recognition. Multimed Syst 26:671–685
160. Tanberk S, Kilimci ZH, Tükel DB, Uysal M, Akyokuş S (n.d.) A Hybrid Deep Model Using Deep Learning and Dense Optical Flow Approaches for Human Activity Recognition. IEEE Access 8:19799–19809
161. Singh T, Vishwakarma DK (n.d.) A deeply coupled ConvNet for human activity recognition using dynamic and RGB images. Neural Comput Applic 33:469–485

162. Mukherjee D, Mondal R, Singh PK (n.d.) EnsemConvNet: a deep learning approach for human activity recognition using smartphone sensors for healthcare applications. Multimed Tools Appl 79:31663–31690
163. Tasnim N, Islam MK, Baek J-H (2021) Deep Learning Based Human Activity Recognition Using Spatio-Temporal Image Formation of Skeleton Joints. Appl Sci 11(6):2675
164. Bilal M, Maqsood M, Yasmin S (n.d.) A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. J Supercomput 78:2873–2908
165. Muhammad K et al (n.d.) Human action recognition using attention-based LSTM network with dilated CNN features. Future Gener Comput Syst 125:820–830, pp. 167–739
166. Andrade-Ambriz YA, Ledesma S, Ibarra-Manzano M-A, Oros-Flores MI, Almanza-Ojeda D-L (2022) Human activity recognition using temporal convolutional neural network architecture. Expert Syst Appl 191:116287
167. Ullah A, Muhammad K, Ding W, Palade V, Haq IU, Baik SW (2021) Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications. Appl Soft Comput 103:107102. https://doi.org/10.1016/J.ASOC.2021.107102
168. Yadav SK, Luthra A, Tiwari K, Pandey HM, Akbar SA (2022) ARFDNet: An efficient activity recognition & fall detection system using latent feature pooling. Knowl Based Syst 239:107948. https://doi.org/10.1016/J.KNOSYS.2021.107948
169. Basak H, Kundu R, Singh PK, Ijaz MF, Woźniak M, Sarkar R (2022) A union of deep learning and swarm-based optimization for 3D human action recognition. Sci Rep 12(1). https://doi.org/10.1038/s41598-022-09293-8
170. Putra PU, Shima K, Shimatani K (n.d.) A deep neural network model for multi-view human activity recognition. PLoS One 17(1):262181
171. Sánchez-Caballero A et al (2022) 3DFCNN: real-time action recognition using 3D deep neural networks with raw depth information. Multimed Tools Appl 81(17):24119–24143. https://doi.org/10.1007/S11042-022-12091-Z/TABLES/7
172. Nasir IM, Raza M, Ulyah SM, Shah JH, Fitriyani NL, Syafrudin M (2023) ENGA: Elastic Net-Based Genetic Algorithm for human action recognition. Expert Syst Appl 227:120311. https://doi.org/10.1016/J.ESWA.2023.120311
173. Nikpour B, Armanfard N (2023) Spatio-temporal hard attention learning for skeleton-based activity recognition. Pattern Recognit 139:109428. https://doi.org/10.1016/J.PATCOG.2023.109428
174. Al-Faris M, Chiverton J, Ndzi D, Ahmed AI (n.d.) A Review on Computer Vision-Based Methods for Human Action Recognition. J Imaging 10;6(6):46

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

Brazilian Journal of Physics (2024) 54:8
https://doi.org/10.1007/s13538-023-01383-w

GENERAL AND APPLIED PHYSICS

Analytical Modelling of Inhomogeneous Energy Density Driven (IEDD) Instability in a Magnetized Dusty Plasma Cylinder

Anshu.[1] · **Jyotsna Sharma**[2] · **Suresh C. Sharma**[1]

Received: 1 August 2023 / Accepted: 24 October 2023
© The Author(s) under exclusive licence to Sociedade Brasileira de Física 2023

Abstract
The theoretical modelling of an inhomogeneous energy density driven (IEDD) instability has been studied while incorporating the effect of dust particles in a collisional magnetized plasma cylinder with a transverse direct current (dc) electric field. Plasma parameters such as finite gyro-radius, relative density ratio, electric field, magnetic field, dust density and size are all taken into account by our model to find out the impact on the growth rate and real frequency of IEDD instability. It was found that with the relative density ratio, gyro-radius and wave vector, the real frequency and instability's growth rate augment. In this work, a comparative analysis of the instability has been done in the presence and absence of dust grains. The impact of different plasma parameters has been studied on the behaviour of the instability i.e., frequency and growth rate. The findings revealed that the introduction of dust grains in the plasma resulted in a decrease in the growth rate of instability. The response of dust grains number density and dust grain size on the IEDD instability was evaluated and it was observed that when the dust grains number density and dust grain size augment, the growth rate of the instability reduces. Further, the effects of an electric field on the IEDD instability were also examined and it was observed that in the presence of an electric field, the growth rate and real frequency augment. The current study's findings align with the existing experimental observations.

Keywords Dusty plasma · IEDD instability · Fluid theory · Growth rate

1 Introduction

The study of plasma has become indispensable due to its pervasive nature. Therefore, it is not surprising that research into "plasmas" [1–6] is proposed in several waves [7–12], instabilities [13–16], and astrophysical phenomena [17–20]. The dust may have a significant impact on the plasma, and it was discovered in the solar wind, comet tails, industries [21–23] and laboratories [24–27]. The nature of plasma as a whole is modified by the presence of these massive particles affecting the overall quasi-neutrality condition in waves [28–31] and instabilities [13]. Due to the potential applications of waves and instabilities [32] in plasma, like Hall thrusters (which are forefront propulsion systems that effectively use plasma to move spacecraft, making them perfect for deep-space missions [33–35], this field of research is of great interest.

Instabilities in plasma have been investigated by numerous researchers [15, 36–39] through diverse methods, including kinetic theory, kappa distribution and fluid theory. Additionally, it has been noted that relative drift or inhomogeneities caused by an electric field tend to drive these instabilities [40–43]. Among these instabilities, the two major instabilities are inhomogeneous energy density driven instability and the density gradient driven instability. The major difference is that the inhomogeneous energy density instability develops when the energy density inside a fluid or plasma is distributed unevenly. Rather than only fluctuations in density, it is also driven by variations in the electric field of the system. This instability can arise as a result of disturbances (localisation in electric field) varying from one region to another region. On the other side, density gradient driven instability is caused when the density of the plasma varies throughout different regions of the plasma. This instability is driven by the buoyant forces which are related with the density gradients [44–47].

✉ Suresh C. Sharma
suresh321sharma@gmail.com

[1] Department of Applied Physics, Delhi Technological University (DTU), Bawana Road Delhi-110042, India

[2] Department of Physics, Amity School of Applied Sciences, Amity University Haryana, Manesar Gurugram-122051, India

Published online: 16 November 2023

An IEDD instability [11, 15, 39, 48] has drawn significant interest in a plethora of space, planetary, and geophysical systems. This instability can occur because of the free energy provided by the inhomogeneous electric field. Inhomogeneities in the background of the plasma may result in an area with negative energy density, and a nonlocal wave packet may couple the areas of both the energies, i.e., positive and negative, which may result in waves to grow [37, 38, 49]. This is the physical mechanism underlying the IEDD instability. IEDD waves have a wide frequency range that is close to the ion cyclotron frequency and typically propagate in the $\vec{E} \times \vec{B}$ direction [15, 50–52]. The key parameter to use in observing the change from one type of wave to another is the parameter R, i.e. $R = k_\theta v_E(r) \big/ k_z v_{ed}$, the ratio of the azimuthal and axial Doppler shifts in cylindrical coordinates [15, 53]. The electrostatic ion cyclotron (EIC) waves are driven by the current, a process which is known as EIC instability [7] for $R << 1$,i.e., small v_E and large v_{ed} cases. For $R >> 1$, i.e. the situation of large v_E and small v_{ed}, the EIC waves are driven by an inhomogeneous electric field which is known as IEDD instability [38, 54]. The IEDD instability is an amalgamation of two regimes, i.e., reactive and dissipative [4, 55]. The field-aligned current is primarily responsible for the IEDD instability in the dissipative regime. A nearby region of positive and negative wave-energy density drives the IEDD instability in the reactive domain, and in most cases no field-aligned current is needed [49, 54].

The IEDD theory was primarily suggested by Ganguli et al. [49, 50, 54] and relied on the nonlocal approximation. Koepke et al. [15] experimentally verified the IEDD instability in Q-machine at West Virginia University, and they have shown that the variation in the electric field affects the IEDD mode amplitude. Koepke et al. [36] experimentally identified the IEDD instability in the laboratory. They reported that the azimuthal and axial Doppler shift ratio has an impact on inhomogeneous energy density driven mode, i.e., frequency of the instability increases with the ratio. Ganguli [37] has reported the IEDD instability using both fluid and kinetic formalism, and a comparative study is shown between them. He depicted that as the wavevector and magnitude of the drift increase, the growth rate of the instability increases. Amatucci et al. [56] observed the reactively produced plasma waves in the ion-cyclotron frequency range in the laboratory, followed by a transverse dc electric field. Koepke et al. [38] investigated the various nonlocal eigenmodes of inhomogeneity-driven plasma instability for the very first time in sodium plasma. They observed that with the gyro-radius parameter increment, the growth rate of the instability augments initially, attains peak value and then decreases.

To the best of the authors' knowledge, no theoretical model has been proposed to study the IEDD instability in the preseence of a transverse dc electric field in a dusty magnetized collisional plasma cylinder via fluid theory. Drawing inspiration from the work described above that only focus on the experimental studies [15, 38] of the IEDD instability in the absence of dust particles, in the present present paper we propose a theoretical model to understand the role of plasma parameters, i.e., dust density, relative density ratio and gyro-radius parameter etc., on the IEDD instability in a collisional magnetized plasma cylinder having negatively charged dust particles in the existence of a transverse dc electric field. Moreover, this current model shows a comparative study for both cases, i.e., in the presence and absence of dust particles, and the impact of different plasma parameters on the frequency and the growth rate of instability is studied.

This manuscript is organized as follows for the succeeding sections: the detailed analysis of inhomogeneous energy density-driven instability using the motion and continuity equations for solving the expression of growth rate and frequency is delineated in Section 2. In Section 3, the outcomes of the theoretical modelling of instability have been compared with the results of the existing experiments. The conclusion is outlined in Section 4.

2 Instability analysis

In this manuscript, we studied the collisional multispecies magnetized dusty plasma cylinder of radius $r_{0,}$ as shown in Fig. 1.

The proposed system consists of three species: electrons, ions and dust particles, and their notations can be tabulated in Table 1.

We take into account a uniform static magnetic field B in the z-direction and a localized dc electric field $E_0(r)$ in the radial direction, such as

$$\begin{aligned} \vec{E}_0(r) &= E_0(\mathrm{r}) \text{ for } \mathrm{r} \le \hat{r}_0 \\ &= 0 \text{ for } \mathrm{r} > \mathrm{r}_0 \end{aligned} \tag{1}$$

Due to an electric field, the $\vec{E} \times \vec{B}$ drift is induced, having a magnitude $\vec{v}_E(r) = \frac{c\vec{E}_0(r) \times \vec{B}}{\vec{B}^2}$, where c is the speed of light. The localization of electric field gives rise to IEDD instability. In equilibrium, the quasi-neutrality condition says

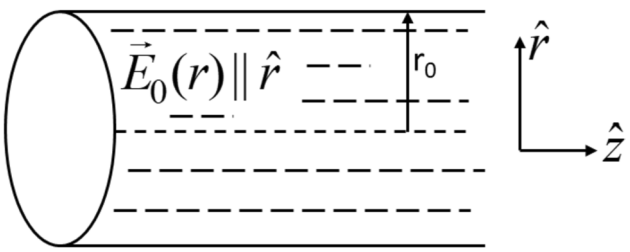


Fig. 1 Schematic of electric field model with negatively charged dust grains

Table 1 Species and their notations considered in the model

S. No	Species	Equilibrium density	Drift Velocity	Mass	Charge	Temperature	Collisional Frequency
1	Electrons	n_e^0	$v_{ed}\|\|\hat{z}$	m_e	$-e$	T_e	υ_e
2	Ions	n_i^0	0	m_i	e	T_i	0
3	Dust Particles	n_d^0	0	m_d	Q_d^0	T_d	0

$$-en_i^0 + en_e^0 + Q_d^0 n_d^0 \approx 0. \tag{2}$$

An electrostatic perturbation is given by

$$\phi_1 = \phi_0(r)\exp[-i(\omega t - \vec{k}.\vec{r})]. \tag{3}$$

The equation of motion

$$m\frac{d\vec{v}}{dt} = -e\vec{E} - \frac{e}{c}\vec{v}\times\vec{B} - \frac{\nabla(nT)}{n} - m\vec{v}v \tag{4}$$

is used to explain the behaviour of electrons in the plasma. On linearization and solving of Eq. (4), we get the perpendicular and parallel components of perturbed velocities,

$$\begin{aligned}\vec{v}_{1\perp} =& \frac{e}{m_e}\frac{[i(\omega_1+i\upsilon_e)\nabla_\perp\phi_1 + \nabla_\perp\phi_1\times\omega_{ec}]}{[(\omega_1+i\upsilon_e)^2-\omega_{ec}^2]} \\ &- \frac{T_e}{m_e n_e^0}\frac{[i(\omega_1+i\upsilon_e)\nabla_\perp n_e^1 + \nabla_\perp n_e^1\times\omega_{ec}]}{[(\omega_1+i\upsilon_e)^2-\omega_{ec}^2]}\end{aligned} \tag{5}$$

$$v_{1z} = \frac{-ek_z\phi_1}{m_e(\omega+i\upsilon_e)} + \frac{T_e}{m_e}\frac{k_z n_e^1}{(\omega+i\upsilon_e)n_e^0}, \tag{6}$$

where $\omega_1 = \omega - k_\theta v_E - k_z v_{ed}$,$\omega_{ec}\left(=\frac{eB}{m_e c}\right)$ is the electron cyclotron frequency; k_z and k_θ are components the propagation vector in z and θ- directions, respectively.

Using perturbed velocities, i.e., Eqs. (5) and (6), in the continuity equation

$$\frac{\partial n}{\partial t} + \nabla.(n\vec{v}) = 0 \tag{7}$$

The perturbed electron density can be evaluated and written as

$$n_e^1 = \frac{\frac{n_e^0 e}{m_e\omega_1}\left(\frac{(\omega_1+i\upsilon_e)\nabla_\perp^2\phi_1}{(\omega_1+i\upsilon_e)^2-\omega_{ec}^2} - \frac{k_z^2\phi_1}{(\omega_1+i\upsilon_e)}\right)}{1 - \frac{v_{et}^2k_z^2}{\omega_1(\omega_1+i\upsilon_e)} + \frac{(\omega_1+i\upsilon_e)\nabla_\perp^2 v_{et}^2}{\omega_1[(\omega_1+i\upsilon_e)^2-\omega_{ec}^2]}}, \tag{8}$$

where $v_{et} = \left(\frac{T_e}{m_e}\right)^{1/2}$ is the thermal velocity of the electrons.

In the short wavelength limit, i.e., $\omega_1 << k_\perp v_{et}$ and using $\omega_1 << \omega_{ec}$, $\omega_1 << i\upsilon_e$, Eq. (8) becomes

$$n_e^1 = \frac{n_e^0 e\phi_1}{T_e\left(1 - \frac{i\upsilon_e\omega_1}{k_z^2\frac{T_e}{m_e}}\right)}. \tag{9}$$

Similarly, the perturbed ion density is obtained in the limit $\omega_2 \sim \omega_{ic}$ and $\frac{k_\perp v_{it}}{\omega_2-\omega_{ic}} << 1$

$$n_i^1 = -\frac{en_i^0}{m_i}\left(\frac{\nabla_\perp^2\phi_1}{\omega_2^2-\omega_{ic}^2} - \frac{k_z^2\phi_1}{\omega_2^2}\right), \tag{10}$$

where $v_{it} = \left(\frac{T_i}{m_i}\right)^{1/2}$ denotes ion thermal velocity, $\omega_2 = \omega - k_\vartheta v_E$ and $\omega_{ic} = \left(\frac{eB}{m_i c}\right)$ is the ion cyclotron frequency.

Case I: In the absence of dust grains.

Substituting Eqs. (9) and (10) in Poisson's equation

$$\nabla^2\phi_1 = 4\pi e\left(n_e^1 - n_i^1\right) \tag{11}$$

we get a second-order differential equation in ϕ_1. Solving it for an axially symmetric case, we obtain

$$\frac{\partial^2\phi_1}{\partial r^2} + \frac{1}{r}\frac{\partial\phi_1}{\partial r} + p_1^2\phi_1 = 0, \tag{12}$$

or

$$\nabla_\perp^2\phi_1 + p_1^2\phi_1 = 0. \tag{13}$$

Here

$$p_1^2 = \frac{-\frac{\omega_{ep}^2}{v_{et}^2}\left(1 - \frac{i\upsilon_e\omega_1}{k_z^2 v_{et}^2}\right) - k_z^2 + \frac{\omega_{ip}^2k_z^2}{\omega_2^2}}{\left(1 - \frac{\omega_{ip}^2}{\omega_2^2-\omega_{ic}^2}\right)}, \tag{14}$$

where $\omega_{ep}\left(=\frac{4\pi n_e^0 e^2}{m_e}\right)^{1/2}$ and $\omega_{ip}\left(=\frac{4\pi n_i^0 e^2}{m_i}\right)^{1/2}$ are the electron and ion plasma frequency, respectively.

The Bessel equation's solution is given by

$$\phi_1 = A_1 J_0(p_{1n} r) p_1 \sim p_{1n}, \tag{15}$$

where $J_0(p_{1n}r)$ denotes the zero-order Bessel functions of argument $(p_{1n}r)$ and A_1 is a constant. At $r = r_0$, $\phi_1 = 0$. Hence, $\phi_1 = A_1 J_0(p_{1n} r_0) = 0$ i.e., $p_{1n} = \frac{x_n}{r_0}$ (n = 1, 2, ….) where x_n represents the Bessel function's zeroes. If $J_0(x) = 0$, $x = 2.404$ and therefore $p_{1n} = \frac{2.404}{r_0}$.

From Eq. (14) and simplifying it, we obtain

$$1 - \frac{\omega_{ip}^2}{\omega_2^2 - \omega_{ic}^2}\frac{p_{1n}^2}{p_{1n}^2 + k_z^2} + \frac{\omega_{ep}^2\left(1 + \frac{i v_e \omega_1}{k_z^2 v_{et}^2}\right)}{v_{et}^2\left(1 + \frac{v_e^2\omega_1^2}{k_z^4 v_{et}^4}\right)\left(p_{1n}^2 + k_z^2\right)} - \frac{\omega_{ip}^2}{\omega_2^2}\frac{k_z^2}{p_{1n}^2 + k_z^2} = 0. \tag{16}$$

Solving Eq. (16) and using $\frac{v_e^2\omega_1^2}{k_z^4 v_{et}^4} << 1$, we obtain

$$\varepsilon_r(\omega, k) + i\varepsilon_i(\omega, k) = 0, \tag{17}$$

where

$$\varepsilon_r = 1 - \frac{\omega_{ip}^2}{\omega_2^2 - \omega_{ic}^2}\frac{p_{1n}^2}{p_{1n}^2 + k_z^2} + \frac{\omega_{ep}^2}{v_{et}^2\left(p_{1n}^2 + k_z^2\right)} - \frac{\omega_{ip}^2}{\omega_2^2}\frac{k_z^2}{p_{1n}^2 + k_z^2}, \tag{18}$$

$$\varepsilon_i = \frac{\omega_{ep}^2 v_e \omega_1}{k_z^2 v_{et}^4\left(p_{1n}^2 + k_z^2\right)}. \tag{19}$$

Let $\omega = \omega_r + i\gamma$ be the solution of Eq. (17), assuming the wave is either growing or damped.

The real part is obtained as follows

$$\varepsilon_r(\omega, k) = 0\big|_{\omega=\omega_r} \tag{20}$$

The imaginary part is given by[57]

$$\gamma = \left.\frac{-\varepsilon_i(\omega, k)}{\frac{\partial\varepsilon_r(\omega,k)}{\partial\omega}}\right|_{\omega=\omega_r} \tag{21}$$

The real frequency can be obtained by solving Eqs. (18) and (20), we obtain

$$1 - \frac{\omega_{ip}^2}{(\omega_r - k_\theta v_E)^2}\frac{k_z^2}{p_{1n}^2}\frac{1}{\alpha} - \frac{\omega_{ip}^2}{\{(\omega_r - k_\theta v_E)^2 - \omega_{ic}^2\}}\frac{1}{\alpha} = 0, \tag{22}$$

where $\alpha \approx \frac{\omega_{ip}^2}{c_s^2}\frac{1}{p_{1n}^2}$; $c_s\left(= \frac{T_e}{m_i}\right)^{1/2}$

Multiply Eq. (22) by $\{(\omega_r - k_\theta v_E)^2\}\{(\omega_r - k_\theta v_E)^2 - \omega_{ic}^2\}$, we get

$$\omega_{2r}^4 - \omega_{2r}^2\left(\omega_{ic}^2 + \frac{\omega_{ip}^2}{\alpha} + \frac{\omega_{ip}^2 k_z^2}{p_{1n}^2\alpha}\right) + \omega_{ic}^2\frac{\omega_{ip}^2 k_z^2}{p_{1n}^2\alpha} = 0. \tag{23}$$

Equation (23) is biquadratic in ω_{2r}, where we consider $\omega_{2r} = \omega_r - k_\theta v_E$, and root is

$$\omega_r = k_\theta v_E + \sqrt{\omega_{ic}^2 + c_s^2\left(p_{1n}^2 + k_z^2\right)} \tag{24}$$

Equation (24) corresponds to the real frequency of IEDD instability with the presence of a transverse dc electric field and absence of dust particles.

Solving Eqs. (19) and (21), the imaginary part is as follows.

$$\gamma = \frac{-1}{2}\frac{\left(\frac{v_e\{1 - k_z v_{ed}(1+R)\}}{k_z^2 v_{et}^4}\frac{m_i}{m_e}\frac{1}{\delta}\right)\left[\left(\omega_r - k_\theta v_E\right)^2 - \omega_{ic}^2\right]^2}{\left(\omega_r - k_\theta v_E\right)\left(p_{1n}^2\right)}. \tag{25}$$

where $R\left(= \frac{k_\theta v_E}{k_z v_{ed}}\right)$; $\delta\left(= \frac{n_i^0}{n_e^0}\right)$.

Equation (25) corresponds to the growth rate of the IEDD instability in the presence of a transverse dc electric field without dust grains.

Case II: In the existence of dust grains.

The perturbed density behaviour for dust is written as

$$n_d^1 = \frac{n_d^0 Q_d^0 \nabla_\perp^2 \phi_1}{m_d \omega_2^2}. \tag{26}$$

Here, it is assumed that dust is unmagnetized as $m_d = 10^{12} m_p$. Therefore, the dust gyro frequency $\omega_{dc}\left(= \frac{Q_d^0 B}{m_d c}\right)$ becomes very small; hence, we can neglect it. Also, $\omega_2 >> \omega_{dc}$. For the dust particle, on the application of probe theory, the acquired charge on the dust particle Q_d is set to get balanced with the plasma currents on the surface, and defined by [58–60]

$$\frac{-dQ_d}{dt} = I_e + I_i. \tag{27}$$

Current due to electrons and ions on the surface of the particle can be understood as

$$I_e = -\pi a^2 e\left(\frac{8T_e}{\pi m_e}\right)^{1/2} n_e^0 \exp\left(\frac{e\left(\phi_g - V\right)}{T_e}\right) \tag{28}$$

and

$$I_i = \pi a^2 e\left(\frac{8T_i}{\pi m_i}\right)^{1/2} n_i^0 \exp\left(\frac{1 - e\left(\phi_g - V\right)}{T_i}\right), \tag{29}$$

where $(\phi_g - V)$ denotes the dust particle surface and plasma potential difference; "a" denotes the radius of the dust grain sphere. In equilibrium, $|I_e^0| = |I_i^0|$.

The fluctuations due to the charge on dust particles can be depicted as

$$\frac{dQ_d^1}{dt}+\eta Q_d^1=-|I_e^0|\left(\frac{n_i^1}{n_i^0}-\frac{n_e^1}{n_e^0}\right), \tag{30}$$

where $\eta=\frac{|I_e^0|e}{C_d}\left(\frac{1}{T_e}+\frac{1}{T_i-e\phi_g^0}\right)$ represents the charging rate of dust; $Q_d^1=Q_d-Q_d^0$ represents the perturbed charge on dust; $C_d=\left(a\left(1+\frac{a}{\lambda_{ed}}\right)\right)$ denotes the capacitance of a dust particle [61]; λ_{ed} represents the electron Debye length.

$$Q_d^1=\frac{e|I_e^0|}{i(\omega+i\eta)}\left(\frac{-1}{m_i}\left(\frac{\nabla_\perp^2\phi_1}{\omega_2^2-\omega_{ic}^2}-\frac{k_z^2\phi_1}{\omega_2^2}\right)-\frac{\phi_1}{T_e\left(1-\frac{iv_e\omega_1}{\frac{T_e}{k_z^2m_e}}\right)}\right). \tag{31}$$

Substituting Eqs. (9), (10), (26) and (31) in Poisson's equation

$$\nabla^2\phi_1=4\pi\left(n_e^1e-n_i^1e+n_d^0Q_d^1+Q_d^0n_d^1\right) \tag{32}$$

simplifying it and using Eq. (13), we get

$$p_1^2=\frac{-\frac{\omega_{ep}^2}{v_{et}^2}\left(1-\frac{iv_e\omega_1}{k_z^2v_{et}^2}\right)-k_z^2+\frac{\omega_{ip}^2k_z^2}{\omega_2^2}+\frac{i\beta\omega_{ip}^2}{\omega+i\eta}\frac{n_e^0}{n_i^0}\frac{k_z^2}{\omega_2^2}-\frac{i\beta\omega_{ep}^2}{\omega+i\eta}\frac{1}{v_{et}^2\left(1-\frac{iv_e\omega_1}{k_z^2v_{et}^2}\right)}}{1-\frac{\omega_{ip}^2}{\omega_2^2-\omega_{ic}^2}-\frac{i\beta\omega_{ip}^2}{\omega+i\eta}\frac{n_e^0}{n_i^0}\frac{1}{\omega_2^2-\omega_{ic}^2}}, \tag{33}$$

where $\beta=0.397\omega_{ip}^2\left(1-\frac{1}{\delta}\right)\left(\frac{a}{v_{et}}\right)\left(\frac{m_i}{m_e}\right)$ is the coupling parameter.

Using the same procedure as was done earlier, we solve Eq. (33) to get

$$\varepsilon_r=1-\frac{\omega_{ip}^2}{\omega_2^2-\omega_{ic}^2}\frac{p_{1n}^2}{p_{1n}^2+k_z^2}+\frac{\omega_{ep}^2}{v_{et}^2\left(p_{1n}^2+k_z^2\right)}-\frac{\omega_{ip}^2}{\omega_2^2}\frac{k_z^2}{p_{1n}^2+k_z^2}+\beta G, \tag{34}$$

where

$$G=\frac{\eta\omega_{ip}^2}{\omega_2^2-\omega_{ic}^2}\frac{p_{1n}^2}{p_{1n}^2+k_z^2}\frac{1}{\omega^2+\eta^2}\frac{1}{\delta}-\frac{\eta\omega_{ip}^2}{\omega_2^2}\frac{k_z^2}{p_{1n}^2+k_z^2}\frac{1}{\omega^2+\eta^2}\frac{1}{\delta}-\frac{\omega\upsilon_e\omega_1\omega_{ep}^2}{k_z^2v_{et}^4}\frac{1}{p_{1n}^2+k_z^2}\frac{1}{\omega^2+\eta^2}-\frac{\eta\omega_{ep}^2}{v_{et}^2}\frac{1}{p_{1n}^2+k_z^2}\frac{1}{\omega^2+\eta^2}.$$

$$\varepsilon_i=\frac{\omega_{ep}^2\upsilon_e\omega_1}{k_z^2v_{et}^4\left(p_{1n}^2+k_z^2\right)}\left(1-\frac{\beta\eta}{\omega^2+\eta^2}\right)+\frac{\beta\omega}{\omega^2+\eta^2}\frac{1}{\delta\left(p_{1n}^2+k_z^2\right)}\left(\frac{\omega_{ip}^2p_{1n}^2}{\omega_2^2-\omega_{ic}^2}-\frac{\omega_{ip}^2k_z^2}{\omega_2^2}\right)+\frac{\beta\omega\omega_{ep}^2}{\omega^2+\eta^2}\frac{1}{v_{et}^2\left(p_{1n}^2+k_z^2\right)}. \tag{35}$$

Following the same method as was done in Case I, the real frequency expression is given as

$$\omega_r=k_\theta v_E+\sqrt{\omega_{ic}^2+c_s^2\left(p_{1n}^2+k_z^2\right)+\frac{\beta Z_d(p_{1n}^2+k_z^2)c_s^2}{n_e^0\pi a^2v_{et}}}, \tag{36}$$

where Z_d is the dust charge state.

Equation (36) represents the real frequency of the instability in the presence of a transverse dc electric field with dust. If $\beta\to0$ in Eq. (36) we retrieve the expression of real frequency without dust, i.e., Eq. (24).

The expression for growth rate can be obtained by solving Eq. (21) and (35),

$$\gamma=\frac{-1}{2}\frac{\left(\frac{\upsilon_e\{1-k_z\upsilon_{ed}(1+R)\}}{k_z^2v_{et}^4}\frac{m_i}{m_e}\frac{1}{\delta}-\beta N\right)\left[\left(\omega_r-k_\theta v_E\right)^2-\omega_{ic}^2\right]^2}{\left(\omega_r-k_\theta v_E\right)\left(p_{1n}^2-\frac{\beta}{\eta}M\right)}, \tag{37}$$

where $N=N_1-N_2$;

$N_1=\frac{\eta\upsilon_e\{1-k_z\upsilon_{ed}(1+R)\}}{k_z^2v_{et}^4}\frac{m_i}{m_e}\frac{1}{\delta(\omega_r^2+\eta^2)}+\frac{\omega_rp_{1n}^2}{(\{\omega_r-k_\theta v_E\}^2-\omega_{ic}^2)}\frac{1}{\delta(\omega_r^2+\eta^2)}$,

$N_2=\frac{\omega_rk_z^2}{(\omega_r-k_\theta v_E)^2}\frac{1}{\delta(\omega_r^2+\eta^2)}+\frac{\omega_r}{\delta(\omega_r^2+\eta^2)v_{et}^2}\frac{m_i}{m_e}$,

$M=\frac{[(\omega_r-k_\theta v_E)^2-\omega_{ic}^2]^2}{(\omega_r^2+\eta^2)\delta}(M_1+M_2+M_3)$,where

$M_1=\eta^2p_{1n}^2[(\omega_r-k_\theta v_E)^2-\omega_{ic}^2]^2+\frac{\omega_r\eta^2p_{1n}^2}{(\omega_r^2+\eta^2)(\omega_r-k_\theta v_E)}+\frac{\eta^2k_z^2}{(\omega_r-k_\theta v_E)^4}$,

$M_2=\frac{\omega_r\eta^2k_z^2}{(\omega_r^2+\eta^2)(\omega_r-k_\theta v_E)^3}+\frac{k_z^2\eta\delta(\omega_r^2+\eta^2)}{\beta(\omega_r-k_\theta v_E)^4}+\frac{\omega_r\eta^2m_i}{(\omega_r^2+\eta^2)(\omega_r-k_\theta v_E)m_ev_{et}^2}$,

$M_3=\frac{\upsilon_e\eta m_i(2\omega_r-k_\theta v_E-k_z v_{ed})}{2k_z^2v_{et}^4m_e(\omega_r-k_\theta v_E)}+\frac{\omega_r^2\upsilon_e\eta m_i(\omega_r-k_\theta v_E-k_z v_{ed})}{k_z^2v_{et}^4m_e(\omega_r^2+\eta^2)(\omega_r-k_\theta v_E)}$.

Equation (37) represents the growth rate of IEDD instability with the inclusion of dust in the presence of a transverse dc electric field. If $\beta\to0$ in Eq. (37) we retrieve the expression of growth rate without dust, i.e., Eq. (25).

3 Results and Discussion

In this paper, an analytical model has been developed in a collisional magnetized plasma cylinder to study the IEDD instability with negatively charged dust grains. The model incorporates an electric field, using the fluid theory in the presence and absence of dust particles. The developed model aims to analyse the relationship between various plasma parameters, i.e., real frequency, growth rate etc. The effect of the electric field, magnetic field and relative density ratio on the instability is studied in both cases. The equations that hold accountable for IEDD instability in Sec. II are used to calculate the frequency and the rate of growth for the instability. The graphs have been plotted using the MATLAB programme for the analytically solved equations i.e., for the real frequency and the growth rate. The modified plasma parameters used in our calculation are obtained from Koepke et al. [15, 38] experimental papers, and are tabulated in Table 2.

Figure 2(a), (b) portray the plot of normalized real frequency and growth rate in the absence of dust grains as a function of the gyro-radius parameter. From Fig. 2(a), it can

Table 2 Plasma parameters specified in this model

Terms	Values
e (Electronic charge)	$4.8 \times 10^{-10} statcoul$
Q_d^0(Dust charge)	$10^3 - 10^4 e$
m_d(Dust mass)	$10^{12} \times 1.67 \times 10^{-24} gm$
m_e(Electron mass)	$9.1 \times 10^{-28} gm$
m_i(Ion mass)	$22 \times 1.67 \times 10^{-24} gm$
T_e(Electron temperature)	$0.2\, eV$
T_i(Ion temperature)	$0.2\, eV$
n_i^0(Initial ion density)	$10^9 cm^{-3}$
n_e^0(Initial electron density)	$\left(1 \times 10^9 - 0.2 \times 10^9\right) cm^{-3}$
n_d^0(Dust density)	$(1-5) \times 10^4 cm^{-3}$
c(Speed of light)	$3 \times 10^{10} cm/\mathrm{s}$
B(Magnetic field)	$(0.1 - 4.5) kG$
$\delta\left(= {}^{n_i}/_{n_e}\right)$(Relative density ratio)	$1 - 10$
$b_i (= k_\theta \rho_i)^2$(Gyro-radius parameter)	$0 - 0.9$
K(Boltzmann constant)	$1.38 \times 10^{-16} erg/Kel$
k_z	$0.65\, cm^{-1}$
k_θ	$2.8\, cm^{-1}$
υ_e(Electron collisional frequency)	$3.22 \times 10^8 \sec^{-1}$
v_{ed}(Electron drift velocity)	$1.76 \times 10^4 cm/\sec$
v_E(Electric field)	$3.3 \times 10^8 / B(Gauss)$
r_0(Plasma radius)	$3\, cm$
a (Dust grain size)	$(1-5) \times 10^{-4} cm$

be inferred that with the increase in the gyro-radius parameter, the frequency of the instability increases. From Fig. 2(b), it can be observed from the graph that with an increase in values of the gyro-radius parameter, the rate of growth of the instability first augments and then attains a maximum value, i.e., $\gamma = 3.2 \times 10^5 \sec^{-1}$ for $b_i = 0.08$. After attaining maximum value, the instability's growth rate decreases. Moreover, it can also be observed that the normalized growth rate of instability decreases by a factor of ~ 3.8 when the gyroradius changes from 0.2 to 0.6 whereas in the experimental observation of Koepke et al. [38], it upsurges by a factor of ~ 3. The results obtained are in good agreement with this experimental observation.

In the presence of dust grains, the normalized real frequency and growth rate variation corresponding to the gyro-radius parameter have been plotted in Fig. 3(a),(b) for different sets of relative density ratios, respectively. Figure 3(a) represents the normalized real frequency plot versus gyro-radius parameter for various relative density ratio i.e., (i) $\delta = 8$, (ii) $\delta = 6$, (iii) $\delta = 4$, and (iv) $\delta = 2$.With the increase in the gyro-radius parameter, the instability frequency augments. Moreover, with the rise in relative density ratio, the frequency also augments. Figure 3(b) represents the normalized growth rate plot versus the gyro-radius parameter for various relative density ratio, i.e., (i) $\delta = 8$, (ii) $\delta = 6$, (iii) $\delta = 4$, and (iv) $\delta = 2$. From the plot, it can be noted that the

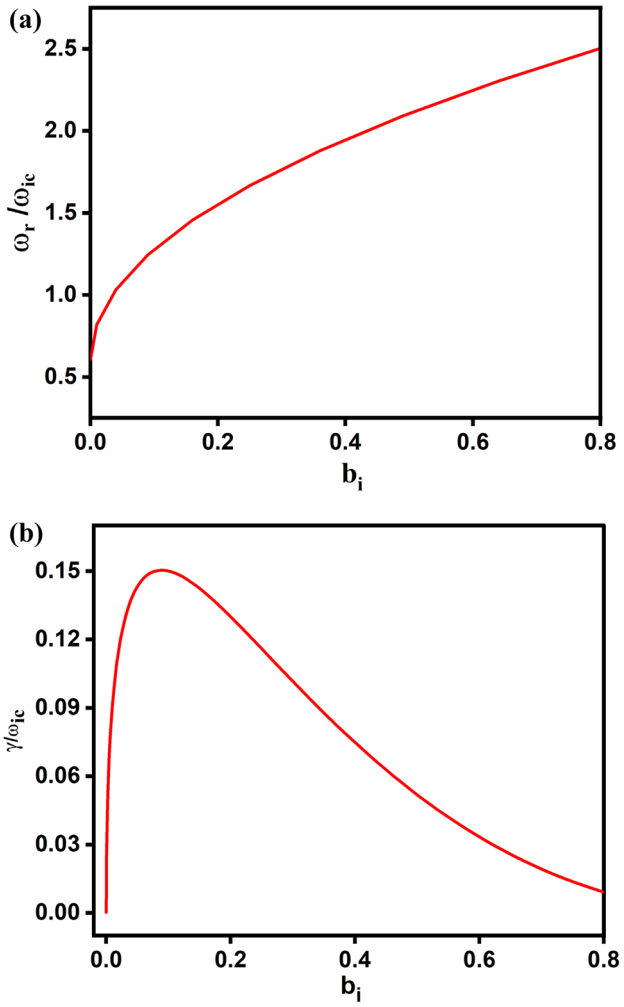


Fig. 2 **a** Variation in normalized real frequency of the IEDD instability ω_r/ω_{ic} with the finite gyro-radius parameter $b_i (= k_\theta \rho_i)^2$, in the absence of dust grains. **b** Variation in normalized growth rate of the IEDD instability γ/ω_{ic} with the finite gyro-radius parameter $b_i (= k_\theta \rho_i)^2$, in the absence of dust grains

growth rate values rise with increasing gyro-radius parameter until they reach their peak value, after which there is a decline in magnitude for all values of the relative density ratio. Hence, we can say that there is a significant decrease in the growth rate by a factor of ~ 6.5 as dust grains are included for $\delta = 4.0$, as compared with the growth rate without the dust grains in the above-mentioned graph. In addition, with the rise in the relative density ratio, the instability growth rate augments and the decline in magnitude becomes less sharp. We can conclude that after the inclusion of dust particles, both real frequency and growth rate follows the similar trend as in the case of absence of the dust, in spite of slight modifications in their values.

In the existence and non-existence of a dc electric field, Fig. 4(a), (b) illustrate the impact of the relative density ratio of ions and electrons on the normalized real frequency and

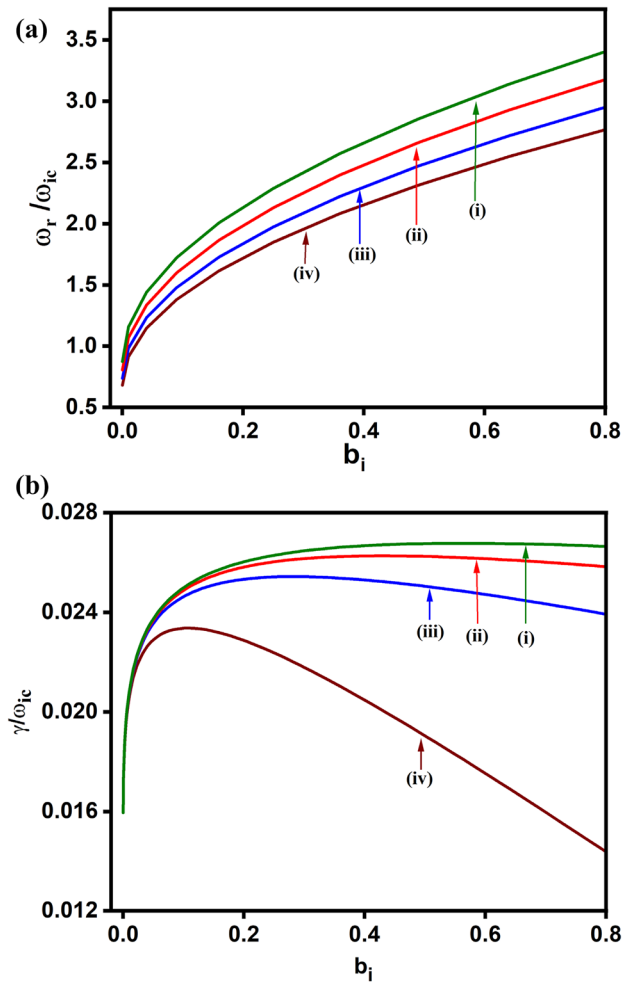


Fig. 3 **a** The normalized real frequency ω_r/ω_{ic} variation for different values of $\delta(=n_i^0/n_e^0)$, i.e., (i) $\delta = 8$, (ii) $\delta = 6$, (iii) $\delta = 4$, and (iv) $\delta = 2$ corresponding to the finite gyro-radius parameter $b_i(=k_\theta\rho_i)^2$ in the presence of dust. **b** The normalized growth rate γ/ω_{ic} variation for different values of $\delta(=n_i^0/n_e^0)$, i.e., (i)$\delta = 8$, (ii) $\delta = 6$, (iii) $\delta = 4$, and (iv) $\delta = 2$ in the existence of dust corresponding to the finite gyro-radius parameter $b_i(=k_\theta\rho_i)^2$

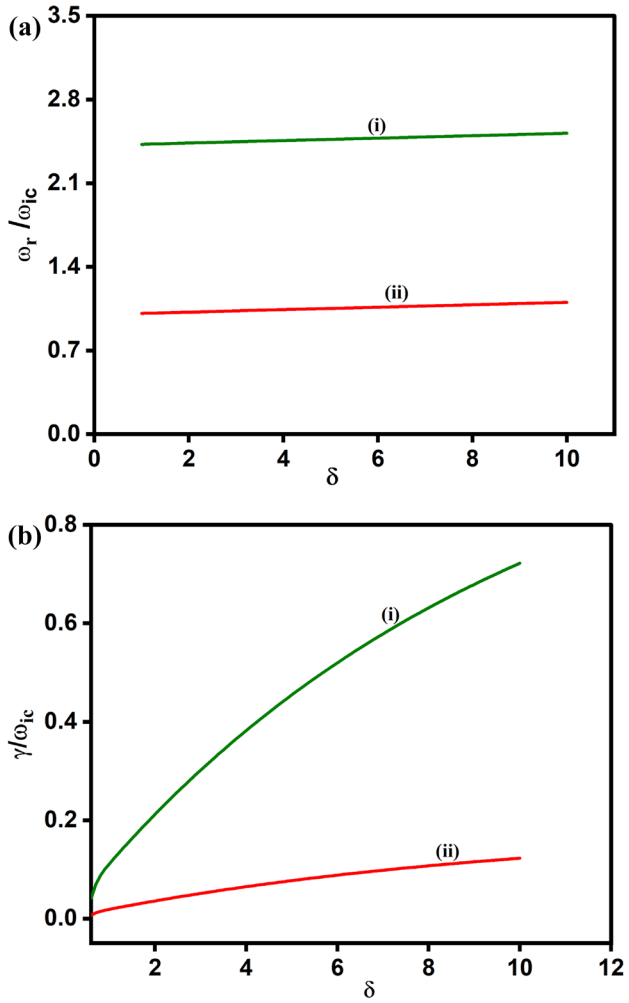


Fig. 4 **a** The normalized real frequency ω_r/ω_{ic} of IEDD instability as a function of relative density ratio $\delta(=n_i^0/n_e^0)$ (i) in the electric field presence and (ii) in the electric field absence. **b** Variation in the normalized growth rate γ/ω_{ic} of IEDD instability with relative density ratio $\delta(=n_i^0/n_e^0)$ in the electric field (i) presence and (ii) absence

growth rate of the relative density ratio of the instability. In the existence of dust charge variations, we observed that the normalized real frequency and growth rate of IEDD instability is in a linear relationship with respect to the $\delta(=n_i^0/n_e^0)$, and it can be verified from Eqs. (36) and (37). As the negatively charged dust grains augment in plasma, the real frequency and the growth rate of instability increase. It might be because as δ increases, the plasma density of electrons decreases relative to the plasma density of ions. For increasing values of δ, an ion's effective mass $m_{ieff} = \frac{m_i}{\delta}$ is smaller than m_i, and because of its higher mobility, the frequency and growth rate of the instability gets enhanced. It can also be compared from the graphs that the addition of an electric field enhanced the instability's real frequency and growth rate. The real frequency maximum value is measured at $\omega_r = 1.64 \times 10^6 rad./\sec$ when there is an electric field involved, and it drops to $\omega_r = 7.2 \times 10^5 rad./\sec$ as $v_E \to 0$ in Eq. (36). The growth rate maximum value is estimated at $\gamma = 3.5 \times 10^5\,\sec^{-1}$ when there is an electric field involved, and it drops to $\gamma = 6.0 \times 10^4\,\sec^{-1}$ as $v_E \to 0$ in Eq. (37). Therefore, we can infer that the frequency and rate of growth have slightly changed as an outcome of the addition of the electric field. This is due to the fact that the addition of an electric field generates $\vec{E} \times \vec{B}$ drift, which results in the Doppler effect, and as a result, the instability's real frequency augments.

Figure 5(a) represents the growth rate variation with k_z/k_θ for different values of v_E/v_{it} in the absence of dust. On increasing the values of k_z/k_θ, the growth rate initially increases significantly; after this initial phase, there is only a tiny bitsy change in the growth rate of IEDD instability. Furthermore, with the increasing values of v_E/v_{it}, the growth

rate augments. The trend of growth rate with k_z/k_θ follows the same trend as observed in experimental results of Ganguli [37] (cf. Figure 5). Figure 5(b) shows the growth rate variation with the k_z/k_θ for different values of v_E/v_{it} in the presence of dust using Eq. (37). With increasing k_z/k_θ values, the growth rate of IEDD instability follows the same increasing trend similar to the case of without dust. But the value of growth rate degrades after the introduction of dust grains when compared to the value of growth rate without dust grains. Furthermore, there is also rise in the value of growth rate with the increasing values of v_E/v_{it}.

We have also portrayed the growth rate as a function of v_E/v_{it}, in Fig. 6, in the occurrence and non-occurrence of dust particles. As we increase v_E/v_{it}, the growth rate of the instability increases in both the cases. In the presence of dust charge variations, the rate of increment of instability's growth rate is small, while the rate of increment of growth rate is larger in the absence of dust. It can be concluded from this graph that the absence of dust charge fluctuations enhances the growth rate of IEDD instability. The growth rate graph displays a trajectory that is consistent with the observed trend described in the experimental investigation done by Ganguli et al. [37] (cf. Figure 3) in the absence of dust grains.

Figure 7 portrays the variation of normalized frequency with respect to v_E/v_{it} in the absence of dust grain particles. It can be noticed that with the increasing magnitude of radial electric field strength, the IEDD instability growth rate increases. This is ascribed to the fact that if $\vec{E} \times \vec{B}$ drift is there, then the effects of Doppler shift may result in IEDD instability to grow. Our result is in good agreement with the experimental observation of Ganguli [37] (cf. Figure 6).

Figure 8 shows the variation of the normalized growth rate with respect to the dust grain size, for varying values of the relative density ratio. The graph shows that, as predicted by Eq. (37), the size of the dust grains has an inverse relationship with the normalized growth rate. This is because the intergrain distance decreases with the increment in dust grain size, which in turn increases the screening of one dust particle charge over another, further lowering the dust grain potential. Because of this, the average dust grain charge Q_d^0 starts to fall, resulting in the decline of the instability's growth rate. It is also possible to deduce from Eqs. (26) to (37) how the dust particles affect the growth rate of the instability. Two parameters are involved in these equations: one is η, the dust charging rate, and the other one is β, the coupling parameter, or the interaction between the plasma and dust particles. Furthermore, with increasing values of δ, the instability rate of growth augments.

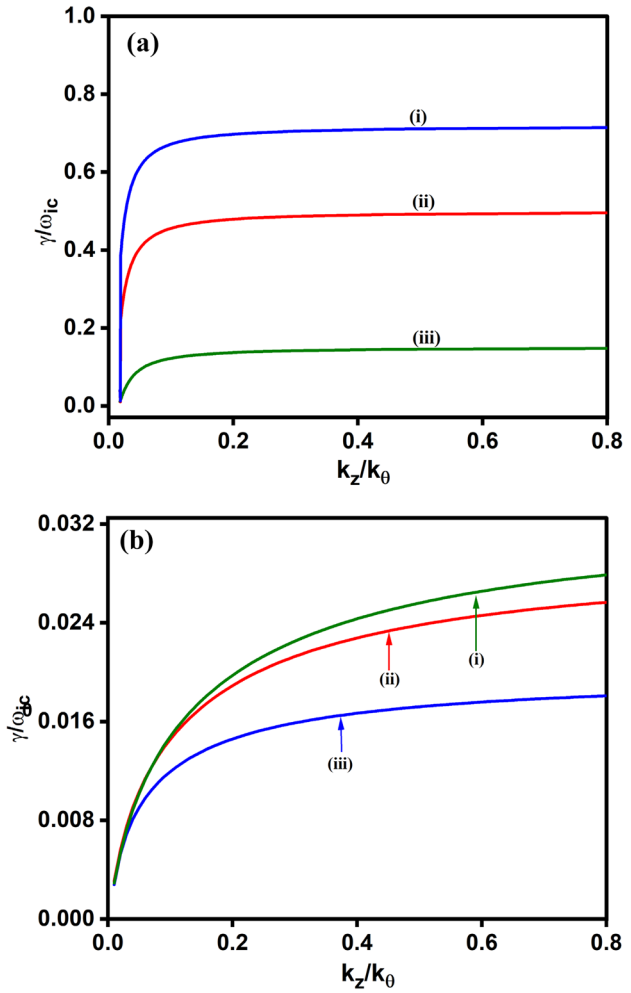


Fig. 5 **a** The normalized growth rate γ/ω_{ic} versus normalized wave vector k_z/k_θ of IEDD instability for varying values of the electric field v_E/v_{it}, i.e., (i) $v_E/v_{it}=5.5$ (ii) $v_E/v_{it}=3.5$ and (iii) $v_E/v_{it}=2$ in the absence of dust. **b** The normalized growth rate γ/ω_{ic} as a function of the normalized wave vector k_z/k_θ of IEDD instability for different values of the electric field v_E/v_{it}, i.e., (i) $v_E/v_{it}=5.5$ (ii) $v_E/v_{it}=3.5$ and (iii) $v_E/v_{it}=2$ in the presence of dust

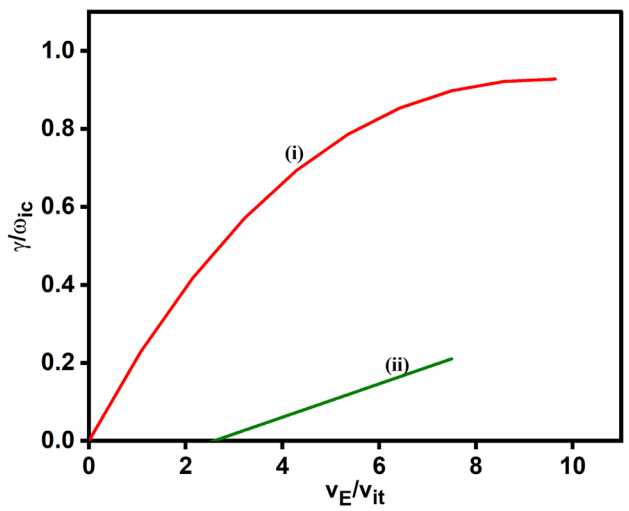


Fig. 6 The normalized growth rate γ/ω_{ic} of IEDD instability with respect to transverse dc electric field v_E/v_{it} for (i) existence of dust grains, and (ii) non-existence of dust grains

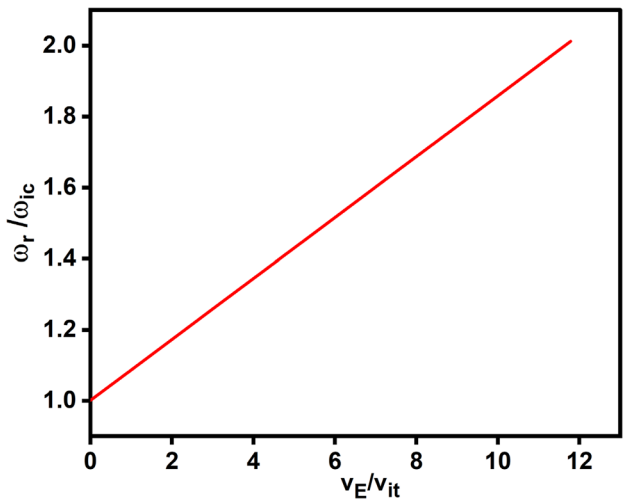


Fig. 7 The normalized real frequency ω_r/ω_{ic} of IEDD instability versus v_E/v_{it} in the absence of dust

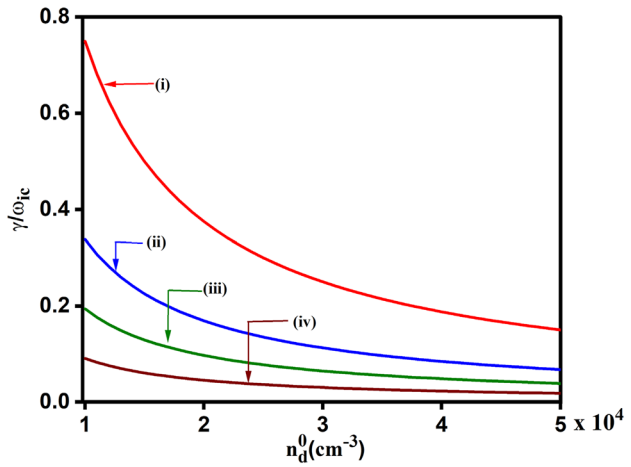


Fig. 9 The normalized growth rate γ/ω_{ic} variation for different values of dust grain size $a(cm)$, i.e., (i) $a = 2 \times 10^{-4} cm$, (ii) $a = 3 \times 10^{-4} cm$, (iii) $a = 4 \times 10^{-4} cm$, and (iv) $a = 6 \times 10^{-4} cm$ with respect to the dust grains $n_d^0(cm^{-3})$ number density

The normalized growth rate graph versus the dust number density for varying values of dust size is illustrated in Fig. 9. It is found that as the dust particle number density rises, the rate of growth of the instability decreases. This is because the number of available electrons per dust grain reduces as the quantity of dust particles rises, revealing that the total demand for electrons among the dust particles is enormous. Reduced average dust grain charge Q_d^0 causes a drop in the growth rate of the instability. In addition to this, from Fig. 9, we can depict that with the increasing values of the size of the dust grains, the growth rate reduces, as explained by Fig. 8. For dust density greater than $3 \times 10^4 cm^{-3}$, the growth rate is not majorly influenced by the number density of the dust grains.

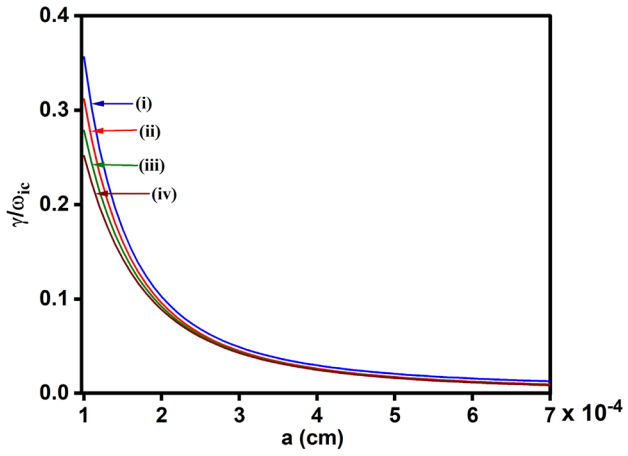


Fig. 8 The normalized growth rate γ/ω_{ic} of IEDD instability variation for different values of $\delta\left(=n_i^0/n_e^0\right)$, i.e., (i) $\delta = 8$, (ii) $\delta = 6$, (iii) $\delta = 4$, and (iv) $\delta = 2$ corresponding to the dust grain size $a(cm)$

There are currently no theoretical findings in the literature that take into account how dust particles affect IEDD instability. As a result, our theoretical findings are crucial for comprehending and illuminating the significance of the number density and dust grain size of the dust particles in the IEDD instability.

4 Conclusion

The IEDD instability in the collisional magnetized plasma cylinder under the outcome of electric field has been observed. Study of absence and presence of dust grains has been studied for various plasma parameters like gyro-radius, relative density ratio etc. It was concluded that the inclusion of dust particles has altered the behaviour of IEDD instability i.e., with the increasing values of dust grain size and dust density, the growth rate of instability decreases. The gyro-radius augments with an increase in relative density ratio. Moreover, the instability's real frequency and growth rate also augment with an augment in the relative density ratio and wave vector. The impact of the gyro-radius parameter on the frequency and the growth rate of instability was also examined. It was discovered that the frequency grows as the gyro-radius parameter augments, but the rate of growth first rises as the gyroradius parameter augments, before starting to fall as the gyroradius parameter further augments. The outcome of an electric field on the frequency and rate of growth is also considered, and it was found that with the addition of an electric field, frequency and rate of growth both rise. The acquired results were compared to the already available experimental observations [15, 37] and a good agreement was found.

The outcomes of our analytical model can be used to explain a wide range of experiments including AMICIST [62], SCIFER [63], and ALEXIS [64] in Earth's ionosphere, and might be beneficial for other space applications [18]. For ionospheric issues connected to broadband electrostatic wave measurements, the use of the IEDD instability mechanism is of special relevance [51]. In the given plasma model, non-linearities are not taken into account, which can play an important role in changing the plasma behaviour. Nonlinear effects can indeed play a crucial role in the generation and propagation of waves, energy transfer etc., of these instabilities. Therefore, it is crucial to investigate and characterise these nonlinear effects in theoretical and experimental studies. So, the same plasma model can further be explored after including these effects which will act as source for fusion reactors and plasma processing tools and will surely help our research move forward.

Author Contribution Statement Miss Anshu: Conceptualization, Methodology, Investigation, Analysis, Writing—Original Draft. Dr. Jyotsna Sharma: Visualization, Revising the manuscript critically for important intellectual content. Prof. Suresh C. Sharma: Conceptualization, Methodology, Writing – Review and Editing at different stages, Supervision.

Data Availability Statement The findings of this research are supported by the data that are presented in the manuscript. Additionally, the corresponding author will disclose the data supporting the manuscript's conclusions upon reasonable request.

Declarations

Conflict of Interests The authors have no conflicts to disclose.

References

1. N.J.P. D'Angelo, S. Science **38**, 1143 (1990)
2. P.K. Shukla, Phys. Plasmas **8**, 1791 (2001)
3. P.K. Shukla, A.A. Mamun, Introduction to dusty plasma physics, CRC press (2015)
4. V. Gavrishchaka, M.E. Koepke, G. Ganguli, Phys. Plasmas **3**, 3091 (1996)
5. F. Schmidt, F. Muecksch, Y. Weisblum, J. Da Silva, E. Bednarski, A. Cho, Z. Wang, C. Gaebler, M. Caskey, M.C. Nussenzweig, N. Engl. J. Med. **386**, 599 (2022)
6. M.E. Koepke, Phys. Plasmas **9**, 2420 (2002)
7. M.E. Koepke, W.E. Amatucci, IEEE Trans. Plasma Sci. **20**, 631 (1992)
8. S.C. Sharma, M. Sugawa, V.K. Jain, Phys. Plasmas **7**, 457 (2000)
9. J. Sharma, S.C. Sharma, V.K. Jain, A. Gahlot, Phys. Plasmas **20**, 033706 (2013)
10. M.E. Koepke, J.J. Carroll Iii, M.W. Zintl, Phys. Plasmas **5**, 1671 (1998)
11. W.E. Amatucci, D.N. Walker, G. Ganguli, D. Duncan, J.A. Antoniades, J.H. Bowles, V. Gavrishchaka, M.E. Koepke, J. Geophys. Res. Space Phys. **103**, 11711 (1998)
12. J. Sharma, S.C. Sharma, Phys. Plasmas **17**, 123701 (2010)
13. K.R. Segwal, S.C. Sharma, IEEE Trans. Plasma Sci. **46**, 775 (2018)
14. S. Seiler, M. Yamada, H. Ikezi, Phys. Rev. Lett. **37**, 700 (1976)
15. M.E. Koepke, W.E. Amatucci, J.J. Carroll Iii, T.E. Sheridan, Phys. Rev. Lett. **72**, 3355 (1994)
16. A.M. DuBois, E. Thomas Jr., W.E. Amatucci, G. Ganguli, Phys. Rev. Lett. **111**, 145002 (2013)
17. M.E. Koepke, J.J. Carroll Iii, M.W. Zintl, J. Geophys. Res. Space Phys. **104**, 14397 (1999)
18. A.A. Chernyshov, A. Spicher, A.A. Ilyasov, W.J. Miloch, L.B.N. Clausen, Y. Saito, Y. Jin, J.I. Moen, Phys. Plasmas **25**, 042902 (2018)
19. D.A. Mendis, M. Rosenberg, Ann. Rev. Astron. Astrophys. **32**, 419 (1994)
20. J. Glanz, Science **264**, 28 (1994)
21. R.L. Merlino, J.A. Goree, Phys. Today **57**, 32 (2004)
22. R.L. Merlino, Plasma Phys. Appl. **81**, 73 (2006)
23. J. Winter, Phys. Plasmas **7**, 3862 (2000)
24. S.H. Kim, J.R. Heinrich, R.L. Merlino, Planet. Space Sci. **56**, 1552 (2008)
25. A. Barkan, N. D'Angelo, R.L. Merlino, Planet. Space Sci. **43**, 905 (1995)
26. R.L. Merlino, A. Barkan, C. Thompson, N. D'Angelo, Phys. Plasmas **5**, 1607 (1998)
27. R.L. Merlino, A. Barkan, C. Thompson, N. D'Angelo, Plasma Phys. Controlled Fusion **39**, A421 (1997)
28. J. Sharma, S.C. Sharma, Contributions to Plasma Phys. e202200073, (2022)
29. K.R. Segwal, S.C. Sharma, IEEE Trans. Plasma Sci. **46**, 797 (2018)
30. S.C. Sharma, M. Sugawa, Phys. Plasmas **6**, 444 (1999)
31. J. Sharma, S. Singh, A. Kumar, M.A. Hossain, S.H. Ahammad, A.N.Z. Rashed, Braz. J. Phys. **53**, 128 (2023)
32. L. Chen, Waves and instabilities in plasmas, World Sci. (1987)
33. J. Tyagi, S. Singh, H.K. Malik, J.Theor. Appl. Phys. **12**, 39 (2018)
34. H.K. Malik, J. Tyagi, D. Sharma, AIP Adv. **9**, 055220 (2019)
35. J. Tyagi, D. Sharma, H.K. Malik, J. Theor. Appl. Phys. **12**, 227 (2018)
36. M.E. Koepke, W.E. Amatucci, J.J. Carroll Iii, V. Gavrishchaka, G. Ganguli, Phys. Plasmas **2**, 2523, (1995)
37. G. Ganguli, Phys. Plasmas **4**, 1544 (1997)
38. M.E. Koepke, J.J. Carroll Iii, M.W. Zintl, C.A. Selcher, V. Gavrishchaka, Phys. Rev. Lett. **80**, 1441 (1998)
39. J.J. Carroll Iii, Experimental investigation of the IEDD instability, West Virginia University (1997)
40. S.C. Sharma, K. Sharma, A. Gahlot, Phys. Plasmas **20**, 053704 (2013)
41. B.P. Pandey, S.V. Vladimirov, A. Samarian, Phys. Plasmas **16**, 113708 (2009)
42. A.A. Galeev, M.N. Rosenbluth, R.Z. Sagdeev, R.N. Sudan, Handbook of plasma physics, North-Holland (1983)
43. A.A. Ilyasov, A.A. Chernyshov, M.M. Mogilevsky, I.V. Golovchanskaya, B.V. Kozelov, Phys. Plasmas **22**, 032906 (2015)
44. Y. Liu, J. Lei, M. Li, Y. Ling, J. Yuan, Phys. Plasmas **25**, 102901 (2018)
45. Munish, R. Dhawan, R. Kumar, H.K. Malik, J. Taibah Univ. Sci. **16**, 725 (2022)
46. M. Rosenberg, P.K. Shukla, Planet. Space Sci. **51**, 1 (2003)
47. R. Dhawan, D. Sharma, H.K. Malik, J. Theor. Appl. Phys. **16**, 1 (2022)
48. W.E. Amatucci, M.E. Koepke, J.J. Carroll Iii, T.E. Sheridan, Geophys. Res. Lett. **21**, 1595, (1994)
49. G. Ganguli, Y.C. Lee, P.J. Palmadesso, The Physics of Fluids **31**, 823 (1988)
50. G. Ganguli, P. Palmadesso, Y.C. Lee, Geophys. Res. Lett. **12**, 643 (1985)
51. G. Ganguli, M.J. Keskinen, H. Romero, R. Heelis, T. Moore, C. Pollock, J. Geophys. Res. Space Phys. **99**, 8873 (1994)
52. K.I. Nishikawa, G. Ganguli, Y.C. Lee, P.J. Palmadesso, J. Geophys. Res. Space Phys. **95**, 1029 (1990)
53. T. Chang, G.B. Crew, J.R. Jasperse, Physics of Space Plasmas (1987): Proceedings of the 1987 Cambridge Workshop in Geoplasma Physics, Ionosphere-Magnetosphere-Solar Wind Coupling Processes, Scientific Publishers (1988)

54. G. Ganguli, Y.C. Lee, P. Palmadesso, The Physics of Fluids **28**, 761 (1985)
55. V.V. Gavrishchaka, M.E. Koepke, G.I. Ganguli, J. Geophys. Res. Space Physics **102**, 11653 (1997)
56. W.E. Amatucci, D.N. Walker, G. Ganguli, J.A. Antoniades, D. Duncan, J.H. Bowles, V. Gavrishchaka, M.E. Koepke, Phys. Rev. Lett. **77**, 1978 (1996)
57. N.A. Krall, A.W. Trivelpiece, Am. J. Phys. **41**, 1380 (1973)
58. E.C. Whipple, T.G. Northrop, D.A. Mendis, J. Geophys. Res. Space Phys. **90**, 7405 (1985)
59. M.R. Jana, A. Sen, P.K. Kaw, Phys. Rev. E **48**, 3930 (1993)
60. R.K. Varma, P.K. Shukla, V. Krishan, Phys. Rev. E **47**, 3612 (1993)
61. M.S. Barnes, J.H. Keller, J.C. Forster, J.A. O'Neill, D.K. Coultas, Phys. Rev. Lett. **68**, 313 (1992)
62. J. Bonnell, P. Kintner, J.E. Wahlund, K. Lynch, R. Arnoldy, Geophys. Res. Lett. **23**, 3297 (1996)
63. P.M. Kintner, J. Bonnell, R. Arnoldy, K. Lynch, C. Pollock, T. Moore, Geophys. Res. Lett. **23**, 1873 (1996)
64. A.M. DuBois, E. Thomas Jr., W.E. Amatucci, G. Ganguli, J. Geophys. Res. Space Phys. **119**, 5624 (2014)

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.

Civil Engineering and Architecture 11(5): 2719-2730, 2023 http://www.hrpub.org
DOI: 10.13189/cea.2023.110535

Application of Gene Expression Programming in Computation of Flow Resistance in Compound Channel with Converging Floodplains

Rahul Kumar*, Vijay Kaushik, Munendra Kumar

Department of Civil Engineering, Delhi Technological University, India

Received February 23, 2023; Revised May 25, 2023; Accepted June 11, 2023

Cite This Paper in the Following Citation Styles

(a): *[1] Rahul Kumar, Vijay Kaushik, Munendra Kumar , "Application of Gene Expression Programming in Computation of Flow Resistance in Compound Channel with Converging Floodplains," Civil Engineering and Architecture, Vol. 11, No. 5, pp. 2719 - 2730, 2023. DOI: 10.13189/cea.2023.110535.*

(b): *Rahul Kumar, Vijay Kaushik, Munendra Kumar (2023). Application of Gene Expression Programming in Computation of Flow Resistance in Compound Channel with Converging Floodplains. Civil Engineering and Architecture, 11(5), 2719 - 2730. DOI: 10.13189/cea.2023.110535.*

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 It is essential to the resolution of a large number of pressing engineering issues to have an accurate estimate of the flow resistance in an open channel flow. When there is an overbank flow on both sides of the river, it breaks through the main channel and pours out into the floodplain. The flow structure in such compound channels may become rather convoluted because of the transfer of momentum that occurs between the principal channel and the floodplains. This has a significant bearing on the flow resistance in the different subsections of the floodplain and the main channel. In addition, activities such as agriculture and construction have been carried out in the floodplain areas of a river system. Because of this, the geometry of the floodplain changes over the length of the flow, which ultimately results in the formation of a converging, diverging or skewed compound channel. Conventional formulae, which are too reliant on empirical methods, are not successful in predicting flow resistance with a high degree of accuracy. As a direct consequence of this, there is a continued high need for methods that are both original and exact. The purpose of this study is to use Gene Expression Programming to make a prediction about the manning's roughness coefficient in a compound channel with converging floodplains. The prediction will be made in terms of the geometric factors as well as the flow variables. Statistical indices are utilized to verify the created models for the experimental study so that the performance and effectiveness of these models can be evaluated. The results show that the GEP-derived manning's roughness coefficient has a strong connection both with the data from experiments and with the results of previous investigations. The GEP model that was developed for the purpose of analyzing compound channel flow has been shown to be credible by the findings of statistically sound research, and as a consequence, it is applicable to this field of study.

Keywords Compound Channel, Converging Floodplains, Manning's Roughness Coefficient, Geometric and Flow Parameters, Statistical Analysis, Gene Expression Programming

1. Introduction

The prediction of an accurate discharge in flooded rivers is one of the most significant challenges in river engineering. This challenge arises from the fact that the geometry and hydraulic parameters of a river may change over time. The quantification of the impediment posed by channels and floodplains to the flow of floodwaters is accomplished by means of the roughness coefficient. In 1D flow analysis, it is common practice to select an appropriate roughness coefficient value to evaluate a channel's conveyance capacity. This particular value of the roughness coefficient is assumed to be constant over the whole of the flow surface and at each and every depth. If

the resistance coefficient is not right, the discharge may be underestimated or overestimated [1]. There have been many studies conducted on straight compound channels, and the difficulties that need to be solved have been approached in a scientific way in order to build equations that can forecast discharge [1-3]. The results of experimental experiments made it abundantly evident that the development of a discontinuity in the overall roughness throughout the bank-full level occurs, and this discontinuity seems to be responsible for stage-discharge discontinuity in some circumstances [4]. As per Pang [5], the Manning's roughness coefficient serves the dual purpose of characterizing the topographical attributes of a channel and influencing the dissipation of energy by the flow. As compared with the structure of the flow in straight channels, the structure of the flow in nonprismatic compound channels is unexpectedly more complicated [6-10]. According to James and Brown [11], the total flow resistance was observed to rise with the skew angle. Studies were conducted on a two-stage channel featuring narrowing floodplains. The findings indicate a reduction in the depth of flow in the converging section from upstream to downstream [6, 9, 12]. Yonesi et al. [10] studied various roughness ratios for a diverging compound channel and noted that as the roughness ratio increased, the water depth along the diverging length decreased. This was seen for a variety of roughness ratios for a diverging compound channel. Despite the fact that a large number of investigations on Manning's n for compound channels have been conducted, it has not been possible to arrive at a solid conclusion about the roughness values for two-stage channels with convergent floodplains [12-15]. The simplest empirical formulae for the estimation of flow resistance are designed for prismatic channels assuming that the flow is uniform and constant. Because of the difference in geometric and hydraulic factors, they are fundamentally erroneous when applied to channels that have converging floodplains. The water depth drops from section to section in a compound channel that has converging floodplains, which results in a flow pattern that is not uniform for a given discharge in a nonprismatic compound. Accurate measurement of flow necessitates an understanding of the contribution of flow in both the primary channel and the zone of flooding. Because of the section-to-section variance that occurs in channels of this kind, the flow changes in the subsections, which in turn causes the roughness to vary throughout the cross-section of the channel. Experiments should be performed in order to calculate the roughness coefficient; however, this is not always possible in real-world situations. It is strongly recommended that this be done. A number of researchers have measured the flow resistance [16-17]. The purpose of these experiments was to study how the redistribution of flow resistance is affected by the width proportions of the main channel to the floodplain. Shiono and Knight [18] investigated the effects that smooth compound channels had on the Darcy-Weisbach frictional factor. Bjerklie et al. [19] formulated several theoretical frameworks employing resistance equations to ascertain river discharge and flow resistance. The study by Yang et al. [14] investigates a variety of factors that influence the resistance coefficients. Nevertheless, the focus of their research was mostly on the resistance to flow that was present in composite channels with prismatic floodplains. Furthermore, in flood scenarios characterized by channel sections exhibiting a variable width ratio, commonly referred to as converging instances, it is common to observe lateral variation in roughness throughout the wetted perimeter in both laboratory experiments and design endeavours. In some circumstances, it is necessary to be aware of the discharge that passes through a cross-section that has a certain angle. As a result, it is essential to develop a model in order to make an estimate of the composite's roughness. It is advised to split the cross-section into subsections in a vertical, horizontal, diagonal, or bisectional way whenever multiple techniques are applied to determine the composite roughness of compound channels at that point. As compared to other ways of division, the vertical division approach yields much fewer errors [14]. Das and Khatua [12] accomplished research to explore the impedance characteristics of overbank flow in nonprismatic sections. The trials were carried out in an oriented composite channel with expanding and narrowing floodplains, and the relative depths of flow were varied. The Manning roughness coefficient is determined by non-dimensional flow and geometric characteristics. In order to get an estimate of the roughness coefficient using Manning's equation for two-stage channels with nonprismatic floodplains, a multivariate model has been devised. The current model produces results that are adequate in comparison to those obtained via previous methods since it produces fewer errors. In previous research, the prediction of flow resistance was determined mostly by the use of multivariate analysis. In light of the fact that the discharge limitations of the earlier investigations were predetermined, fresh tests are being carried out in order to investigate how the flow resistance varies at greater flow rates. In this study, the researchers employed a new technique called gene expression programming to predict the resistance of flow in a composite channel with narrowing floodplains based on flow and geometric factors. The results of the study are shown below. Using statistical analysis, a comparison was made between the proposed model and previously existing approaches. This allowed for an evaluation of the usefulness of the model.

2. Materials and Methods

2.1. Data Source

The present study utilizes data from prior investigations conducted by Rezaei [20], Naik and Khatua [21], in addition to the author's own experimental findings. Table 1

contains the information that can be discovered on the experimental channel diameters of these compound channels that have a range of geometric characteristics.

2.2. Experimental Setup

At Delhi Technological University's Department of Civil Engineering's Hydraulic laboratory, all of the tests were carried out in masonry flume to gather data. The masonry flume was 12m in length, 1m in width, and 0.8m in depth. This flume has a compound cross-section that was built out of brick masonry. The main channel is 0.5m in width and 0.25m in depth. Brick masonry was used to assist in the construction of the segment of the channel that converges. The converging angle of the channel, which was 4 degrees, ensured that the geometry remained unchanged. The length of the converging section of the nonprismatic compound channel is 3.6m. Under certain situations of the two-stage channel, a subcritical flow regime was achieved with a longitudinal bed slope of 0.001. The Manning's n value was determined through the analysis of data obtained from both in and over bank flows that took place in the primary channel and flood zones. A water reclamation mechanism was utilized to convey water from a subterranean sump to a central reservoir situated within the experimental channel. The aqueous solution originating from the channel was channeled into a container with a volumetric capacity, featuring a V-shaped notch, and subsequently guided toward a subterranean reservoir. The geometric attributes of the two-stage channel, along with a cross-sectional view, are illustrated and explained in figure 1.

The overhead perspective of the experimental setup that has been created to carry out the planned study is shown in figure 2. A tailboard was erected at the upstream end of the flume to manage the water surface profile and ensure a consistent flow depth across the flume section. While measuring the profile of the water's surface, a point gauge was used, and its accuracy was 0.1 mm. For the purpose of determining three-dimensional velocity distributions, an Acoustic Doppler Velocimeter (ADV) was used. The ADV data were processed with the Horizon ADV software after collection. In addition to analyzing velocity distributions, boundary shear stress lateral distributions were also measured at the same points using a 5 mm outer diameter Preston tube. A digital manometer was utilized to ensure accurate pressure differential readings. After that, the calibration equations found in Patel [22] were used in order to calculate the shear stress values. The study involved the acquisition of local wall shear stress and velocity measurements at regular intervals of 2.5 cm and 10 cm along the vertical and horizontal axes, respectively, around the wetted perimeter (figure 1).

2.3. Theoretical Background

In a river that is flooded, the roughness of the channel section changes laterally throughout the perimeter of the wet area. It is not possible to establish the conveyance capacity of a compound channel without first determining its composite roughness. Due to the potential correlation between Manning's n, Chezy's C, and the Darcy-Weisbach f, the equations are exclusively evaluated and summarized based on their predictive capability of the composite roughness coefficient. The equations exhibit diverse structures and forms. However, they can be classified into distinct groups based on various criteria. These groups include various methods for the computation of composite roughness in two-stage channels are presented in table 2. The equation (4) for a compound channel with converging floodplains was established by Das and Khatua [12] by the use of multivariate analysis in terms of geometric and flow characteristics for the compound channel. The composite roughness equations are

$$n = \frac{1}{A}\sum_{i=1}^{i=N} A_i n_i \tag{1}$$

$$n = \left(\frac{1}{P}\sum_{i=1}^{i=N} P_i \;\; {n_i}^{3/2}\right)^{2/3} \tag{2}$$

$$n = PR^{5/3}\left(\sum_{i=1}^{i=N} \frac{P_i \;\; {R_i}^{5/3}}{n_i}\right)^{-1} \tag{3}$$

$$n = 0.096\{1 - 0.0131e^{0.211(\alpha)} - 0.068e^{0.3(\beta)} - 0.082(\theta)^{0.046} - 0.235e^{0.101(x)} - 0.791(R)^{-0.214} + ln\, ln\;\; (\delta^{0.0114} S_0^{0.073} F^{-0.167})\;\}\tag{4}$$

Composite roughness for compound channel is calculated by the following expression:

$$n = \frac{AR^{2/3}\sqrt{S_e}}{Q} \tag{5}$$

Various factors such as width ratio (α), relative depth ratio (β), aspect ratio (δ), converging angle (θ), relative distance (X_r), longitudinal slope (S_o), energy slope (S_e), discharge ratio (Q_r), Froude's number (F_r), and Reynolds's number (R_e) have an impact on Manning's n values. Multiple parameters are employed to facilitate the application of the model equation to various compound channels.

The necessary dimensionless equation may be expressed as follows:

$$n = F(\alpha, \beta, \delta, \theta, X_r, S_o, S_e, Q_r, F_r, R_e) \tag{6}$$

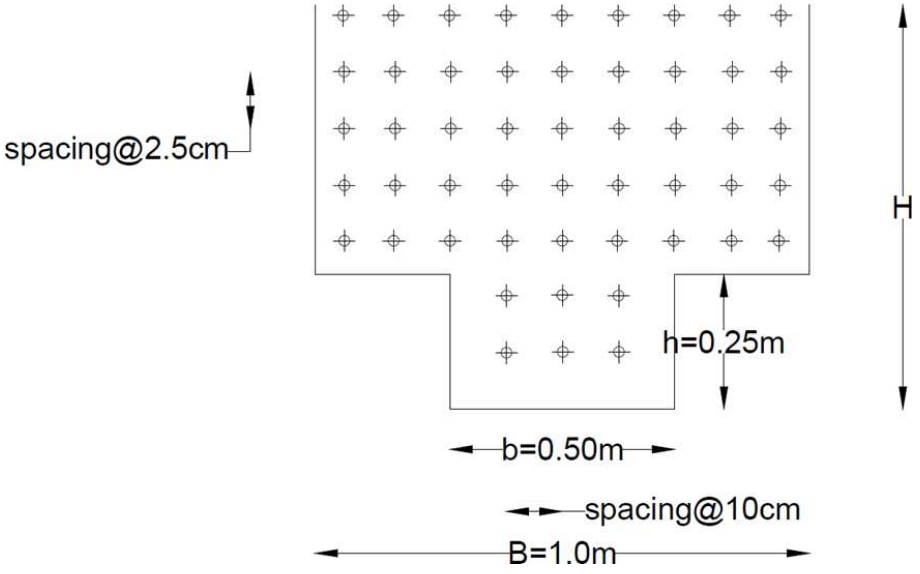


Figure 1. Cross-section of a two-stage channel

Table 1. Channel dimensions used in the present experimental study

Verified Test Channel	Type of Channel	Angle of Convergent (θ)	Longitudinal Slope (S)	Cross-sectional Geometry	Total Channel Width (B) (m)	Main Channel Width (b) (m)	Main Channel Depth (h) (m)	Converging Length (L) (m)	Aspect ratio (δ)
Channel 1 Rezaei (2006)	Converging	11.31°	0.002	Rectangular	1.2	0.398	0.05	2	7.96
Channel 2 Rezaei (2006)	Converging	3.81°	0.002	Rectangular	1.2	0.398	0.05	6	7.96
Channel 3 Rezaei (2006)	Converging	1.91°	0.002	Rectangular	1.2	0.398	0.05	6	7.96
Channel 1 Naik and Khatua (2016)	Converging	5°	0.0011	Rectangular	0.9	0.5	0.1	2.28	5
Channel 2 Naik and Khatua (2016)	Converging	9°	0.0011	Rectangular	0.9	0.5	0.1	1.26	5
Channel 3 Naik and Khatua (2016)	Converging	12.38°	0.0011	Rectangular	0.9	0.5	0.1	0.84	5
Present Channel	Converging	4°	0.001	Rectangular	1.0	0.5	0.25	3.60	2

Table 2. The equations pertaining to the resistance coefficient of a two stage channel [12]

Composite roughness (*n*) equation	Assumptions	References
Eq. (1)	(1) The sum of all the subsection shear force equals to the total shear force, (2) for all the sections the friction slope is same, (3) the velocities of the considered subsection is varying in proportion to the depth to a seven sixth power law	Cox method [23]
Eq. (2)	(1) For all the subsections the friction slope is same, (2) for all the subsections the velocities are equal	Einstein and Bank method [24]
Eq. (3)	(1) The sum of the constituent discharges equals to the total discharge, (2) for all subsections the friction slope is the same	Lotter method [25]

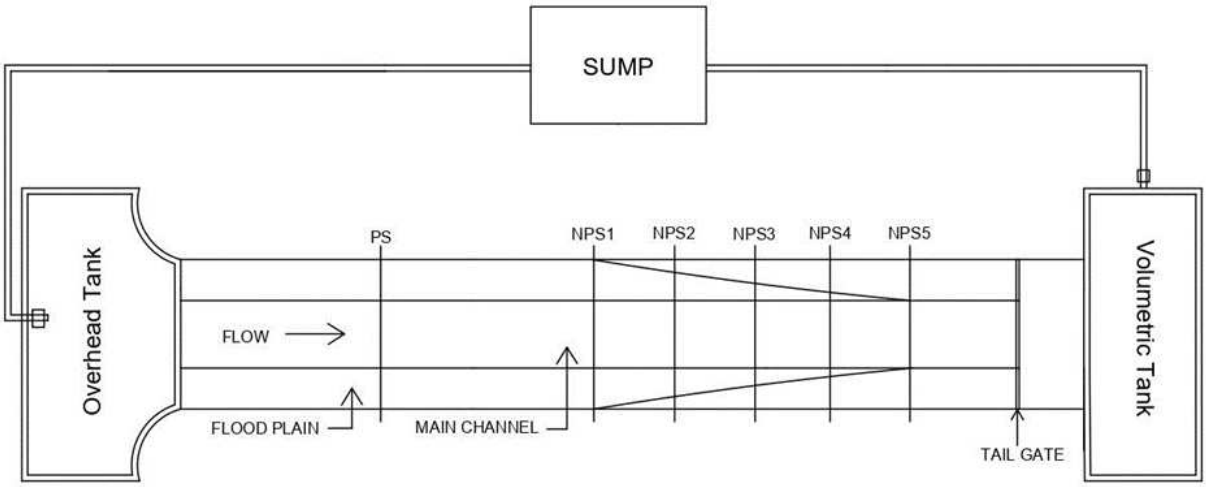


Figure 2. Experimental setup

2.4. GEP Model

The GEP methodology was initially introduced by Ferreira [26] and is a combination of Genetic Programming (GP) and Genetic Algorithms (GA) in computer programming. The approach entails utilizing linear chromosomes that possess a predetermined length and branching. The chromosomes, which may be either simple or linear, are linked together by structures that exhibit variability in both shape and size. This linkage mechanism bears a resemblance to the parse tree observed in Genetic Programming. The proposed methodology effectively segregates phenotype and genotype by organizing all branch structures into linear chromosomes of uniform length, regardless of their diverse dimensions and configurations. Gene-expression programming (GEP) is a unique multigene coding language that partitions complex equations into numerous sub-equations for modification. The technique employs gene generations, gene selection based on fitness, and genetic operators to introduce genetic variation. These elements are achieved through the use of genetic operators. To develop an equation for the manning's roughness coefficient (n) of a nonprismatic compound channel, the author employs a Gene-expression programming (GEP) strategy and GeneXproTools 5.0 [27] as a resource. The creation of models is based on the suitability of datasets used for both training and testing. The chosen models are duplicated in Gene Expression Programming (GEP) through the implementation of one or multiple genetic operators, such as mutation or recombination. A detailed and all-encompassing conceptual summary of GEP can be found in prior research depictions [28]. The sixth equation depicts the variability of the Manning's roughness coefficient in a two-stage channel with narrowing floodplains, as a result of changes in geometric and flow parameters. The present investigation employs ten distinct input variables as independent factors and n as the target value, as explicated in equation (6). The distribution of 396 data sets is randomized between two modelling stages. The initial stage of the model involves a solitary gene and two distinct cranial dimensions. Throughout the modelling procedure, there was a gradual increase in the number of genes and heads in each iteration, and the outcomes from both the training and testing datasets were documented after each run. However, the performance of the model did not show a significant improvement beyond twelve head lengths and six genes when comparing the results from the training data phase to those from the testing data phase. Due to the lack of significant improvement in performance beyond twelve head lengths and six genes during the testing and training phases, they were selected as the optimum values for the GEP model. These genes were linked using the addition function, which was essential to bind the genes. After

695000 generations, it was observed that the fitness function value and the coefficient of determination of both the training and testing data remained constant. This suggests that the generational progress may have reached its conclusion at this juncture. Table 3 presents an overview of the key parameters that influence the efficacy of GEP modelling in developing a predictive model for estimating the manning's n of a compound channel featuring converging floodplains. Figure 3 illustrates the GEP model for the manning's n, representing input variables from d0 to d9 and the gene two constant by G2c0. An algebraic equation can be reconstructed from this expression tree by decoding it.

Table 3. Selection criteria and parameters of the GEP proposed model

Variables	Selected Values
Function Set	+, −, ×, /
Number of Chromosomes	30
Head Size	12
Number of Genes	6
Gene Size	38
Linking Function	Addition
Fitness Function	RMSE
Program Size	114
Literals	41
Number of Generations	695000
Constants per Gene	10
Data Type	Floating-point
Mutation	0.00138
Inversion	0.00546
Gene recombination rate	0.00277
Insertion sequence (IS) transposition rate	0.00546
Root insertion sequence (RIS) transposition rate	0.00546

3. Results and Discussion

The fluctuation in Manning's roughness coefficient (n) that occurs with increasing longitudinal distance is seen in figure 4. Due to a reduction in hydraulic radius throughout the longitudinal distance, the n values drop in the prismatic section and up to the middle of the nonprismatic portion. This happens along the longitudinal distance. After the midpoint of the nonprismatic section, there is a sharp rise in the values of n as a result of the converging geometry of the channel. This results in a sharp increase in the energy slope in comparison to the hydraulic radius. Following then, n values drop in the downstream section because the hydraulic radius and energy slope both become smaller as they go farther down the length of the flow. Higher relative flow depths exhibit the same pattern of fluctuation as lower relative flow depths. The fluctuation of Manning's n with nondimensional characteristics such as width ratio, energy slope, discharge ratio, and Froude number can be observed in figure 5. The values of n rise in a nonlinear fashion in response to increases in both the width ratio and the discharge ratio. When the n values have dropped to their lowest point, an upward trend in accordance with the energy slope will begin. The Froude number is correlated with a nonlinear decline in the n values. The GEP approach's ability to make accurate predictions is demonstrated by the scatter plots comparing the anticipated and observed n values in figure 6. The proximity of the values to the line denoting exceptional concurrence is a noteworthy indication of the potential of the GEP model. The dispersed GEP predicted values over the best-fitting line demonstrate the highest level of agreement with the experimental data, yielding high R^2 values. To evaluate the efficacy of the GEP model, statistical measures such as R^2, MSE, RMSE, MAE, and MAPE were used, and their values are presented in table 4. The GEP model gives the best predicting results (R^2 = 0.84 and RMSE = 0.0006 in training, R^2 = 0.86 and RMSE = 0.0005 in testing) for the manning's roughness coefficient of the composite waterway with nonprismatic floodplains compared to previous methods such as Das and Khatua [12] and the theoretical method for composite roughness computation. The GEP model with the lowest MAPE is the technique best suited for predicting flow resistance in compound channels with converging floodplains when compared to other methodologies.

Table 4. Statistical analysis of predicted n by various approaches

Parameters	GEP Model		Das and Khatua Method (2018)	Lotter method
	Testing	Training		
R^2	0.84	0.86	0.70	0.78
MSE	3.6×10^{-7}	2.9×10^{-7}	0.0019	3.9×10^{-7}
RMSE	0.0006	0.0005	0.043	0.0006
MAE	0.0003	0.0003	0.002	0.0005
MAPE (%)	3.25	2.84	10.25	5.68

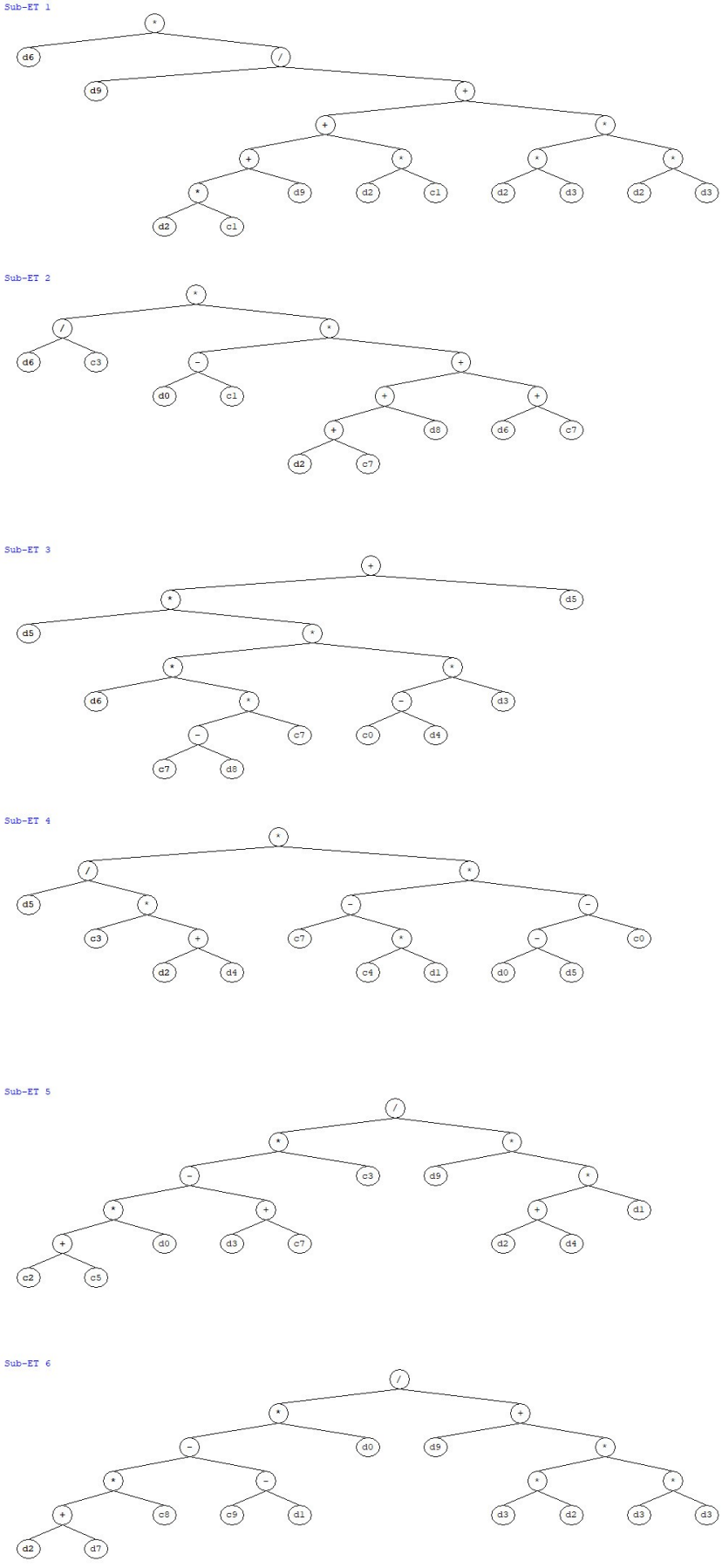


Figure 3. GEP proposed expression tree

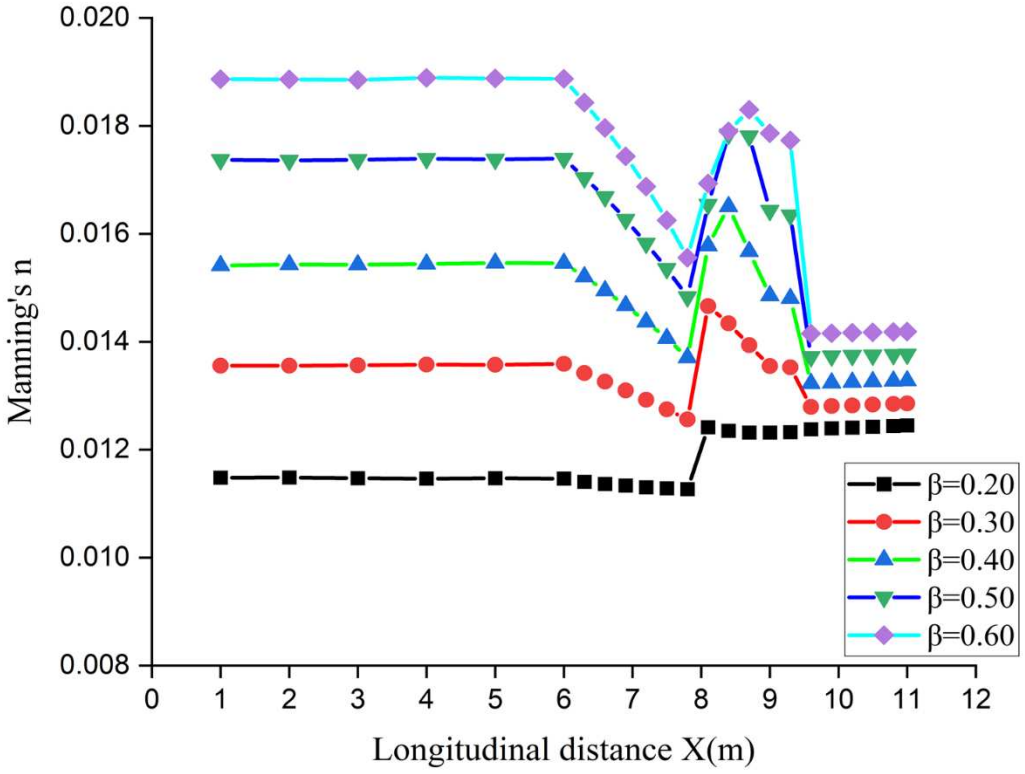
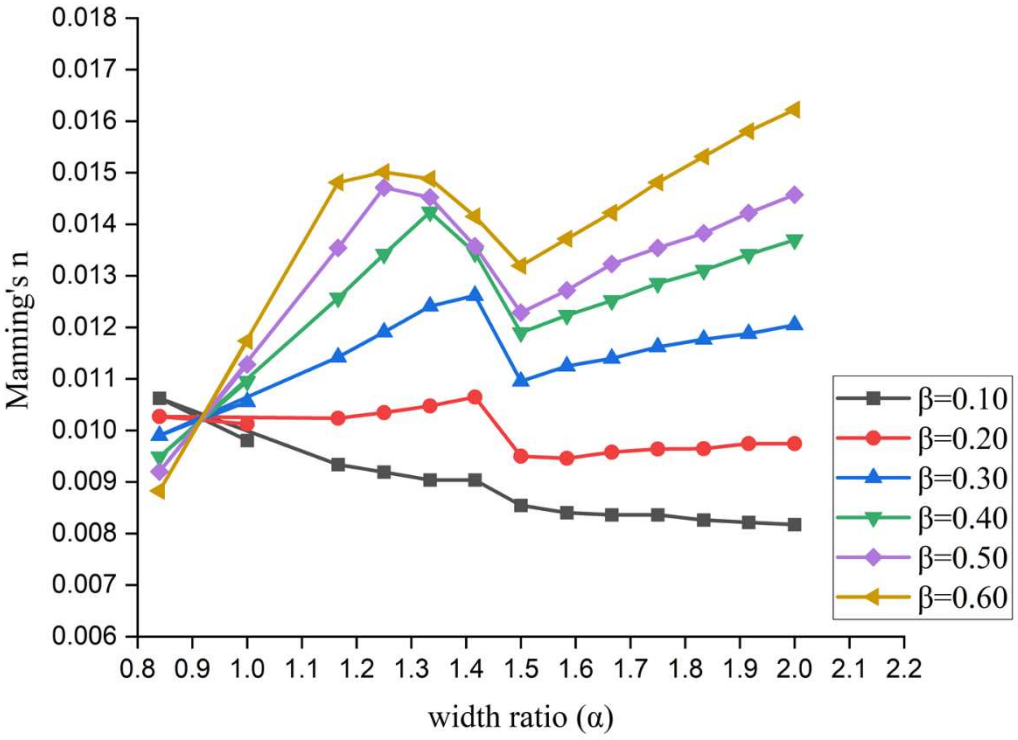
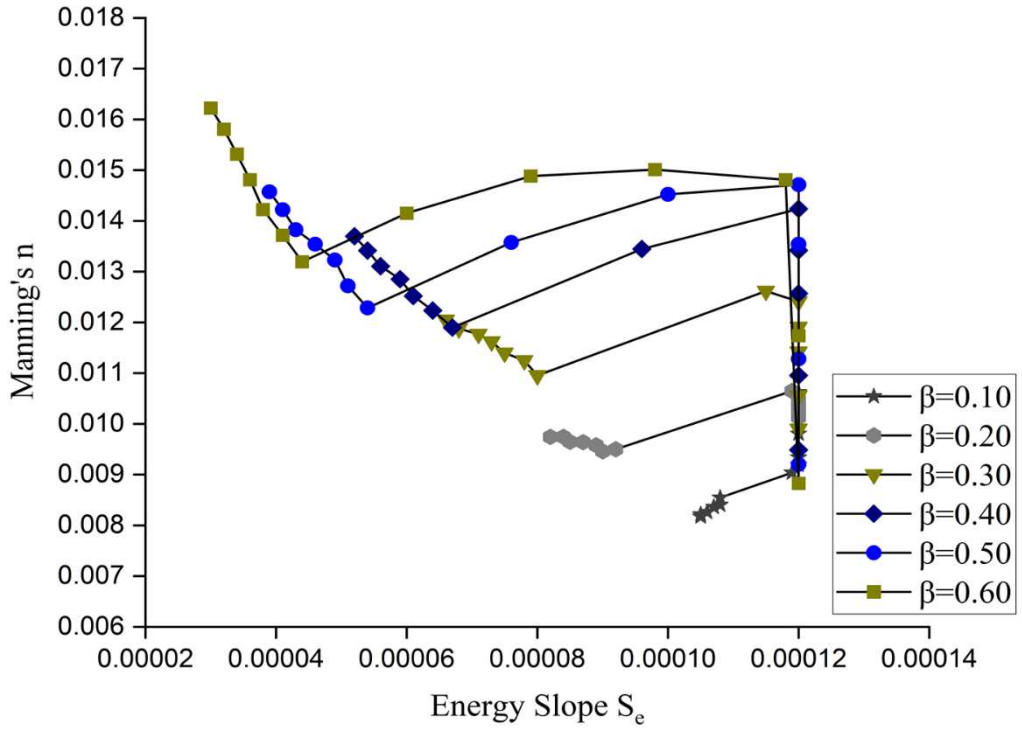


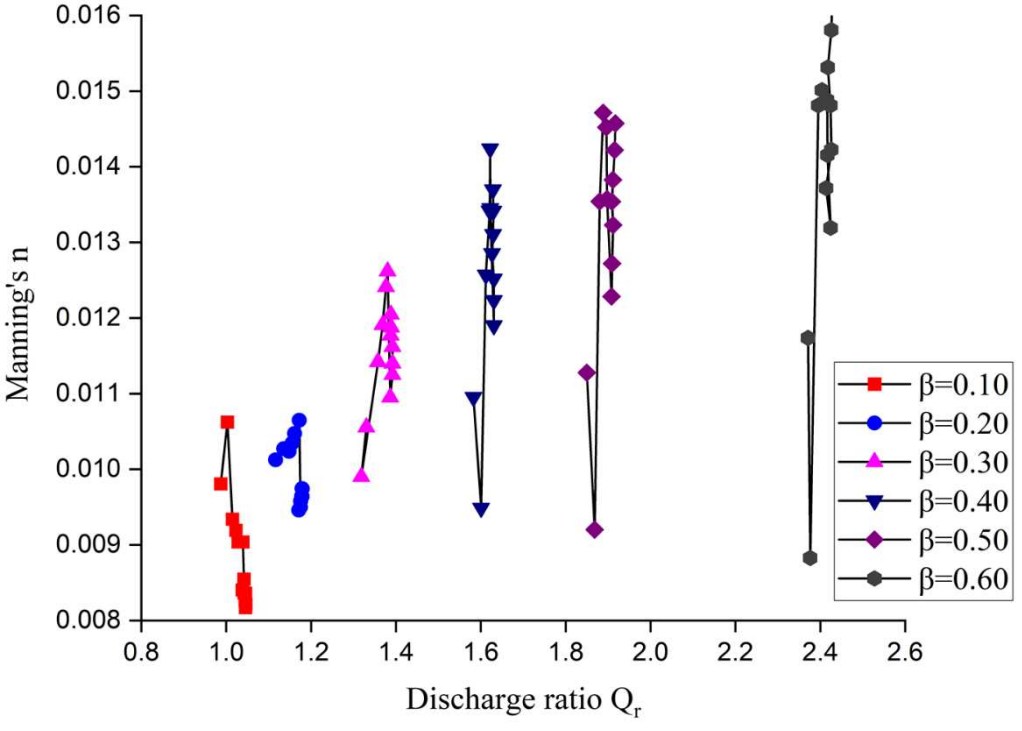
Figure 4. Variation of Manning's n with longitudinal distance



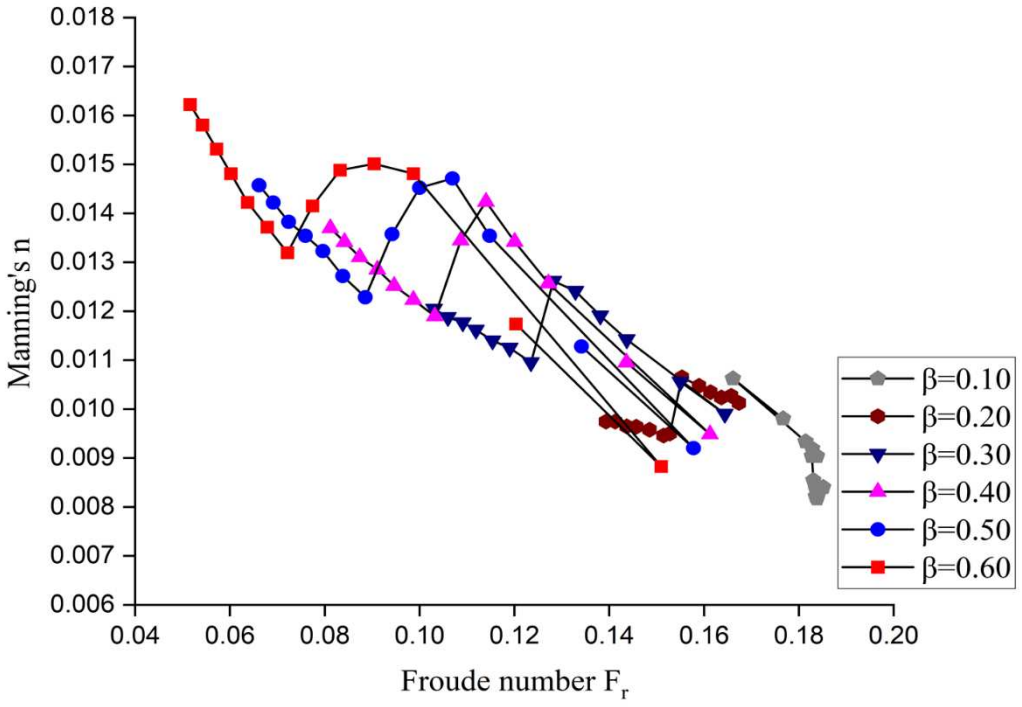
(a)


Manning's n
Energy Slope S_e
β=0.10
β=0.20
β=0.30
β=0.40
β=0.50
β=0.60

(b)


Manning's n
Discharge ratio Q_r
β=0.10
β=0.20
β=0.30
β=0.40
β=0.50
β=0.60

(c)



(d)

Figure 5. Variation of Manning's n with (a) width ratio (b) energy slope (c) discharge ratio (d) Froude number

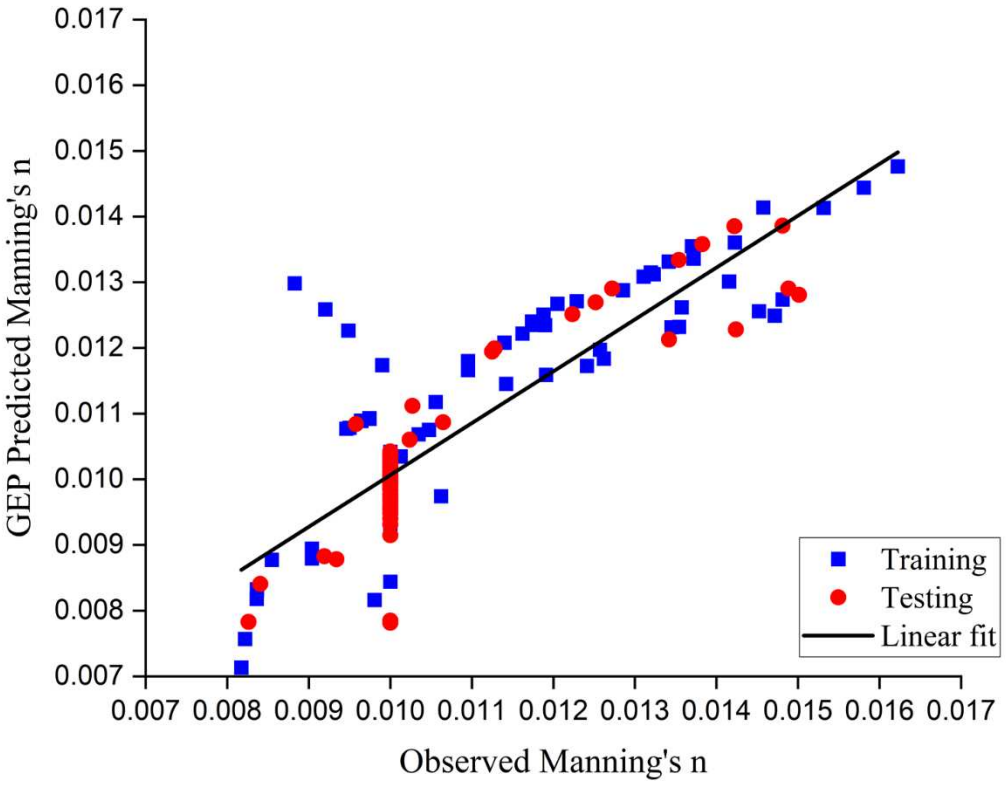


Figure 6. Comparison between actual and predicted n for the GEP Training and Testing phase

4. Conclusions

The study employs Gene Expression Programming (GEP) to estimate the Manning's roughness coefficient (n) in composite waterways with narrowing floodplains. The GEP model is developed using high-quality experimental datasets with varying geometrical and flow variables. The channels had an angle of convergence and relative depth of flow ranging from 1.91° to 12.38° and 0.10 to 0.60, respectively. Based on the study results, the following conclusions and inferences can be made:

The study indicates that the proposed model is influenced by a range of features, such as the width ratio, relative depth of flow, aspect ratio, angle of convergence, relative longitudinal distance, longitudinal bed slope, energy slope, discharge ratio, Froude number, and Reynolds number. In the prismatic portion, the Manning roughness coefficient decreases, and in the nonprismatic section, it decreases up to the middle. After then, a precipitous increase in the roughness value is seen. In a converging compound channel, the value of Manning's roughness goes up nonlinearly with width ratio, energy slope, and discharge ratio, while it goes down with the Froude number. It has been noticed that the Manning's n in nonprismatic compound channels follows the same pattern of change regardless of the angle of convergence. The present study aims to examine the correlation between the nondimensional Manning's n coefficient and the non-dimensional geometric and flow variables of a composite waterway with narrowing floodplains. It has been discovered that there is a connection that is not linear between each of the parameters. In comparison to earlier techniques such as Das & Khatua [12] and the Lotter method, the newly created model demonstrates superior performance in terms of R^2, MAE, RMSE, and MAPE for a variety of datasets. These metrics are measured by mean absolute percentage error. The results showed that, in line with the evaluation criteria, the GEP approach could fairly forecast the manning's in compound channels with converging floodplains. This was shown by the fact that the technique was successful. Since it had the best R^2 and the lowest RMSE, MAE, and MAPE, the GEP model demonstrated superior performance. On the other hand, a fresh tree technique that utilizes the GEP method was created in order to estimate the manning's roughness coefficient of a compound channel that has converging floodplains. A constraint of the model lies in its applicability solely to the estimation of the Manning's roughness coefficient for a composite channel featuring a floodplain that converges and exhibits uniform roughness. This is the model's sole use. Estimating the Manning's roughness coefficient of nonprismatic compound channels with rugged floodplains and innovative methodologies should be the primary focus of research that will be conducted in the future.

Acknowledgements

The authors are grateful for the encouragement offered by the Department of Civil Engineering at Delhi Technological University in Delhi, India.

REFERENCES

[1] Lambert M. F., Sellin R. H. J., "Discharge prediction in straight compound channels using the mixing length concept," J. Hydraul. Res., 34(3), 381–394, 1996. https://doi.org/10.1080/00221689609498487

[2] Khatua K. K., Patra K. C., Mohanty P. K., "Stage-discharge prediction for straight and smooth compound channels with wide floodplains," J. Hydraul. Eng., 138(1), 93–99, 2012. https://doi.org/10.1061/(ASCE)HY.1943-7900.0000491

[3] Devi K., Khatua K. K., Das B. S., "A numerical solution for depth-averaged velocity distribution in an open channel flow," ISH J. Hydraul. Eng., 22(3), 262–271, 2016. https://doi.org/10.1080/09715010.2016.1184104.

[4] Knight D. W., "Flow mechanics and sediment transport in compound channels," Int. J. Sediment Res., 14(2), 217–236, 1999.

[5] Pang B., "River flood flow and its energy loss," J. Hydraul. Eng., 124(2), 228–231, 1998. https://doi.org/10.1061/(ASCE)0733-9429(1998)124:2(228)

[6] Bousmar D., Wilkin N., Jacquemart J. H., Zech Y., "Overbank flow in symmetrically narrowing floodplains," J. Hydraul. Eng. ASCE, 130(4), 305–312, 2004.

[7] Proust S., Rivière N., Bousmar D., Paquier A., Zech Y., "Flow in the compound channel with abrupt floodplain contraction," J. Hydraul. Eng., 132(9), 958–970, 2006.

[8] Chlebek J., Bousmar D., Knight D.W., Sterling M. A., "Comparison of overbank flow conditions in skewed and converging/diverging channels," In: River Flows International Conference, pp. 503–511, 2010.

[9] Rezaei B., Knight D. W., "Overbank flow in compound channels with non-prismatic floodplains," J. Hydraul. Eng., 137, 815–824, 2011.

[10] Yonesi H. A., Omid M. H., Ayyoubzadeh S. A., "The hydraulics of flow in nonprismatic compound channels," J. Civ. Eng. Urban., 3(6), 342-356, 2013.

[11] James M., Brown R. J., "Geometric parameters that influence floodplain flow," In U.S. Army Engineer Waterways Experimental Station, June, Vicksburg Miss. Research report H 77, 1977.

[12] Das B. S., Khatua K. K., "Flow Resistance in a Compound Channel with Diverging and Converging Floodplains," Journal of Hydraulic Engineering, 144(8), 2018b. https://doi.org/10.1061/(asce)hy.1943-7900.0001496

[13] Yang K., Cao S., Liu X., "Study on resistance coefficient in compound channels," Acta Mech. Sin., 21(4), 353–361, 2005. https://doi.org/10.1007/s10409-005-0042-x

[14] Yang K., Cao S., Liu X., "Flow resistance and its prediction methods in compound channels," Acta Mech. Sin., 23(1), 23–31, 2007b. https://doi.org/10.1007/s10409-006-0043-4

[15] Morvan H., Knight D., Wright N., Tang X., Crossley A., "The concept of roughness in fluvial hydraulics and its formulation in 1D, 2D, and 3D numerical simulation models," J. Hydraul. Res., 46(2), 191–208, 2008. https://doi.org/10.1080/00221686.2008.9521855

[16] Nalluri C., Adepoju B. A., "Shape effects on resistance to flow in smooth channels of circular cross-section," J. Hydraul. Res., 23(1), 37–46, 2010. https://doi.org/10.1080/00221688509499375

[17] Shiono K., Al-Romaih J. S., Knight D. W., "Stage-discharge assessment in compound meandering channels," J. Hydraul. Eng., 125(1), 66–77, 1999. https://doi.org/10.1061/(ASCE)0733-9429(1999)125:1(66)

[18] Shiono K., Knight D. W., "Turbulent open-channel flows with variable depth across the channel," J. Fluid Mech., 222, 617–646, 1991. https://doi.org/10.1017/S0022112091001246

[19] Bjerklie D. M., Dingman S. L., Bolster C. H., "Comparison of constitutive flow resistance equations based on the Manning and Chezy equations applied to natural rivers," Water Resour. Res., 41(11), W11502, 2005. https://doi.org/10.1029/2004WR003776

[20] Rezaei B., "Overbank Flow in Compound Channels with Prismatic and non-Prismatic Floodplains," Ph.D. Thesis, University of Birmingham, Birmingham, UK, 2006.

[21] Naik B., Khatua K. K., "Boundary shear stress distribution for a converging compound channel," ISH Journal of Hydraulic Engineering, 22(2), 212-219, 2016. doi:10.1080/09715010.2016.1165633

[22] Patel V. C., "Calibration of the Preston tube and limitations on its use in pressure gradients," J. Fluid Mech. 231, 85–208, 1965.

[23] Cox R. G., "Effective hydraulic roughness for channels having bed roughness different from bank roughness," Miscellaneous paper H-73-2. Vicksburg, MS: U.S. Army Corps of Engineers Waterways Experiment Station, 1973.

[24] Einstein H. A., Banks R. B., "Fluid resistance of composite roughness," Eos, Trans. Am. Geophys. Union, 31(4), 603–610, 1950. https://doi.org/10.1029/TR031i004p00603

[25] Lotter G. K., "Considerations on hydraulic design of channels with the different roughness of walls," Trans. All-Union Sci. Res. Inst. Hydraul. Eng., 9, 238–241, 1933.

[26] Ferreira C., "Gene expression programming: a new adaptive algorithm for solving problems," Complex Syst., 13(2), 87–129, 2001.

[27] Gepsoft GeneXproTools 5.0., "Data Modeling & Analysis Software," 2014. https://www.gepsoft.com/

[28] Mallick M., Mohanta A., Kumar A., Patra K. C., "Gene-expression programming for the assessment of surface mean pressure coefficient on building surfaces," Build. Simul, 13, 401–418, 2020.

Application of Response Surface Method and Genetic Algorithm in the Design of High-Efficiency Prototype Vehicle

Paras Singh[1], Harshit Gupta[1*], Ojas Vinayak[1], Aryan Tyagi[1]

[1] Department of Mechanical Engineering, Delhi Technological University, New Delhi, India

[*] hgh1602@gmail.com

ABSTRACT

Breakthroughs in aerodynamic optimization have made it possible to develop efficient modes of transport with lesser exploitation of valuable resources. This makes it crucial for technical professionals such as engineers and scientists to understand the methodologies behind carrying out such optimizations. A common approach towards improving the aerodynamic properties of a vehicle is to alter its physical shape, which has concurrently been a very strenuous process given the time consumed to remodel the vehicle for each simulation process. This research aims to tackle this problem by using intelligent techniques to automate the step-by-step process of remodeling the car and arriving at a final optimized solution with a significantly lower drag coefficient, a quantity used to measure the amount of drag force acting on a vehicle. This is achieved by assigning particular parameters to ensure guided improvement of the airfoil in a process known as parametrization, followed by implementing a response surface methodology primarily to circumvent the strenuous task of performing a large number of CFD simulations by employing surrogate models to generate a response surface between selected independent variables. Further, evolutionary algorithms such as Genetic Algorithm have gained momentum in the optimization studies carried out during product design by selecting the optimum parameters from the available design spaces on the basis of natural evolution. The proposed method of optimization has been successfully implemented on a prototype vehicle with an improvement of 26.6% and 51.1% in the drag coefficient and drag area respectively.

Keywords: Optimization, Genetic Algorithm, Efficiency, Vehicle Aerodynamics, Computational Fluid Dynamics

1 Introduction

Understanding airflow dynamics in the automotive sector is crucial for effective aerodynamic design within technological constraints [1]. A substantial amount of fuel can be saved provided a vehicle is aerodynamically compliant. Vehicle optimization during the design process aims to enhance desired aerodynamic properties, with minimizing drag through improved body design being a prominent strategy. This improves fuel efficiency, increases top speeds, and passenger comfort. Crucial developments in the past have led to the technological advancements in aerodynamic design optimization that we employ today.

Past studies have demonstrated the efficiency of optimization frameworks in various fields [2,3,4]. For example, Kaveh et al. [5] emphasize the benefits of employing the CS algorithm in the creation of two-dimensional steel frames. Their research demonstrates that Levy flights outperformed Gaussian or uniform distributions, which are frequently employed by alternative metaheuristic algorithms when it comes to navigating extensive design spaces.

Furthermore, work has been done on researching optimization algorithms and their roles in the aerodynamic optimization of various components, specifically, the use of sensitivity algorithms has also been studied in relation with computational fluid dynamics to optimize shapes [6], and the use of benchmarking optimization algorithms for wing design optimization [7,8].

The present study adds to the existing literature by exploring the application of genetic algorithms, surrogate modeling, and computational fluid dynamics (CFD) to optimize the aerodynamic characteristics of a prototype vehicle for Shell Eco Marathon, a global-level competition in which collegiate teams compete to make the most energy-efficient vehicles. It has 2 different vehicle classes - (1) Urban Concept and (2) Prototype. Urban concept cars are required to resemble the original on-road cars, whereas prototypes can have any design (must be compliant with Shell Eco Marathon official rules) whether 3 or 4-wheeled which is aimed at achieving ultimate performance in terms of mileage. By comparing it with conventional hit-and-trial design processes involving repeatedly modifying the vehicle body and carrying out simulations, the superiority of this method is demonstrated in achieving enhanced efficiency, reduced drag, and improved overall performance.

The present study breaks down the overall vehicle geometry into two segments- the side and the top profile. These profiles are composed of three-degree Bezier curves, the top profile being symmetric and the side profile having an asymmetric shape about the longitudinal axis. The control points of these curves are used as the parameters of the optimization study with the space for mounting the powertrain and steering components providing constraints on the value of these parameters. In the next step, a Response Surface Methodology is employed to circumvent the large number of simulations by creating a response surface between the control points and the objective function (Drag Coefficient). After this, the genetic algorithm is employed to ensure a globally optimal airfoil shape. This is achieved through probabilistic selection based on fitness, combining genetic information through crossover, introducing diversity through mutation, and iteratively evolving a population of potential solutions. The obtained results are analyzed and presented in the later stages of the research paper through aerodynamic visualizations and velocity/pressure distribution graphs of selected shapes of the vehicle.

2 Methodology

2.1 Numerical Model Validation

In the current work, wind tunnel data for Ahmed body [9] is utilized as a reference to validate the CFD model. The Ahmed body is a well-established benchmark used for validating CFD solvers, as it has all the necessary flow features encountered by modern automotive designs. Since the current study aims to propose a new vehicle design, there is no experimental data for validating the same model. Therefore, steady-state RANS simulations with the k-ω SST model [10] carried out using the commercial CFD code Ansys Fluent version 22.1 are utilized to first compute drag coefficients for the Ahmed Body, and in the next step, the validated model would be used for generating the required data for the optimization algorithm. The dimensions of the Ahmed body used for the study are presented in Figure 1. For the simulations, a freestream velocity of 60 m/s is assigned at the inlet corresponding to a Reynolds number of 4.25×10^6 based on the model length.

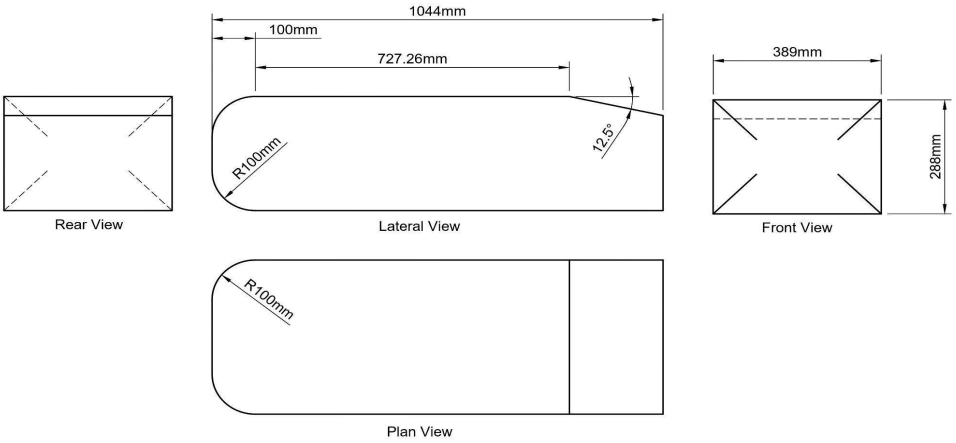


Figure 1: Ahmed Body dimensions

To examine the sensitivity of the numerical solution on the grid size, a grid independence test is conducted and the results for each grid are compared with the experimental data in Table 1 shown below.

Table 1: Results of Grid Independence and Validation study for the Ahmed Body

Parameter	Coarse	Medium	Fine	Experiment
Cd [-]	0.297	0.268	0.248	0.230
Cell Count [-]	$0.9x10^6$	$1.4x10^6$	$1.8x10^6$	-
Error [%]	29.13	16.52	7.82	-

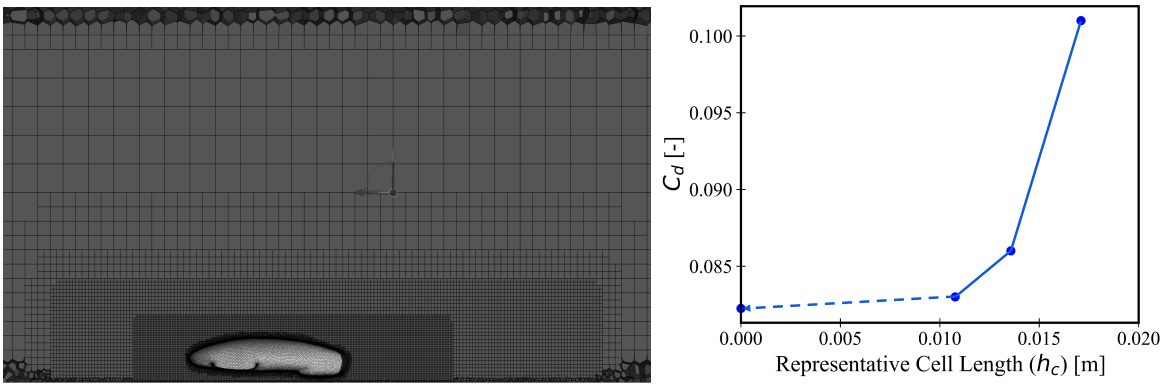
The accuracy of the numerical model can be clearly observed from Table 1. The fine grid is able to capture the quantity of interest, i.e., drag coefficient (C_d) accurately, and hence would be used for further parts of the study.

2.2 Grid Convergence Study

To ensure accuracy as well as efficiency in the simulations, a high-quality Polyhex core grid, shown in Figure 2(a), is used for discretization of the flow domain. The origin of the domain is at the nose of the vehicle body and the inlet is kept at a distance of 8 times the vehicle length and the outlet at a distance of 20 times the vehicle length. The side and the top walls are kept at a distance of 8 times the vehicle length. To save computational resources, finer cells are used to model the wake of the vehicle, and relatively coarser cells are used in the surrounding regions, For capturing the effect of the boundary

layer, 20 prismatic cells are used with a first layer height of 1mm and a growth rate of 1.12. The gradient discretization is done using the least square cell-based method and the second-order upwind scheme is used for spatial discretization of other transport quantities. The convergence criteria for residuals is set to 10^{-6} and a total of 1000 iterations are used to ensure the convergence of residuals and the drag coefficient.

To ensure solution independence with respect to the grid count, a grid convergence study is carried out using the Richardson extrapolation method [11]. The results of the study are presented in Figure 2(b). Three different grids were used with cell counts similar to Table 1. It can be observed that the solution for the fine grid is very close to the extrapolated solution for an infinitely fine grid. The error metrics for the grid are as follows: extrapolated relative error is 0.91%, relative error is 3.61%, and finally, the grid convergence index for the medium and the fine grid is 1.13.



(a) (b)

Figure 2: (a) Computational Grid, (b) Variation of C_d and Representative Cell Length

2.3 Parametrization of Airfoils

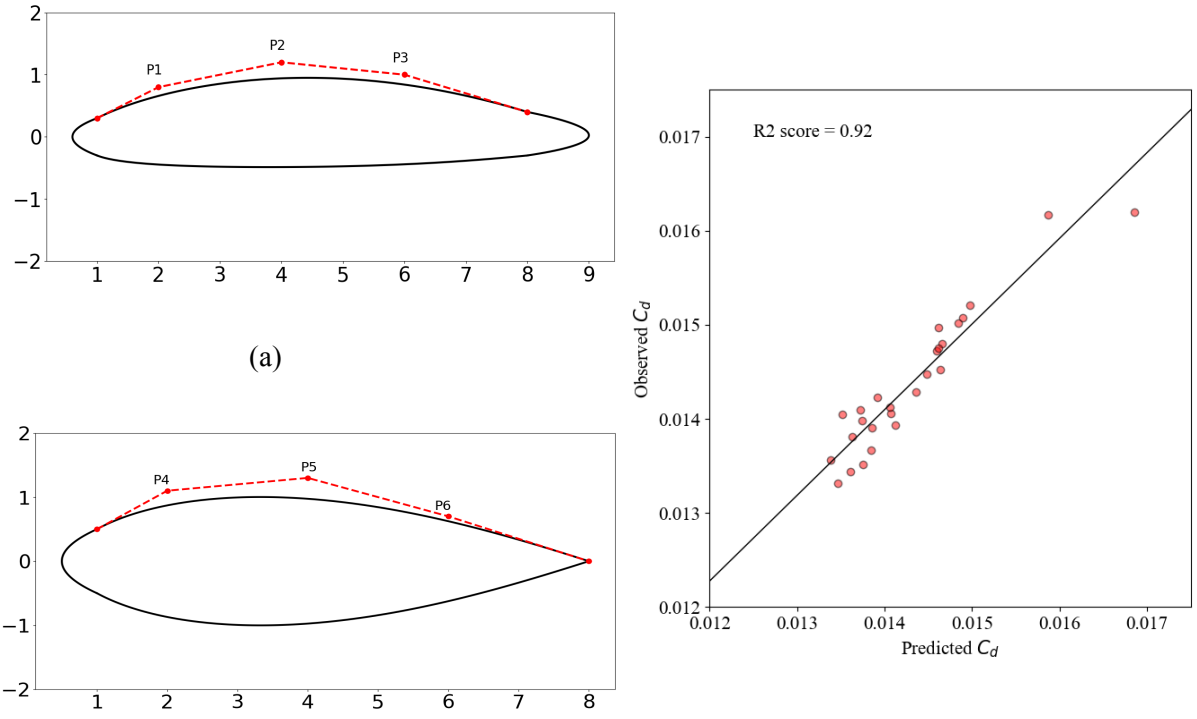
Parameterization of airfoils refers to the process of representing the complex geometry of an airfoil using a set of numerical parameters that are adjusted to satisfy an objective from the desired shape. These parameters are used to define the shape of the airfoil mathematically, allowing engineers and researchers to describe, analyze, and manipulate the airfoil's geometry in a systematic and standardized manner. The process is essential in the field of aerodynamics and vehicle design because it simplifies the representation of airfoil shapes, making it easier to perform computational simulations, conduct optimizations, and compare different airfoil designs.

Parametrization of airfoils is crucial for several reasons in the field of the development of aerodynamically efficient design. It allows engineers to optimize the shape of airfoils for specific aerodynamic performance metrics, such as lift-to-drag ratio. In our case, the drag coefficient serves as the foundational performance metric. The process allows researchers to study the effect of varying the different parameters on the behavior and structure of the object that is being designed thereby helping the designers in their decision-making process. Parametric models and techniques foster innovation and research by allowing designers and researchers to experiment with unconventional airfoil shapes. The efficiency of the technique is also extremely crucial as efficient parameterization methods help in quickly exploring a wide range of airfoil shapes, enabling engineers to identify the most efficient designs for various applications. The effect of airfoil shape parameterization in airfoil optimum shape and its convergence rate has been investigated by [12] and it was shown that a suitable parameterization method can improve the convergence rate of the optimization algorithm.

Among the various methods available for the process of parametrization, smooth parametrization of the airfoils is accomplished using B-spline curves. The Bezier curve is found to be suitable for airfoil parameterization, and the shape of the upper and lower curves can be precisely and visually controlled by Bezier control points [13]. Based on the vehicle's design constraints, the airfoils' chord is maintained at 3.2 meters. As the airfoil's lower surface represents the vehicle's floor in the side profile, it is left unchanged. Three control points, as seen in Figure 3, govern the shape of the top and side profiles. Only one side of the airfoil contains control points since the top profile is kept symmetrical about the airfoil's chord; the other side is produced by mirroring the top portion about the chord line, as illustrated in Figure 3 (b). The x-coordinates of the control points are fixed for both profiles. Table 2 lists the normalized values of the upper and lower limits for each control point. These values were decided based on the space constraints of other vehicle components, like the steering assembly, powertrain, and the space occupied by the driver. These control points are varied through various iterations between the upper and the lower limits to search for an optimum design with minimum C_d.

Table 2: Limits of control points for side and top profiles of the vehicle

Point ID/Constraints	X/c	Y/c Lower	Y/c Upper
P1	0.1250	0.0625	0.1875
P2	0.5000	0.1094	0.1875
P3	0.8750	0.0625	0.1875
P4	0.3125	0.1093	0.1406
P5	0.5312	0.0937	0.1406
P6	0.6875	0.0625	0.1406



(a)

(b)

(c)

Figure 3**:** Control Points of (a) side profile (b) top profile (c) Validation of Kriging surrogate model with ground truth CFD data

2.3 Response Surface Methodology

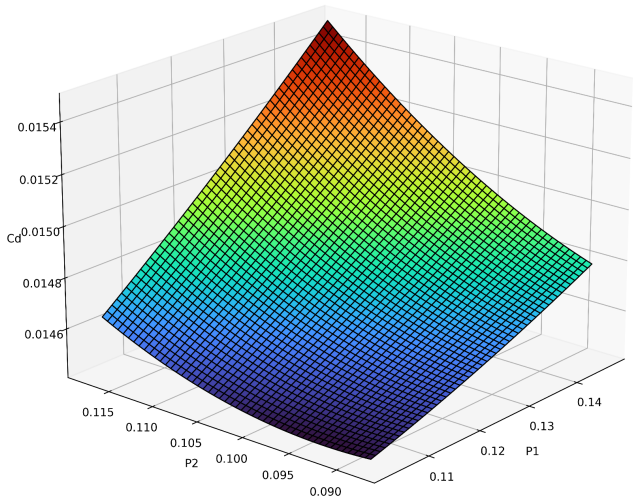


Figure 4: Response Surface Generated by Kriging Surrogate Model for Top Profile for P3 = 0.08

Response Surface Methodology (RSM) is used in problems where a large number of simulations are required for determining the optimal solution. In the present study, for the top and side views of the vehicle, the positions of two control points, each having an abscissa and an ordinate, will have to be optimized in order to obtain the most efficient design. Considering all the possible airfoils that can be generated, this leads to a large number of samples and it is impractical and inefficient to run thousands of CFD simulations. Thus, RSM is used to construct a response surface between the independent variables (control points) and the objective function (drag coefficient). Surrogate models are employed in RSM to construct the response surface. Kriging surrogate model is chosen for this objective. Kriging is an interpolation technique that involves forming a linear combination of a known function $f_i(\mathbf{x})$ and incorporating it with a realization of a stochastic process $Z(\mathbf{x})$

$$\sum_{i=1}^{k} \beta i f i(x) \; + \; Z(x)$$

where $Z(\mathbf{x})$ is a realization of a stochastic process with mean zero given by

$$cov\left[Z(x^{(i)}), Z(x^{(j)})\right] \; = \; \sigma^2 R(x^{(i)}, x^{(j)})$$

where σ^2 is the process variance, and R is the correlation.

The accuracy of the surrogate model is shown in Fig. 3 (c).

2.4 Optimization Algorithm - The Genetic Algorithm (GA)

Evolutionary algorithms (EAs) are a family of optimization algorithms inspired by the principles of biological evolution. They are part of the broader class of evolutionary computation techniques, which also include genetic algorithms, genetic programming, evolutionary strategies, and differential evolution, among others. EAs are used to find approximate solutions to complex optimization and search problems, especially in cases where the search space is large, multidimensional, or poorly understood. Evolutionary algorithms such as GA are used to obtain the globally optimum solution from the design space. The GA method imitates natural evolution wherein an initial population is chosen before crossover and mutation operations are often carried out depending on each individual's score for the objective function [14,15]. The genetic algorithm technique can be employed efficiently and accurately to produce globally optimal airfoils with excellent aerodynamic properties using desired objective values such as that of lift coefficient, drag coefficient, or the ratio of the two [16]. The primary objective of this study is the minimization of the net drag force experienced by the vehicle by optimizing its shape. The mathematical principles involve probabilistic selection based on fitness, combining genetic information through crossover, introducing diversity through mutation, and iteratively evolving a population of potential solutions. Over generations, the algorithm refines the solutions, converging towards an optimal or near-optimal solution for the given problem. The choice of appropriate parameters, such as selection methods, crossover points, and mutation rates, significantly influences the performance of the genetic algorithm. Intelligent and advanced GA also exist in the state-of-the-art that incorporates a fractional factorial crossover and SA mutation, which effectively improves the evolutionary efficiency of the conventional GA [17].

The process starts with “initialization” i.e. generating an initial population of candidate airfoil shapes by randomly generating or selecting individuals from the search space defined by the parameterization process.

$$x = (x1, x2, x3, \ldots..)$$

Further, the performance of each individual is evaluated using appropriate CFD simulations or analytical methods. The fitness value for each profile is evaluated based on the defined objective (drag minimization of prototype vehicle). The fitness function f(x) evaluates the quality of a solution represented by a chromosome x. It assigns a numerical value indicating how well the solution solves the problem. Based on the fitness values, individuals from the population are selected to serve as parents for the next generation.

$$P(x) = \frac{f(x)}{\sum f(xi)}$$

Individuals with a high fitness probability (lower fitness score) will be more likely to be chosen. Crossover and mutation are two genetic operations that are used to produce offspring from the chosen parents. While mutation causes random changes to the offspring, crossover entails merging the genetic information from two parents. The parents and offspring from the previous steps are combined to create a new population with random new samples that could better fit the objective function. The optimization process terminates on reaching a maximum number of generations or achieving a satisfactory solution depending on the level of fitness at the optimum solution. The best individual in the final population represents the optimized airfoil shape. The comprehensive flow of operation of the genetic algorithm has been represented in the flowchart in Figure 5.

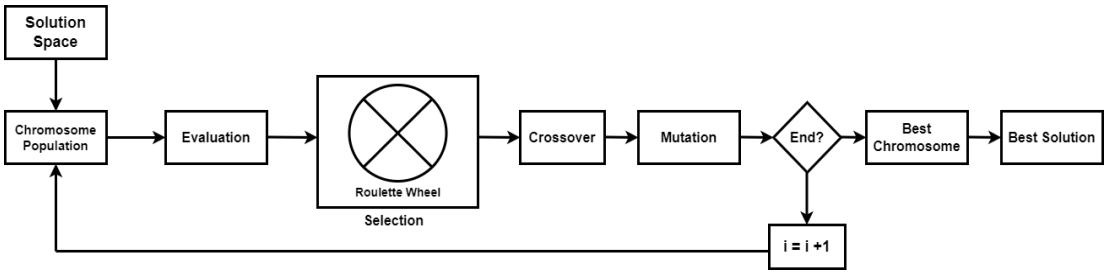


Figure 5: Genetic algorithm process flowchart

3 Results

3.1 Optimum Parameters

The results of the study are summarized in Table 3. Here initialized design refers to the vehicle body designed using the initial side and top profiles fed into the optimization algorithm and the optimum design refers to the final vehicle body designed using the optimized side and top profiles. An improvement of 26.6% can be observed in the overall Drag Coefficient and an improvement of 51.1 % in the Drag Area demonstrating the effectiveness of the Genetic algorithm in carrying out multi-parameter optimization studies. The optimum and initialized design are presented in Figure 6.

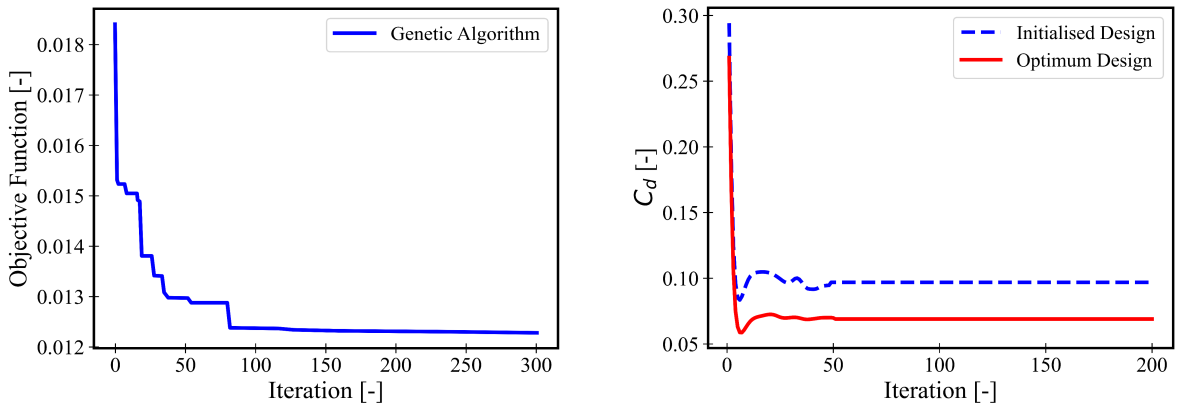
Table 3: Comparison of Initialized and Optimized Vehicle Design

Design	Drag Coefficient (C_d) [-]	Frontal Area (A) [m^2]	Drag Area (C_d. A) [m^2]
Initialised Design	0.094	0.72	0.0677
Optimum Design	0.069	0.48	0.0331

Figure 6: Comparison between the Initial and Optimized design

The graph in Figure 7 (a) provides insights into how the optimization process is progressing over time and is used to compare the convergence histories throughout the optimization process. The objective function represents the quality or fitness of the solutions generated by the algorithm and the number of iterations refers to the generations the algorithm has gone through. As the number of iterations increases, the objective function achieves a steady value. The rate of convergence towards the optimum solution can also be inferred from the graph showing the efficiency of the algorithm to obtain optimum solutions. At the end of the optimization process, the objective function value stabilizes and this final value gives an idea of the quality of the solution that the genetic algorithm has found. In

addition, the graph in Figure 7 (b) represents the convergence monitor for C_d for both the initial and optimum design as given by the CFD solver.

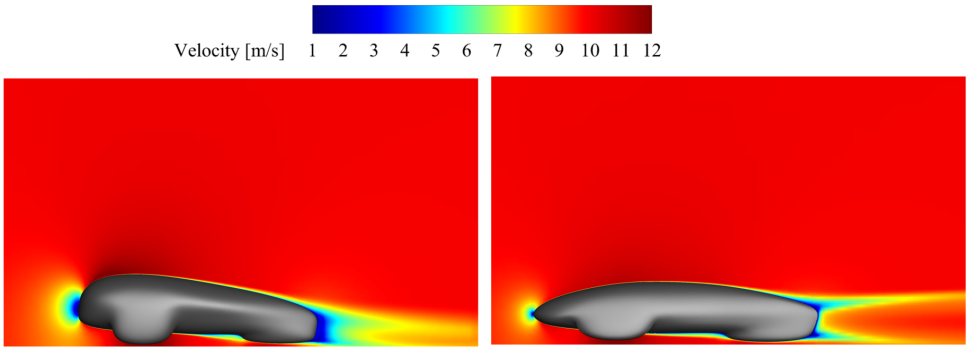


(a) (b)

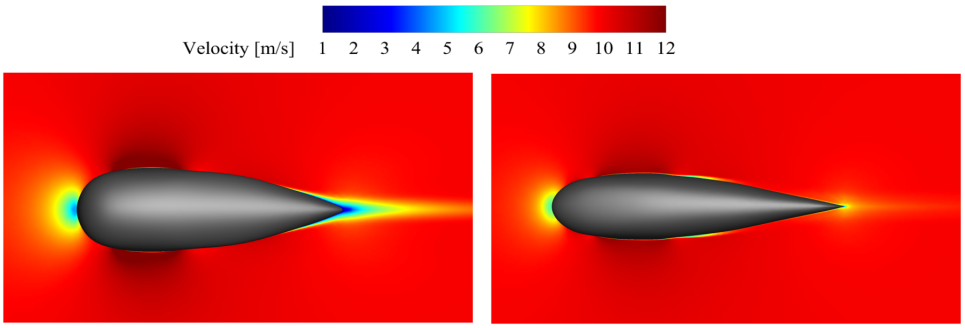
Figure 7: (a) Convergence of Objective Function for top profile using GA, (b) Convergence of C_d monitor obtained from the RANS simulation of initialized and optimum Design

3.2 Velocity Distribution

The velocity distribution around the initial and optimum geometry are presented side by side for comparison in Figure 8. The incoming air from the left-hand side impinges upon the nose of the vehicle and stagnates. The stagnation region is more prominent for the initial geometry due to its bluff shape and steep change in the curvature. On the other hand, the incoming air striking the nose of the optimum design flows smoothly over it due to its gradually changing curvature leading to a relatively smaller stagnation region as evident in Figure 8. Also, the wake of the initial geometry is more pronounced, implying a greater momentum deficit and thus larger drag force pulling the vehicle to the leeward side. The larger momentum deficit in the wake and a more prominent stagnation zone increase the imbalance of pressure in the longitudinal direction leading to a larger form drag for the initialized design which is also evident from Table 3.



(a)

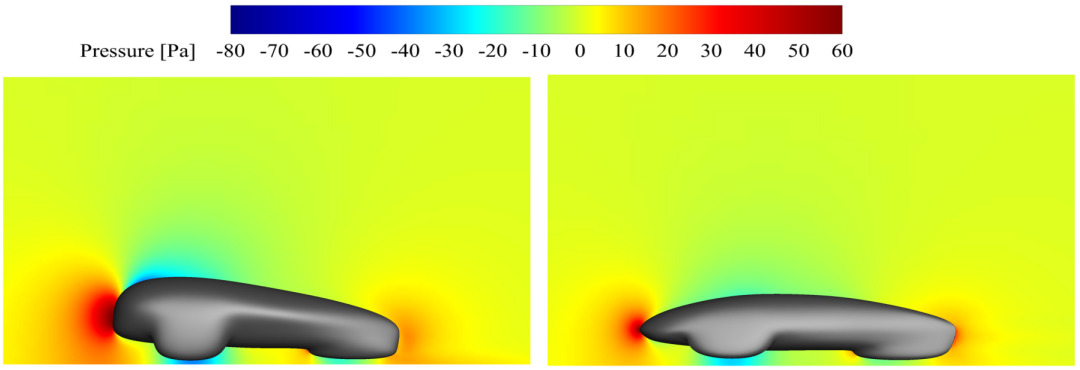


(b)

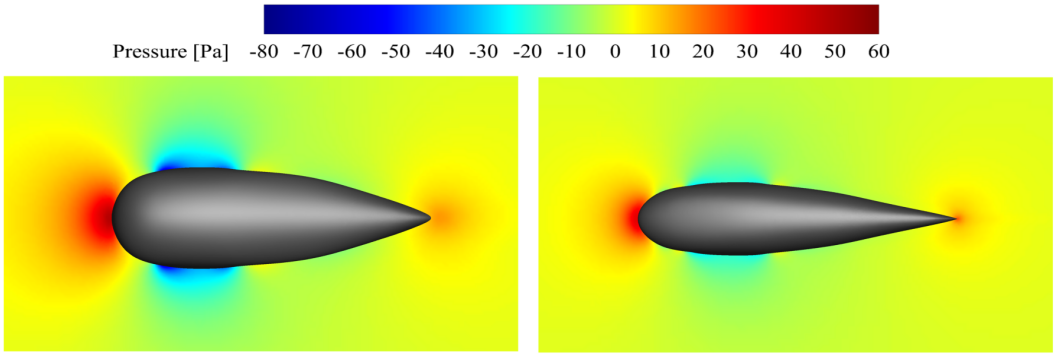
Figure 8: Velocity Distribution around the vehicle (a) Side View, (b) Top View

3.2 Pressure Distribution

A comparison has been made in the pressure distribution around both the geometries in Figure 9. The larger stagnation zone for the initial geometry as discussed previously, leads to a larger static pressure at the nose of the vehicle. This buildup of incoming air is mitigated for the optimum design due to its streamlined nose which aids the incoming air to flow over it easily without much hindrance. As the incoming air strikes the nose of the vehicle and starts to flow it, the pressure decreases due to the increase in the flow velocity due to the convexity of the curvature. As the air reaches the top of the vehicle and starts to flow down the curvature towards the rear end, the pressure starts to increase. This creates a pressure differential along the longitudinal direction contributing towards the form drag or pressure-induced drag force. Due to the symmetry of curvature for the optimum design as evident from the side view shown in Figure 9 (a), there is a greater pressure recovery at the rear end of the vehicle resulting in a much lower C_d compared to the initialized geometry.



(a)



(b)

Figure 9: Pressure Distribution around the vehicle (a) Side View, (b) Top View

4 Conclusion and Future Work

In conclusion, the proposed study demonstrates the effectiveness of utilizing Genetic Algorithm to optimize the design of a fuel-efficient prototype vehicle. This is achieved by employing parametrization to simplify simulations and a response surface methodology to create a response surface. Further, genetic algorithms are used to obtain globally optimum solutions. The minimization of drag coefficient through this approach led to a significant improvement in aerodynamic efficiency. These findings highlight the potential of the present methodology in achieving enhanced performance and fuel economy through improvements in vehicle design. The current approach only focuses on the side and top profile, but to cater to design optimization for more complex geometries further developments can be done in the area of parameterization of three-dimensional geometries.

5 References

[1] Sarkar, Subhasis & Thummar, Kunj & Shah, Neel & Vagrecha, Vishal. (2019). A Review paper on Aerodynamic Drag Reduction and CFD Analysis of Vehicles. 10.13140/RG.2.2.31506.91849.

[2] H.-L. Minh, T. Sang-To, M. Abdel Wahab, T. Cuong-Le, Structural damage identification in thin-shell structures using a new technique combining finite element model updating and improved cuckoo search algorithm, Advances in Engineering Software 173 (2022) 103206. doi:https://doi.org/10.1016/j.advengsoft.2022.103206.

[3] G. Hu, J. Wu, H. Li, X. Hu, Shape optimization of generalized developable H-Bezier surfaces using adaptive cuckoo search algorithm, Advances in Engineering Software 149 (2020) 102889. doi:https://doi.org/10.1016/j.advengsoft.2020.102889.

[4] A. Tsipianitis, Y. Tsompanakis, Improved cuckoo search algorithmic variants for constrained nonlinear optimization, Advances in Engineering Software 149 (2020) doi:https://doi.org/10.1016/j.advengsoft.2020.102865.

[5] A. Kaveh, T. Bakhshpoori, M. Azimi, Seismic optimal design of 3d steel frames using cuckoo search algorithm, The Structural Design of Tall and Special Buildings (2015). doi:10.1002/tal.1162.

[6] Baysal, O. and Eleshaky, M.E., 1992. Aerodynamic design optimization using sensitivity analysis and computational fluid dynamics. *AIAA journal*, *30*(3), pp.718-725.

[7] Lyu, Z., Xu, Z. and Martins, J.R.R.A., 2014, July. Benchmarking optimization algorithms for wing aerodynamic design optimization. In *Proceedings of the 8th International Conference on Computational Fluid Dynamics, Chengdu, Sichuan, China* (Vol. 11, p. 585).

[8] Aryan Tyagi, Paras Singh, Aryaman Rao, Gaurav Kumar, & Raj Kumar Singh. (2023). A Novel Framework for Optimizing Gurney Flaps using RBF Neural Network and Cuckoo Search Algorithm.

[9] Franck, Gerardo & Nigro, Norberto & Storti, Mario & D'Elía, Jorge. (2009). D'ELIA, J.. Numerical simulation of the flow around the ahmed vehicle model. Latin American Applied Research. 39. 295-306.

[10] Menter, Florian. (1993). Zonal Two Equation k-w Turbulence Models For Aerodynamic Flows. AIAA Paper. 1993. 10.2514/6.1993-2906.

[11] Ismail Celik, U Ghia, P.J. Roache, Chris Freitas, H Coloman, and Peter Raad. Procedure of estimation and reporting of uncertainty due to discretization in cfd applications. J. Fluids Eng., 130:078001, 07 2008. doi: 10.1115/1.2960953.

[12] Shahrokhi, A. and Jahangirian, A., 2007. Airfoil shape parameterization for optimum Navier–Stokes design with genetic algorithm. *Aerospace science and technology*, *11*(6), pp.443-450.

[13] Wei, X., Wang, X. and Chen, S., 2020. Research on parameterization and optimization procedure of low-Reynolds-number airfoils based on genetic algorithm and Bezier curve. *Advances in Engineering Software*, *149*, p.102864.

[14] John H, Holland (1975). Adaptation in natural and artificial systems. The U. of Michigan Press

[15] John H, Holland. (1992). Genetic Algorithms. Scientific American 267, no. 1. 66–73

[16] Chen, X. and Agarwal, R.K., 2013. Optimization of wind turbine blade airfoils using a multi-objective genetic algorithm. *Journal of Aircraft*, *50*(2), pp.519-527.

[17] Liu, J.L. and Chen, C.M., 2009. Improved intelligent genetic algorithm applied to long-endurance airfoil optimization design. *Engineering Optimization*, *41*(2), pp.137-154.

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 05 | May 2023 www.irjet.net p-ISSN: 2395-0072

Bibliometric review of application of blockchain in supply chain management

Akash Mishra[1], Ayush Saini[2], Ansh Jini[3], N. Yuvraj[4]

[1]Student, B.Tech (Mechanical Engineering), Delhi Technological University, New Delhi, India
[2] Student, B.Tech (Mechanical Engineering), Delhi Technological University, New Delhi, India
[3] Student, B.Tech (Mechanical Engineering), Delhi Technological University, New Delhi, India
[4]Associate Professor, Dept. of Mechanical Engineering, Delhi Technological University, New Delhi, India

---

Abstract - *This paper analyzes the current state and potential of blockchain technology in supply chain management through bibliometric analysis. The study employs bibliometric data-driven analysis using the Web of Science and Scopus databases to collect relevant literature from 2013 to 2023. The Vos Viewer software is used to visualize the collaborative relationships and themes among the sampled literature. The co-word analysis reveals that the development of blockchain technology is centered on supply chain management, traceability and transparency, supply chain financing, and sustainability. The study highlights the potential of blockchain technology in facilitating developed supply chains, and its combination with the Internet of Things, that can drive significant changes across various industries and supply chains. This paper provides valuable insights for academic researchers and industrial managers in adopting blockchain to improve supply chain management performance. The study identifies the technical features of blockchain that can address challenges in SCM and highlights its potential to improve consumer trust, secure transactions, and efficient management. However, technical limitations and the need for ongoing research and development remain challenges for large-scale implementation.*

Key Words: **Supply Chain Management, Blockchain Technology, Smart Contract, Vos Viewer Software, Scopus, IOT, financing, sustainability**

1. INTRODUCTION

Efficient supply chain management practices are critical for boosting performance and efficiency, but a fundamental challenge remains the lack of trust across various parties[1]. Blockchain technology, with its openness, visibility, and disintermediation qualities, offers enormous promise for addressing security issues in the trade process and building mutual confidence among all supply chain participants. However, its present supply chain use is restricted, and additional study is required to incorporate it into e-commerce supply chain management.

Blockchain has been widely adopted in many areas of life and industrial sectors in China as a strategic frontier technology. It efficiently tackles the issue of information opacity and asymmetry while ensuring the confidentiality of operational data[2]. While there are several studies on blockchain industrial applications, there has been minimal study on its integration into e-commerce supply chain management.

This paper illustrates bibliometric analysis to witness how blockchain technology has been used into supply chain management. We gathered academic articles on blockchain and supply chain management that had been published internationally using Scopus, a bibliographic database. The search returned 19,062 research articles with the keywords "blockchain" and "supply chain management," all of which were written between 2016 and 2023.

The paper analyses how blockchain technology has the potential to improve supply chain management, notably in the areas of payment and supply chain management. The Vos Viewer software was used to build visualizations of the linkages and themes in the gathered literature on blockchain-powered supply chains.

The analysis discovered that the number of blockchain studies was modest in the years after 2016, but that the number of relevant studies increased substantially from 2018 to 2022. This increase was most likely caused by the advent of new data connected to the properties of blockchain technology, such as the Internet of Things and big data. Furthermore, extensive blockchain research and implementation are related with the introduction of new technologies.

A co-word analysis diagram was developed using 4,106 keywords to comprehend the new advancements in the domains being addressed and to uncover new links between pre-existing scientific fields. According to the findings, blockchain technology development is centred in the areas of supply chain management, traceability and transparency, supply chain financing, and sustainability.

The methodology and study strategy of a bibliometric analysis examining the integration of blockchain technology in supply chain management are discussed in this paper. The research on blockchain technology is also centered on the sectors of supply chain management, traceability and transparency, supply chain financing, and sustainability.

2. Theoretical Background

A digital transaction ledger called a blockchain allows numerous parties to safely and openly share and access data without the need for middlemen. Blockchain technology allows users to validate and add transactions to a safe, impenetrable database.

2.1 Blockchain Technology

Transactions in a blockchain are organized into blocks and added to the ledger in chronological order. Each block contains a one-of-a-kind code known as a "hash" that uniquely identifies the block and all its transactions [3]. This hash also includes the hash of the preceding block in the chain, forming a chain of blocks and giving rise to the term "blockchain."

One of the key benefits of blockchain technology is its decentralized nature. The blockchain, rather being depending on a central authority or middleman to run the network, is controlled by a distributed network of nodes that collaborate to validate transactions and maintain the ledger [4]. Because of this decentralization, the blockchain is more resistant to hacking and fraud, while simultaneously increasing transparency and trust between parties.

Numerous uses for blockchain technology are available, including voting systems, supply chain management, cryptocurrencies, and more. It offers a secure and effective method of storing and moving data without the involvement of middlemen, which has the potential to change how we interact and conduct business. Blockchain technology, however, also has drawbacks and restrictions.

2.2 Integration of Blockchain in Supply Chain

As blockchain technology continues to expand, it is finding its way into various industries, and the supply chain sector is not left out. Blockchain integration in supply chain management is becoming increasingly popular as it provides a way to improve transparency, accountability, and security in the supply chain.

Blockchain technology is essentially a decentralized digital ledger that records and stores data in a secure and transparent way. When integrated into the supply chain, it helps to create a decentralized system that allows all stakeholders in the supply chain to have access to the same information simultaneously, without the need for intermediaries. One of the key benefits of blockchain technology in the supply chain is the ability to increase transparency. With blockchain, all participants in the supply chain can view the entire transaction history of a product, from the manufacturer to the end consumer. This level of transparency helps to prevent fraud, counterfeiting, and other illegal activities that can occur within the supply chain[5].

Another significant benefit of blockchain technology in the supply chain is improved traceability. By using blockchain, manufacturers and suppliers can easily track products from their origin to their destination, providing customers with more information about the product's quality and authenticity.

Additionally, blockchain technology also offers improved security. By using cryptography, blockchain can ensure that all transactions are secure and tamper-proof, which reduces the risk of cyberattacks and data breaches. Several companies have already implemented blockchain technology in their supply chain management[6]. For example, Walmart, one of the world's largest retailers, uses blockchain technology to track its food supply chain. The system allows the company to quickly trace the origin of contaminated food products and remove them from store shelves before they cause any harm to customers.

Another example is Maersk, the world's largest container shipping company, which partnered with IBM to develop a blockchain-based platform called Trade Lens. The platform enables all stakeholders in the supply chain to access real-time information about the location of their shipments, reducing the risk of delays, theft, and other disruptions.

In conclusion, the integration of blockchain technology in the supply chain sector provides several benefits, including increased transparency, traceability, and security. Companies that have already adopted this technology are reaping the benefits of reduced costs, improved efficiency, and enhanced customer satisfaction. As more companies continue to integrate blockchain into their supply chain management, we can expect to see even more advancements in this field.

2.3 Hash Algorithm

Blockchain is a decentralized digital ledger that securely and transparently records transactions. The hash function, which is responsible for transforming data into a fixed-length value known as a hash, is a key component of blockchain technology. The hash function is an important feature of blockchain because it protects data integrity by making it impossible to change the data without changing the hash value.

A hash function takes any size input and returns a fixed size output. Because the output, or hash, is unique to the input data, even minor changes in the input data result in an entirely new hash value. The hash function is supposed to be a one-way function, which indicates that determining the original input data from the hash result is very difficult[7].

In blockchain, the hash function is used to generate a hash for each block, which is subsequently put to the block's header. The block header provides information such as the block's timestamp, the preceding block's hash, and the

current block's hash. By connecting the blocks with the hash function, a tamper-resistant chain of blocks is formed.

The hash function is also used to verify the validity of blockchain transactions. Each transaction is assigned a one-of-a-kind identification known as a transaction hash, which is generated by applying the hash function to the transaction data. The hash value of the transaction is then added to the blockchain, ensuring that the transaction cannot be changed without modifying the hash value.

To summarize, the hash function is an important part of blockchain technology since it maintains the data's integrity and security. It is used to generate a unique identifier for each block and transaction on the blockchain, making tampering with the data almost difficult. The hash function is an essential component in the creation of a decentralized and secure digital ledger system.

2.4 Smart Contract

Smart contracts have been around for a long time, but their use in industrial practices has grown in popularity with the arrival of blockchain technology[8]. A smart contract is a script recorded on a blockchain that executes automatically without the intervention of the contract's signer. It is highly secure and does not require middlemen, making it a suitable option for a wide range of sectors like insurance, healthcare, and smart cities.

Blockchain-based smart contracts can be used in the insurance business to handle and analyze pay-as-you-go vehicle insurance data. Smart contracts can help the healthcare business by assuring safe access to electronic health information. Smart contracts may also be used to promote interoperability between institutions and national healthcare delivery capacities, as well as to secure healthcare privacy[9]. Smart contracts in the supply chain sector can shorten cash flow cycles and minimize supply chain operating risks in traditional supply chain finance models. Smart contract transaction information is commonly employed in this industry because to its benefits such as traceability and irreversibility.

Smart contracts are also being used in the agricultural supply chain to boost transaction reliability and operational efficiency. Furthermore, smart contracts may be employed in smart city contexts to perform autonomous distributed administration of community electricity grids and smart metre technologies[10].

Overall, smart contracts provide a highly safe and efficient method of conducting transactions without the use of middlemen, making them a suitable option for a wide range of sectors. With the continuous advancement of blockchain technology, smart contract use cases are expected to evolve and spread even more.

3. Research Design

The systematic review analysis is performed in this study by employing a bibliometric analysis. This part of the paper presents the research design and methodology of this study, and the idea of application of the blockchain technology in the area of supply chain management has been profoundly explored and discussed to help to better understand this budding integration of the two massive areas of study.

3.1 Data Collection

A data of all research papers published across the world on our topic of study on integration of blockchain and supply chain management were collected via Scopus. For the purpose of research, previous publications concerning blockchain and supply chain management were gathered as samples. Scopus is a crucial tool for accessing global academic information. In order to gain a deeper understanding of the development trends of blockchain in the field of supply chain management, this study utilized the core database of Scopus. Scopus is a bibliographic database and a tool for accessing global research literature in various disciplines, including natural sciences, social sciences, engineering, and health sciences. The search for relevant research papers published between 2016 to 2023 utilized the keywords "blockchain" and "supply chain management", yielding a total of 19,062 research papers. In the next section these research papers will be analyzed thoroughly in order to drive key insights and share areas of budding interest and research.

3.2 Bibliometric Study

Bibliometric analysis is a quantitative approach used to analyze and evaluate scholarly publications and their impact. It involves the statistical analysis of bibliographic data, including citation counts, authorship patterns, and publication trends, among others. Bibliometric analysis is commonly used in the evaluation of research productivity, impact, and collaboration, as well as in identifying emerging research trends and gaps. Bibliometric indicators such as the h-index and impact factor are widely used to measure research impact and inform decision-making.

Therefore, this paper tries to disclose the development trend of blockchain technology and its application in supply chain management by employing a bibliometric analysis. The complicated interactions of objective items could be depicted and reflected through the similarity definition realized by the VosViewer mapping technology. VosViewer enables us to explore the structure and characteristics of the research field or a set of publications through bibliometric maps, which show relationships between authors, institutions, keywords, and other bibliographic elements. VOSviewer allows us to create maps based on co-citation analysis, co-authorship analysis, and keyword analysis, among others. The purpose of the VosViewer technology is to reflect the similarity

between each pair of items in a two-dimensional space based on their spacing distance as accurately as possible.

3.2.1 Co Word Analysis

Co-word analysis is a bibliometric method used to identify and analyze the relationships between words or concepts within a set of documents. Co-word analysis can be used to map the structure of a research field or to identify emerging trends and topics. It involves several steps, including selecting a relevant set of documents, identifying the key concepts or keywords in those documents, creating a co-occurrence matrix that counts the frequency of pairs of words, and using various statistical and visualization techniques to analyze the relationships between the words. To help better understand this topic, the VosViewer analysis is employed to draw a keyword co-occurrence network diagram to explore the evolution of blockchain technology, contributing to summarizing the overall development trend and future research opportunities

3.2.2 Network Analysis

Network analysis is a method used to study the relationships between entities in a network or system. It involves analyzing the connections between nodes or entities, to identify patterns, trends, and important features of the network. The analysis typically involves identifying the nodes and links between them, and then measuring various properties of the network, such as centrality, clustering, and connectivity. Centrality measures the importance of a node within the network, while clustering measures the degree to which nodes tend to cluster together in subgroups. Connectivity measures the overall connectedness of the network, such as the number of links between nodes. Network analysis in our research involves creating visual representations of the network, using VosViewer, to help identify patterns and trends. Through the network analysis, we can judge the diversified development of the application of blockchain in different fields. Furthermore, through the co-occurrence matrix of high-frequency feature words construction, the development trend and research hotspots can be figured out.

3.2.3 Research Framework

This paper performed and bibliometric analysis by addressing the collected literature on blockchain and supply chain management from 2016 to 2023. According to research, the use of blockchain technology has the potential to facilitate developed supply chains, particularly in the areas of payment and supply chain management. The study employed bibliometric analysis with VosViewer software to reveal the intricate visualizations and interconnections of blockchain-powered supply chains. The research results and themes are then summarized.

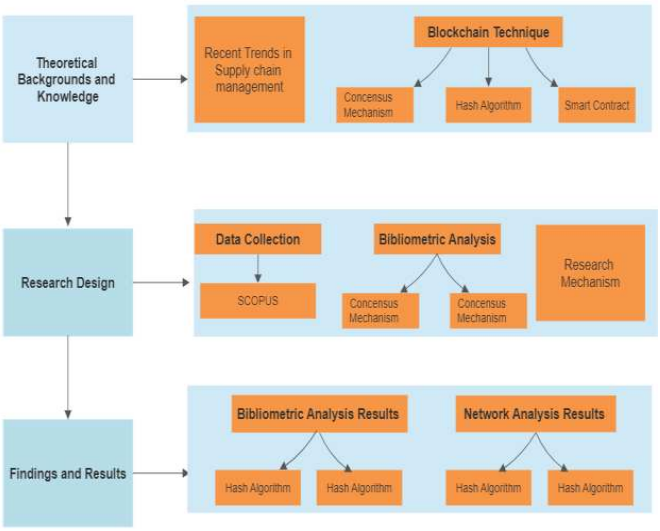


Fig 3.1: Research Framework

4. Findings and Results

4.1 Overview Analysis Results

4.1.1 Source Level Analysis

To examine academic research on blockchain, we searched for related papers in the core database of Scopus using the keywords "blockchain," and "supply chain management[13]." This bibliometric analysis collected 21,148 references. Our findings indicate that academic research on blockchain began in 2016 and has steadily increased until 2022, as illustrated in Figure 2. Although the number of studies on blockchain was minimal in the following years after 2016, the relevant studies have grown rapidly since 2018, with the most rapid growth occurring from 2018 to 2022. We can speculate that the emergence of new data, such as the Internet of Things and big data, which are related to blockchain technology's characteristics, has led to blockchain's development. Additionally, we found that the widespread research and application of blockchain are associated with the emergence of new technology.

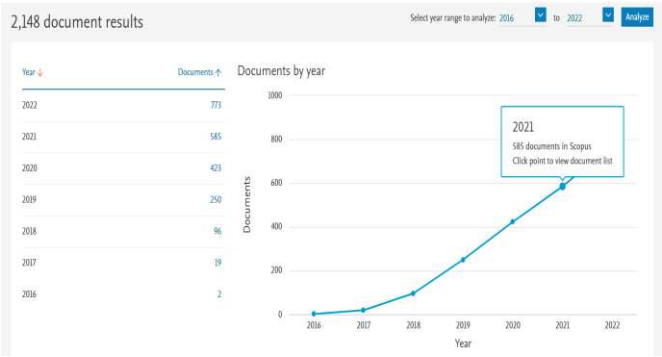


Fig 4.1: Year Wise Research Publications

Figure 4.2 presents that the top 10 countries in blockchain research based on the number of publications. India exhibits its predominance in production and influence with 446 papers, followed distantly by the USA and China. USA has less than half as many publications as India. The worldwide distribution of blockchain publications shows that academic research on blockchain is being published and pursued aggressively in all regions of the world.

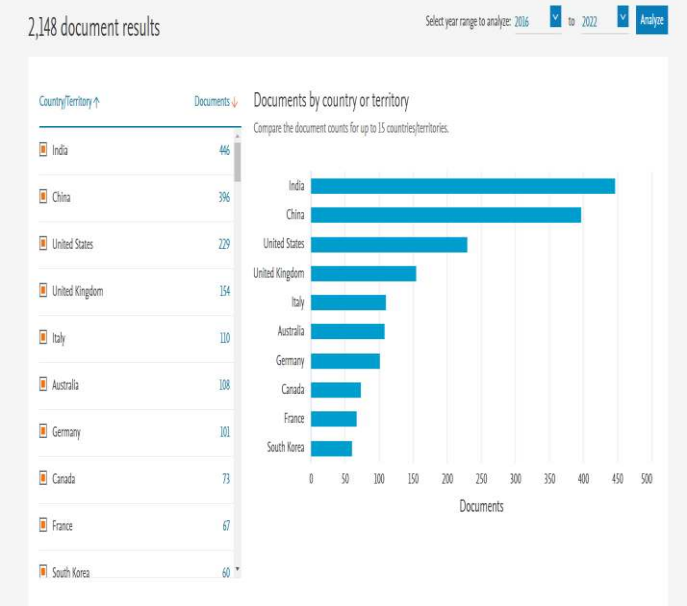


Figure 4.2: Country Wise Research Publications

This figure shows the country wise distribution of research papers published per country since 2016 till 2022 on the overlap of topics 'Supply Chain Management' and 'Blockchain'.

This graph depicts the countries in the world who are heading the innovations and research in this unique overlap of these big fields of science. Moreover, it depicts how the top 3 countries' researches are easily outnumbering the other countries quantitatively.

All of these key insights were fetched using the database of scopus which provided us with the data of these publications vis a vis their countries and publication year.

According to Figure 4.3, China leads in blockchain research with a total link strength of 214, followed by the United Kingdom with a total link strength of 191. Other countries such as India, Australia, USA, and France also show a significant interest in blockchain technology. This global attention and investment in blockchain technology suggest its potential to revolutionize various industries.

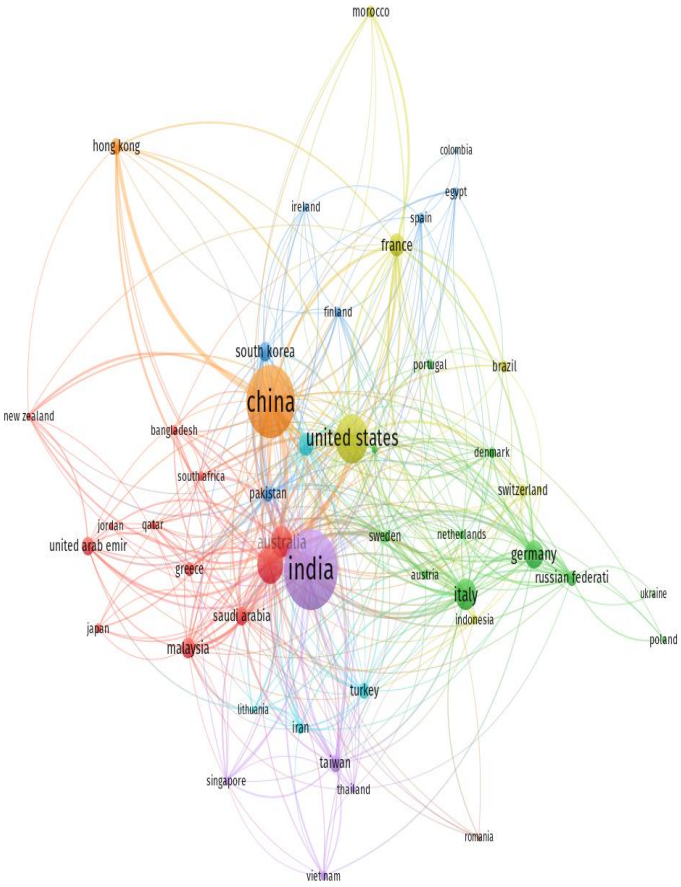


Figure 4.3 : Analysis of Cooperation in different countries on the topic under discussion

As a result, numerous countries are prioritizing its development as a strategic technology for the future, particularly China and the United Kingdom.

4.2 Network Analysis Results

4.2.1 Literature Analysis

Table 1 presents the top 5 high cited papers in blockchain research. According to the highly cited papers, the mainstream research direction was explored thoroughly. Most of these articles discuss the application of blockchain and AI in various aspects of supply chain management. Blockchain has great development prospects in supply chain, agriculture, and other fields. The role of blockchain is examined in achieving key Supply chain management goals for enterprises by establishing a framework. The combination of blockchain and the Internet of Things will drive major changes across multiple industries and their supply chains, as well as the development of new business models and distributed applications.

The following table shows us the most cited papers in the domain being studied.

Item	Journal Name	Theme	Total Citations
1.	International Journal of Production Research 57(7), pp. 2117-2135	Blockchain technology and its relationships to sustainable supply chain management	1250
2.	Telematics and Informatics 36, pp. 55-81	A systematic literature review of blockchain-based applications: Current status, classification, and open issues	967
3.	International Journal of Information Management 39, pp. 80-89	Blockchain's roles in meeting key supply chain management objectives	859
4.	International Journal of Production Research 57(3), pp. 829-846	The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics	704
5.	2018 IoT Vertical and Topical Summit on Agriculture - Tuscany, IOT Tuscany 2018 pp. 1-4	Blockchain-based traceability in Agri-Food supply chain management: A practical implementation	569

Table 4.1: Top 5 most cited papers

4.2.2 Co-Word Analysis

Using the 4,106 keywords gathered by the authors, we created a co-word analysis diagram depicted in Figure 5. Co-word analysis is a content analysis technique that can proficiently showcase how keywords co-occur in datasets and expose the structural network relationships. This helps us understand the new developments in the fields being discussed as well as find the new linkages between pre-existing scientific fields. This gives us a clear understanding on the thought and ideas of existing researches in the discussed fields[14]. The co-word analysis is based on the proposed formula T = (-1 + sqrt(-1 + 8I))/2 where I1 represents the number of words with frequency 1, and T denotes the minimum frequency value of high frequency words. The T index is used to distinguish the high frequency and low frequency words, and the T value calculated in this study is 41. Therefore, we chose keywords with more than 41 occurrences as the analysis content to ensure the effectiveness of the network diagram. VosViewer utilizes the VOS clustering algorithm for conducting clustering analysis. To capture the prominent areas of research, only nodes with a frequency of occurrence exceeding 41 were considered for clustering, resulting in the identification of three distinct clustering groups. Figure 5 depicts the nodes in proportion to their frequency of occurrence, with larger nodes indicating higher frequencies. Each node is color-coded to reflect the research hotspot at a particular stage, while the connecting lines depict the knowledge flow transitioning from purple to yellow, representing the chronological order.

Cluster 1 pertains to the use of blockchain technology in diverse industrial sectors such as big data, AI, logistics, monitoring, and supply chain. The decentralized and immutable nature of blockchain enables its potential for broad applications in several fields. For instance, blockchain enables transparent tracking of logistics processes, thereby allowing both businesses and consumers to access real-time information, ultimately improving information flow efficiency. Additionally, blockchain facilitates the traceability of products in the supply chain, ensuring confidentiality for all parties involved. As a result, blockchain technology helps to establish a robust network with a complete trust mechanism in the supply chain.

Cluster 2 shows the in depth analysis of the use of blockchain technology in the varied fields and the technology behind it. It talks about the concept of smart ledgers, consensus mechanism and the entire tech side of blockchain and its application. It also talks about the integration of game theory and blockchain in order to enhance the efficiencies of modern supply chains. These parts of this cluster are simply informing us about all the technologies and theories that are being discussed under the conjunction of these topics. This gets us to pace with all the research being carried out from 2016.

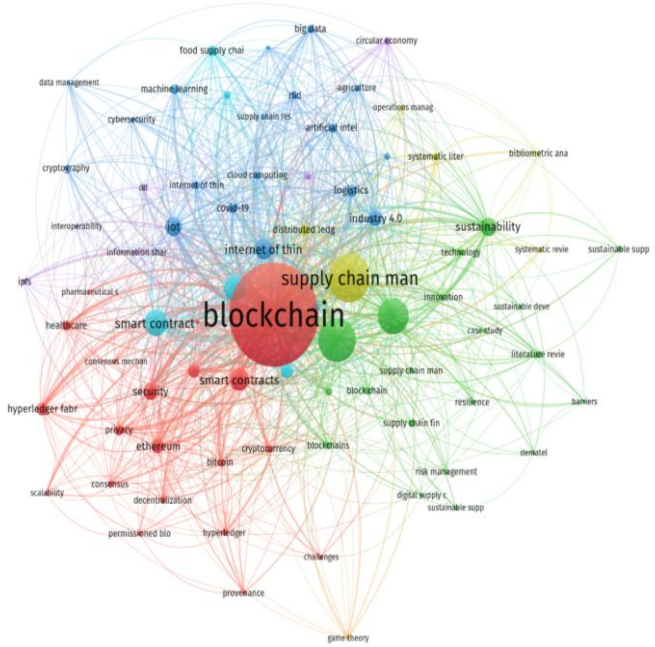


Figure 4.4: Co-occurrence of keywods by author-keywords

Cluster 3 lastly discusses sustainability and other case studies on the application of blockchain to supply chain management discussing the execution and implementation side of things. It talks about the application of blockchain and big data in food and high value goods supply chain. It also discusses the risk management aspect of the supply chains as blockchain will be a big driving force to making these

supply chains risk averse and safe for all of its beneficiaries. The logistics part of this amalgamation is being discussed in this cluster. Moreover the talks and moves towards making these supply chains more sustainable has also been discussed.

Our findings indicate that most of the research conducted between early 2019 and April 2019 predominantly focused on blockchain technologies, specifically exploring consensus mechanisms and smart contracts. Subsequently, from April to August 2019, research interest shifted towards the integration and implementation of other technologies, such as the Internet of Things, supply chain management, personal privacy protection, cloud computing, data sharing, and traceability. In 2020, the research focus was on exploring the industrial applications of blockchain in diverse sectors, including but not limited to personal analysis, services, resource management, trust management, smart cities, public key, and Industry 4.0.

4.3 Research Themes

Based on the bibliometric analysis of blockchain development trends, it is evident that research on blockchain technology is particularly concentrated in the domains supply chain management, traceability and transparency, supply chain financing, and sustainability. This section aims to examine the research trends of blockchain technology in these topics.

4.3.1 Traceability and Transparency in Supply Chains

Supply chain management involves the coordination of various activities and partners involved in the production, transportation, and delivery of goods and services. Ensuring transparency and traceability in supply chains is crucial for improving the efficiency, accountability, and sustainability of the supply chain. Blockchain technology, with its inherent characteristics of decentralization, immutability, and transparency, can be used to provide a secure and transparent platform for supply chain management. By utilizing blockchain technology, businesses can create a digital ledger of all transactions and activities involved in the supply chain, from the point of origin to the destination.

With blockchain technology, every participant in the supply chain can have access to the same information, and every transaction and activity can be recorded in real-time. This enables stakeholders to have complete visibility of the supply chain, including the origin of goods, their movement throughout the supply chain,and their final destination. Transparency in supply chains provided by blockchain technology can help businesses identify and address inefficiencies and bottlenecks in the supply chain. It also promotes ethical practices and accountability by enabling businesses to track the environmental and social impact of their supply chain activities. By maintaining a transparent and traceable supply chain, businesses can also mitigate risks associated with counterfeiting, fraud, and product recalls. Blockchain technology can also facilitate the implementation of end-to-end traceability in supply chains. End-to-end traceability involves tracking the entire journey of a product from its origin to the end consumer. This enables businesses to monitor the quality of the product, ensure compliance with regulations, and provide consumers with information on the origin and safety of the product.

End-to-end traceability provided by blockchain technology can be particularly beneficial in industries such as food and pharmaceuticals, where safety and quality are of utmost importance. By utilizing blockchain technology, businesses can trace the origin of raw materials, monitor the production process, and track the distribution and sale of the final product.

4.3.2 Supply Chain Financing Using Blockchain

Supply chain financing refers to the practice of providing financial solutions to businesses that are involved in the supply chain process. The goal of supply chain financing is to improve the efficiency of the supply chain by optimizing cash flow and minimizing the risks associated with the supply chain. One technology that has the potential to revolutionize supply chain financing is blockchain. Blockchain technology is a decentralized, distributed ledger that allows for secure and transparent transactions without the need for intermediaries. It provides a way for businesses to share information and conduct transactions in a secure and efficient manner, without the need for a central authority to verify and process the transactions. This makes it an ideal technology for supply chain financing, where there are many parties involved in the process and a need for secure, transparent transactions.

- → One of the main benefits of using blockchain for supply chain financing is that it can improve transparency and reduce the risk of fraud. By using a decentralized ledger to record transactions, all parties involved in the supply chain can have access to the same information in real-time. This allows for greater visibility into the supply chain process and can help to prevent fraud by identifying and addressing any discrepancies or errors in the supply chain.

- → Another benefit of using blockchain for supply chain financing is that it can improve the efficiency of the process by reducing the need for intermediaries. Currently, many businesses involved in the supply chain process rely on banks and other financial institutions to provide financing solutions. However, with blockchain, businesses can directly interact with each other, without the need for intermediaries. This can help to reduce costs and streamline the supply chain process. Overall, the

use of blockchain for supply chain financing has the potential to revolutionize the supply chain industry. By improving transparency, reducing the need for intermediaries, and providing greater security, blockchain can help to streamline the supply chain process and make it more efficient and cost-effective for businesses.

4.3.3 Sustainability in Supply Chain using Blockchain

Sustainability has become an increasingly important focus for businesses across the globe, and the supply chain is no exception. Supply chain sustainability refers to the practice of creating environmentally and socially responsible supply chains that are able to meet the needs of the present without compromising the ability of future generations to meet their own needs. Blockchain technology has the potential to play a significant role in enhancing the sustainability of supply chains in the following ways:-

→ by providing greater transparency and traceability. Blockchain enables the recording of data in a tamper-proof manner, meaning that once data is recorded, it cannot be altered or deleted. This makes it an ideal technology for tracking the movement of goods and materials through the supply chain, from their origin to their destination. By providing greater transparency and traceability, blockchain can help to ensure that products are sourced and transported in an environmentally and socially responsible manner. For example, a blockchain-based platform could be used to track the movement of goods and materials from their origin to their destination, providing real-time visibility into the entire supply chain. This could help to identify any potential sustainability issues, such as excessive use of resources, unethical labor practices, or environmental damage. By identifying these issues early, businesses can take proactive steps to address them and ensure that their supply chains are more sustainable.

→ Another way in which blockchain can support supply chain sustainability is by providing greater accountability. By recording data in a tamper-proof manner, blockchain can help to ensure that businesses are held accountable for their sustainability practices. For example, a blockchain-based platform could be used to record information about a business's sustainability practices, such as their use of renewable energy sources or their efforts to reduce waste. This information could then be shared with stakeholders, such as customers and investors, providing greater transparency and accountability.

→ Finally, blockchain can also support supply chain sustainability by enabling the creation of decentralized networks. Decentralized networks are networks that are not controlled by a single entity, but rather are governed by a consensus of all participants. This can help to promote greater collaboration and cooperation between businesses in the supply chain, enabling them to work together to address sustainability challenges. For example, businesses could use a blockchain-based platform to share information and resources related to sustainability, such as best practices for reducing carbon emissions or strategies for reducing waste. In conclusion, blockchain technology has the potential to significantly enhance the sustainability of supply chains.

By providing greater transparency and traceability, enabling greater accountability, and promoting greater collaboration and cooperation, blockchain can help to ensure that supply chains are more environmentally and socially responsible. As businesses continue to prioritize sustainability, we can expect to see an increasing use of blockchain technology in the supply chain industry.

5. CONCLUSIONS

- This paper provides a comprehensive analysis of the existing literature on the use of blockchain in supply chain management, highlighting its potential to address challenges in this area.

- The research hotspots in recent years mainly focused on the Internet of Things, supply chain, intelligent community, cloud computing, chemical industry, aviation, and other fields. In addition, we focused on analyzing the research trends of blockchain technology in supply chain management domain mainly pertaining to financing, traceability, transparency and sustainability. We also found that application of blockchain in these areas can help resolve issues pertaining to consumer trust, awareness, secure transactions, cost savings and efficient management

- In addition, this study also reveals some theoretical enlightenments and practical significances. The blockchain technology can help supply chain to carry out flexible management and efficient allocation of resources by innovative man-agreement practices.

6. REFERENCES

[1] Hsu, C.-C., Li, Y.-C., & Chen, M.-C. (2019). A blockchain-based approach to enhancing supply chain network efficiency. Journal of Industrial Information Integration, 15, 27-34.

[2] Bai, Y., Lu, Y., & Liang, C. (2018). Blockchain-based data sharing and collaboration in supply chain management. International Journal of Production Research, 56(1-2), 431-447.

[3] Wu, Z., Zhao, Y., & Luo, X. (2020). A bibliometric analysis of blockchain research in supply chain management. Sustainability, 12(9), 3864.

[4] Zhang, Y., Zhang, C., Yang, M., & Zhu, Y. (2019). A novel blockchain-based integrity verification system for medical data. Journal of Medical Systems, 43(7), 173.

[5] Wang, X., Liu, Y., Wang, H., & Li, K. (2019). Blockchain-based secure and efficient data sharing for supply chain management. IEEE Transactions on Industrial Informatics, 15(1), 354-362.

[6] Liu, Y., Liu, H., Li, W., & Chen, M. (2018). Blockchain-enabled secure and dynamic supply chain management. IEEE Transactions on Engineering Management, 65(2), 240-252.

[7] Biryukov, A., Khovratovich, D., & Pustogarov, I. (2016). Deanonymisation of clients in Bitcoin P2P network. Proceedings of the 25th USENIX Security Symposium, 1271-1288.

[8] Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2017). An overview of blockchain technology: Architecture, consensus, and future trends. IEEE International Congress on Big Data, 557-564.

[9] Domingo-Ferrer, J., & Martínez-Balleste, A. (2018). Blockchain and smart contracts for privacy-enhancing technologies. IEEE Security & Privacy, 16(4), 14-21.

[10] Khodabandelou, G., Abbasi, M. S., & Mirtaheri, S. L. (2019). A survey of smart contracts: Challenges, advances, and applications. Journal of Network and Computer Applications, 132, 95-113.

[11] Li, Z.; Zhong, R.Y.; Tian, Z.G.; Dai, H.-N.; Barenji, A.V.; Huang, G.Q. Industrial Blockchain: A state-of-the-art Survey. Robot. Comput. -Integr. Manuf. 2021, 70, 102-124.

[12] Feng, Y.; Zhu, Q.; Lai, K.-H. Corporate social responsibility for supply chain management: A literature review and bibliometric analysis. J. Clean. Prod. 2017, 158, 296–307.

[13] Esmaeilian, B.; Sarkis, J.; Lewis, K.; Behdad, S. Blockchain for the future of sustainable supply chain management in Industry 4.0. Resour. Conserv. Recycl. 2020, 163, 105064.

[14] Wang, C.; Lim, M.K.; Zhao, L.; Tseng, M.-L.; Chien, C.-F.; Lev, B. The evolution of Omega-The International Journal of Management Science over the past 40 years: A bibliometric overview. Omega 2020, 93, 102098

[15] S. Saberi, M. Kouhizadeh, J. Sarkis & L. Shen, "Blockchain technology and its relationships to sustainable supply chain management, International Journal of Production Research, Pages Vol. 57, Pages 2117-2135

[16] Rosanna Cole, Mark Stevenson, James Aitken "Blockchain technology: implications for operations and supply chain management", Emerald insight, ISSN: 1359-8546, Vol. 24 No. 4, pp. 469-483.

[17] Yulin Luo, Shenxing Xu, Q. Zhou, "Application of blockchain technology in supply chain management", ICIMTECH 21, Article No.: 205

[18] Pankaj Dutta, Surabhi Somani, Richa Butala & Tsan-Ming Choi, "Blockchain technology in supply chain operations: applications, challenges and research opportunities", Elsevier, 1366-5545, Volume 142

[19] Abderahman Rejeb, John G. Keogh & Horst Treiblmaier, "Leveraging the internet of things and blockchain technology in supply chain management", MDPI, Volume 11, Issue 7.

This article has been accepted for publication in IEEE Transactions on Multimedia. 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/TMM.2023.3325993

Bit-plane based Reversible Data Hiding in Encrypted Images using Multi-Level Blocking with Quad-Tree

Ankur, Rajeev Kumar, Ajay K Sharma, *Member, IEEE*

***Abstract*—Reversible data hiding in encrypted images (RDHEI) has gained significant popularity among security and privacy researchers as well as users, because of its features such as reversibility, embedding capacity (EC), and security. To enlarge the EC while ensuring the complete reversibility and security, we propose a bit-plane based RDHEI method based on multi-level blocking with quad-tree. The proposed method uses median edge detector (MED) as well as difference predictor to transform the original input image into a low-magnitude difference matrix. The difference matrix is then encoded by first employing a novel quad-tree based bit-plane representation strategy to exploit the intra-bit plane correlation and subsequently by inter bit-plane redundancy mitigation strategy to exploit inter bit-plane level correlation, for significantly condensing their size. Thus, a bigger room is reserved inside the cover image for embedding, so that a large amount of secret data can be hidden while ensuring the complete reversibility of the image. Experimental results validate the superiority of the proposed method over the state-of-the-art methods.**

***Index Terms*—Reversible Data Hiding, Encrypted Image, Bit-plane, Quad-Tree, MBQ, RDHEI.**

I. Introduction

The art of covert communication with the help of data hiding (DH) in images has been a popular alternative for secret communication because of its ability to undetectably embed data in a cover image [1], [2]. However, the process of conventional embedding may lead to permanent loss of the original cover image's contents, posing a serious issue on its applicability in a wide variety of domains such as medical, military, etc. To tackle this issue, reversible data hiding (RDH) techniques were introduced to enable both lossless extraction of the embedded data and complete recovery of the original image [3]–[6]. This ability of RDH methods makes them highly popular among security researchers [7].

To further enhance security in terms of confidentiality and privacy, RDHEI was introduced, which allows embedding in an encrypted image. In recent years, RDHEI has been the subject of numerous research works because of its wide applicability in the popular cloud domain [8]–[36]. The existing RDHEI methods are broadly classified into two categories: Vacating Room After Encryption (VRAE) and Reserve the Room Before Encryption (RRBE) based on the approach adopted for embedding. The VRAE methods [8]–[13] encrypt the original image first and then try to create a room for embedding, while the RRBE methods [15]–[21], [24]–[29], [31], [32], [34], [35] exploit the correlation of the original images by pre-processing them to reserve the room for embedding by exploiting the prevalent correlation, before its encryption. As a result, RRBE methods generally offer a much higher EC than VRAE methods.

As far as we know, the first high capacity RRBE method was discussed by Ma *et al.* [15] in 2013. This method involves creating a room in the image using histogram shifting, where the first three least significant bits (LSBs) are shifted from one region to another. Thus, the method achieves an average embedding rate (AER) of 0.5 bits per pixel (bpp), however, does not guarantee a perfect reconstruction of the original image. In 2015, Zhang *et al.* [16] discussed an RDHEI method that takes advantage of the Paillier crypto-system for image encryption and utilizes low-density parity-check coding for embedding. Though this method also achieves AER only up to 0.5 *bpp* approximately, but, keeps the distortion lower than the previous methods. However, the method may result in compromised data extraction and image recovery in case of exceedingly large payloads. To increase the EC, Cao *et al.* [17] exploited the redundancy present in the original image by compressing the pixels in the local patches. For this, a patch-level sparse representation is used to investigate the correlation between neighbouring pixels, allowing for efficient sparse coding and leaving enough space for the DH. Experimental results show an AER of about 1 *bpp*, outperforming other erstwhile methods. However, the PSNR decreases to approximately 40 dB for smooth images.

In 2017, Yi *et al.* [18] discussed a novel RDHEI method by capitalizing on the most significant bits (MSBs) of pixels, given their heightened predictability in natural images. The method exploits the redundancy in blocks through binary block embedding, however, fails to ensure optimal performance in every situation, as it solely investigates local correlation. In a similar vein, Puteaux *et al.* [19] introduced the first prediction-error (PE) based RDHEI method, which employs the MED to compute the PEs and perform embedding in the MSBs of the image pixels. Although the method attains an AER of approximately 1 *bpp*, the process of reconstructing the original image requires identifying non-embeddable MSBs through the use of the 8-bit flag, which may pose risks to data recovery. Building upon the work of [19], Puyang *et al.* [20] capitalized on neighbourhood correlation by employing two MSBs for

Ankur is with the Delhi Technological University, Delhi, India (e-mail: ankur_2k20phdco507@dtu.ac.in)

Rajeev Kumar is with the Delhi Technological University, Delhi, India (e-mail: rajeevkumar@dtu.ac.in)

Ajay K. Sharma is with the National Institute of Technology, Delhi, India (e-mail: director@nitdelhi.ac.in)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993

data embedding to increase the EC. However, the redundancy present in the remaining six bits of each pixel and their inter-pixel correlation is not considered. In 2020, Mohammadi *et al.* [21] investigated redundancy in small blocks and performed embedding into multiple MSBs after calculating prediction errors (PEs) using a difference predictor. This method yields an AER of approximately 2 *bpp*. Nevertheless, the method falls short of effectively capitalizing on the correlation among the blocks. Similarly, a method proposed by Ren *et al.* [22] explores the redundancy present within small blocks. Diverging from [19], [21], Puteaux *et al.* [23] eschewed label maps for non-embeddable pixels, opting instead to demarcate the predicted error list from the embeddable space with an end mark, thereby achieving reversibility and avoiding the flags. In 2022, Yu *et al.* [24] introduced a hierarchical embedding based RDHEI method, wherein the prediction results generate a hierarchical label map for each bit-plane to increase the number of embeddable bits. Thus, the method achieves higher EC but still fails to fully exploit the image correlation present in small regions. In order to exploit the large redundancy of MSB-bit planes, Chen and Chang [25] first employed a block-based MSB bit-plane rearrangement technique to generate a more compressible bit-stream and then utilized the extended run-length encoding to compress them so that bigger room for embedding can be created.

On the other hand, Yi *et al.* [26] discussed an RDHEI method that first divides the image into blocks to capitalize on local correlations and then encodes the pixels using parametric binary tree labelling (PBTL) for reserving a room inside the image. Building on Yi *et al.*'s [26] PBTL-RDHEI, Wu *et al.* [27] made use of entire image spatial correlation to further enhance the EC by incorporating the MED predictor. On similar lines, Yin *et al.* [28] introduced a new method that first utilizes the MED to predict pixel values similar to [27] and then make bit-wise comparisons to calculate PEs which are then encoded using Huffman compression to reserve bigger room for embedding. To further increase the EC, Gao *et al.* [29] employed geometric transformation to reduce the size of the label map. However, the method yields limited EC in case the difference between predicted and original values is minimal, like in [28]. Gao *et al.* [30] employed multilinear regression prediction for pixel estimation and manage the absolute PEs using two levels of label-maps. Although accurate PEs can enhance the EC, the regression method notably prolongs the processing time.

Chen *et al.* [31] introduced a method that employs iterative MSBs-inversion prediction and adjacent prediction plane XOR. The prediction is made by taking a weighted average of the MSBs of neighbouring pixels and XOR the values of neighbouring pixels in a specific pattern while adding a constant offset. Subsequently, the bits of each block are compressed using block variable length coding. Thus, the method achieves an AER of 2.3 *bpp*. To increase the EC of Chen and Chang [25], Yin *et al.* [32] leveraged the sparsity of the PE's bit-planes. However, the method fails to utilize the residual redundancy due to the consistent block size, like [25]. Yang *et al.* [33] introduced an encryption method that shuffles the Huffman codes using timestamps, leading to weaker encryption but a substantial EC due to the direct implementation of Huffman encoding on PEs.

In 2021, Liu *et al.* [34] discussed a new RDHEI method using quad-tree to compress the bit-plane in different size blocks by assigning the code to each block with the help of four components, namely depth coding, value coding, number coding, and path coding. Thus, the method is able to save more space inside the cover image for embedding. To further increase the EC, Xu *et al.* [35] discussed a hierarchical block variable length coding (HBVLC) based RDHEI method by selecting the optimal compressible block size so that variable sparsity can be effectively captured. A more detailed explanation of the HBVLC is provided in the next subsection. Contrary to [35], Yao *et al.* [36] focused on discerning similarities within the LSB-planes of block, eschewing exploration of varied levels. This method leverages the lower LSB-plane for embedding, which aids in EC enlargement. However, the method does not consider the redundancy present among the bit-planes, which would have further increased the EC.

To address the aforementioned challenges of existing RDHEI methods, the proposed work introduces a new RDHEI method based on multi-level blocking with quad-tree (MBQ-RDHEI). The contribution of the proposed MBQ-RDHEI method can be summarized as follows:

- The proposed MBQ-RDHEI method employs MED to leverage the high correlation typically present among adjacent pixels of meaningful images, for generating a low magnitude difference image having high sparsity in MSB-planes.
- The bit-planes of the image are encoded by employing a novel quad-tree based bit-plane representation technique to condense their size by exploiting the sparsity. This process leads to the quad-tree generation for each bit-plane, where each node of the tree is associated with a variable-sized block of the image. Next, the nodes are mapped with binary bits and traversed in label order to generate the compressed bit-stream.
- In order to further reduce the size of bit-planes' representation, a novel inter bit-plane redundancy mitigation (IBRM) strategy is utilized to recursively reduce the sub-trees from the quad-tree by exploiting the inter bit-plane correlation, which ensures that the final bit-stream is compact and devoid of any redundant information. The resulting bit-stream is then embedded into the cover image.
- The experimental results indicate that the MBQ-RDHEI offers a significantly higher EC compared to the state-of-the-art (SOTA) methods while ensuring the perfect reconstruction of the original image.

The rest of the paper is structured as follows: Section II and III discuss the preliminary in the context of the proposed work and the proposed work itself, respectively. Next, Section IV presents the experimental results and their analysis. At last in Section V, the paper is concluded while giving directions to future work.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993

II. Preliminary

In this section, we will discuss some of the existing quad-tree inspired image compression techniques which include the original quad-tree compression [37], hierarchical quad-tree coding [34], and hierarchical block variable length coding [35], in detail. These techniques utilize the spatial correlation of the original images and divide the image/bit-planes into blocks for compression. The detailed compression process of each of the aforementioned techniques is explained with suitable examples in the following subsections.

A. Quad-tree Compression

In 1992, Markas and Reif [37] discussed an image compression method by decomposing the image into a hierarchical tree configuration, where each node of the tree signifies a specific square region/block of the image and possesses four distinct child nodes, hence the name "quad" tree. The process of image decomposition begins by determining whether the image block (which is initially the entire image itself), satisfies a predefined criterion (e.g., minimum threshold of the difference between pixel values, in the context of image compression) or not. If it does, its node is marked '$Leaf$', and no further subdivision is done as the process is assumed to be completed for that block, otherwise, its node is marked '$Internal$' and the image block is recursively divided into four smaller blocks until each sub-block satisfies the criterion or reaches a predetermined minimum size. Thus, the root node of the tree represents the entire image, and each of its child nodes represents a quadrant of the image. Next, each node of the tree is assigned a four-bit code signifying the number of child nodes it has. Additionally, each '$Leaf$' node is linked with a fixed size code to denote the average intensity of its corresponding block. To encode the image, the assigned codes are amassed following any traversal sequence, which will give a compressed bit-stream representing the complete image.

It is to be noted that the quad-tree encoding permits the use of larger blocks for representation in homogeneous areas. This adaptability is a crucial aspect of the compression process, as it allows these regions to be represented more compactly, thereby optimizing space usage. In other words, the quad-tree encoding essentially 'summarizes' these uniform areas into larger blocks, achieving significant data reduction and contributing to the overall efficiency of the compression.

B. Hierarchical Quad-tree Coding

In 2021, Liu *et al.* [34] introduced an RDHEI method based on hierarchical quad-tree coding (HQTC). The method encodes the cover image in a bit-plane wise manner starting from the MSB to the LSB, to create room for embedding. For this, all the bits of a given bit-plane are checked, if these bits are the same, or in other words, all the bits are either 0 or 1, then the complete bit-plane is encoded by the corresponding bit only as either $Leaf-0$ or $Leaf-1$, and no subdivision is executed like in II-A. Otherwise, the bit-plane is recursively subdivided into four sub-blocks till the aforementioned condition is not met for each sub-block. Next,

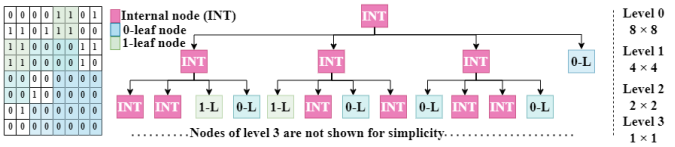


Fig. 1: Hierarchical quad-tree coding.

all the leaf nodes are coded with the help of four unique codes which include $depthcode$ representing the number of levels, $valuecode$ denoting the category of the leaf node ($Leaf-0$ or $Leaf-1$), $numbercode$ indicating the total number of leaf nodes, and $pathcode$ suggesting the route to the particular node. Finally, the codes of each leaf i.e., either $Leaf-0$ or $Leaf-1$, are collected for each meaningful level of the tree, to encode the complete image. Thus, a significantly big space for embedding is reserved.

In order to better elucidate the process of the HQTC, an illustrative example is presented in Fig. 1. The example considers an 8×8 bit-plane and generates its corresponding quad-tree structure. The tree is coded in a level-wise manner, starting with level 1, which contains only a $Leaf-0$. This leaf is represented by '01 0 01 11', where the blue-colored bits represent the level number or $depth$ code, the yellow-colored bit represents the type of $Leaf$ or $value$ code, the green-colored bits represent the one $Leaf-0$ node at the given level number or $number$ code, and the red-colored bits represent the path from the root node or $path$ code. Similarly, for the $Leaf-0$ and $Leaf-1$ of level 2, two bit-streams are generated as follows: '10 0 0100 0011 0110 100011' and '10 1 0010 0010 0100', respectively.

C. Hierarchical Block Variable Length Coding

In 2022, Xu *et al.* [35] discussed a HBVLC for RDHEI. The working of HBVLC seems to be inspired by the quad-tree structure, as the image (bit-plane) is partitioned into a 3-level hierarchical structure having adaptive block sizes. HBVLC initially considers the entire bit-plane of size $\tau\times\tau$ as a level-1 block and checks for the '1' bit distribution. If the distribution is greater than a defined threshold (t_1) which is calculated using Eq. 1, then the level-1 block is partitioned into four level-2 sub-blocks of size $\tau/2\times\tau/2$.

$$t_i = \lfloor(\tau_i^2 - 2)/\log_2\tau_i^2\rfloor, \tag{1}$$

where $1\leq i\leq 3$, and τ_i is the i^{th}-level block. Next, the distribution of '1' is checked in all the four level-2 blocks and found to be greater than a defined threshold (t_2) as stated in Eq. 1, for any of the sub-blocks then each of the four sub-blocks is divided into level-3 sub-blocks of size $\tau/4\times\tau/4$. Next, the distribution of '1' is checked in level-3 blocks and if found to be greater than a defined threshold (t_3) as given in Eq. 1, then no further partitioning is done but the bit-plane itself is considered as its code. However, the level-1 block is encoded using its level-label concatenated by the code of its sub-blocks, where the code of the sub-blocks includes a type-label followed by its structure information. Thus, HBVLC tries to keep the total code length of a block less than the direct

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993

recording of the bit-plane and achieves significant performance benefits as far as image compression is concerned.

III. Proposed RDHEI using Multi-Level Blocking with Quad-Tree (MBQ-RDHEI) Method

This section provides a detailed description of our proposed MBQ-RDHEI method, which aims to mainly improve the EC while ensuring reversibility and confidentiality of the cover image data. The proposed MBQ-RDHEI works in four stages that include 1) Room reservation, 2) Image encryption and supporting information embedding, 3) Data hiding, and 4) Data extraction with image recovery, as illustrated in Fig. 2.

A. Room Reservation

To reserve a room inside the original cover image (I) for embedding, I is first transformed into a difference matrix with the help of difference predictor and MED predictor [38], which is later converted into bit-planes. Next, the obtained bit-planes are represented into compressible bit-streams guided by quad-tree structure to condense their size. Afterwards, a novel IBRM strategy is employed to further optimize the bit-plane representation. Lastly, the supporting information, generated during the aforementioned process which is essentially an overhead, is prudently organized to lessen its impact on the created room. The detailed process to execute the aforementioned steps is provided as follows.

1) Image Transformation: The original gray-scale image I of size $R \times C$ is transformed into a difference matrix (D) of the same size i.e., $R \times C$ using the difference and Median Edge Detection (MED) predictors, both. This process of the transformation explores the redundancy of the input image I and can be executed by Eq. (2), defined below:

$$D_{(i,j)} = I_{(i,j)} - \begin{cases} I_{(i,j)}, & \text{if } i = 1 \ \& \ j = 1 \\ I_{(i,j-1)}, & \text{else if } i = 1, \\ I_{(i-1,j)}, & \text{else if } j = 1, \\ I_{med(i,j)}, & \text{otherwise} \end{cases} \tag{2}$$

where $D_{(i,j)}$ represents the difference matrix element corresponding to the original image pixel $I_{(i,j)}$ and I_{med} is determined using Eq. (3), $1 \leq i \leq R$ and $1 \leq j \leq C$.

$$I_{med(i,j)} = \begin{cases} max(B_p, U_p), & \text{if } O_p \leq min(B_p, U_p) \\ min(B_p, U_p), & \text{elseif } O_p \geq max(B_p, U_p), \\ B_p + U_p - O_p, & \text{otherwise} \end{cases} \tag{3}$$

where $B_p = I_{(i,j-1)}$, $U_p = I_{(i-1,j)}$ and $O_p = I_{(i-1,j-1)}$ denote back-pixel, upper-pixel, oblique-pixel, respectively, with respect to $I_{(i,j)}$. It is to be noted that the original image (I) can be reconstructed at the recipient end by having access to the difference image (D) and the value of the first pixel (i.e., $I_{(1,1)}$), which is chosen as the reference point. By designating the $I_{(1,1)}$ as the reference pixel (RP), the corresponding difference pixel $D_{(1,1)}$ can be initialized to zero. Since I is a gray-scale image, then $-255 \leq D_{(i,j)} \leq +255$ holds which in turn indicates that $D_{(i,j)}$ must be represented by 9 bits, calculated as follows:

TABLE I: Count of '0' Values in $b^{(k)}$ Bit-Plane

Test image	k=8	k=7	k=6	k=5	k=4	k=3	k=2	k=1	k=0
Lena	262144	262021	260296	250944	222282	154565	78493	29207	146894
Airplane	262139	261711	260217	257182	248793	222200	107790	103924	183771
Baboon	262140	258013	230106	178159	114437	60158	26465	8950	134330

$$b^{(k)}_{(i,j)} = \begin{cases} \begin{cases} 0, & \text{if } D_{(i,j)} < 0 \\ 1, & \text{otherwise} \end{cases}, & \text{if } k = 0 \\ \left\lfloor \frac{|(D_{(i,j)}| mod(2^k)}{2^{k-1}} \right\rfloor, & \text{Otherwise} \end{cases} \tag{4}$$

where $b^{(k)}$ denotes the k^{th} bit-plane for D and $8 \geq k \geq 0$. In other words, 8 bit-plane ($b^{(8)}$, $b^{(7)}$, ..., $b^{(1)}$) are required to store the magnitude of D and 1 bit-plane ($b^{(0)}$) for the sign-bit of each element of D. Due to prevalent spatial correlation within the natural input images, it has been commonly observed that the sparsity of a given bit-plane $b^{(k)}$ decreases with the value of k for magnitude bit-planes. This observation is also validated by the experimental finding presented in Table I for three different characteristics and frequently used test images (i.e., $Lena$, $Airplane$, and $Baboon$), where the number of '0' bits representing the magnitude of sparsity, keeps on decreasing with the bit-plane number. However, it can also be observed that the highest level of randomness is present in the sign bit-plane $b^{(0)}$ which makes the embedding a challenging task.

To demystify the process of the bit-plane generation, an illustrative example is presented in Fig. 3. The example considers an 8×8 image, as shown in Fig. 3a, and follows the rules for generating a difference matrix as depicted in Fig. 3b. The resulting difference matrix is shown in Fig. 3c, which is then separated into nine bit-planes, with the first eight bit-planes representing the magnitude and the last bit-plane representing the sign-bit. This separation of bit-planes is depicted in Fig. 3d.

2) Quad-Tree based Bit-plane Representation: To condense the size of the obtained bit-planes, each bit-planes ($b^{(k)}$) is converted into a compressed bit-stream ($s^{(k)}$) one by one, starting from the MSB to the LSB-plane. For this, a quad-tree $q^{(k)}$ specifically tailored for each individual bit-plane $b^{(k)}$ is generated adhering to the principles of a traditional quad-tree, as described in II-A. However, the partitioning of a block (or node) into four sub-blocks (or child nodes) is done only if any bit associated with the corresponding block is non-zero, such blocks are also termed as non-embeddable block or node (NB) and remaining ones as an embeddable block (EB).

To provide further insight into the quad-tree generation process, a bit-plane $b^{(k)}$ with dimensions $\tau \times \tau$ has been considered, where $\tau = 2^t$ and $t \in \mathbb{N}$. Fig. 4 provides a visual representation of a complete quad-tree $q^{(k)}$ for $b^{(k)}$. It shows that a $\tau \times \tau$ bit-plane can have up to $\log_2 \tau$ possible levels, with each level containing a maximum of 4^{r-1} nodes, where r ranges from 1 to $\log_2 \tau$. The step-by-step algorithm to generate a complete quad-tree ($q^{(k)}$) for bit-plane ($b^{(k)}$) is provided in Algorithm 1.

The presented Algorithm 1 generates a quad-tree starting with a $root$ node representing the entire bit-plane of dimen-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993

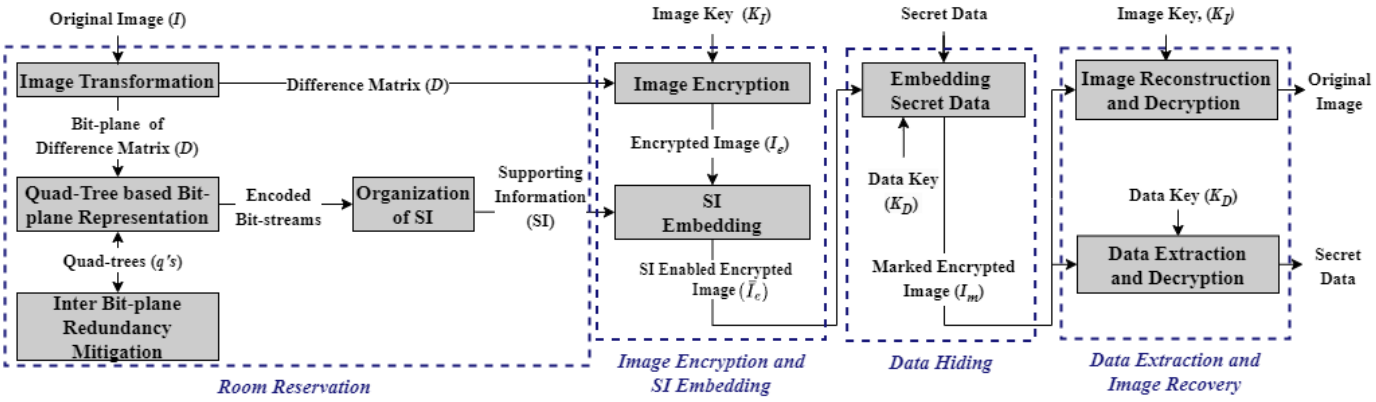
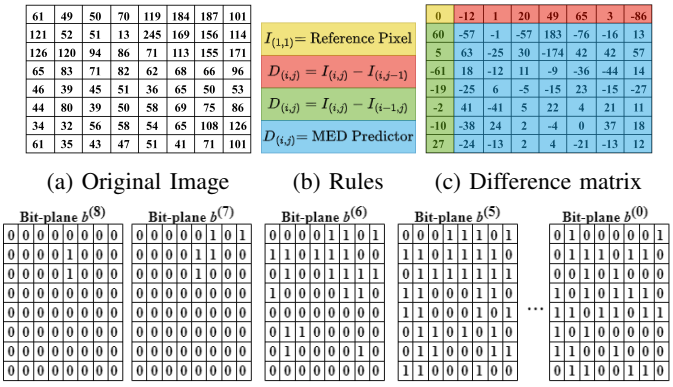


Fig. 2: Workflow of the proposed RDHEI using multi-level blocking with quad-tree method.



(a) Original Image (b) Rules (c) Difference matrix

(d) 9 bit-planes ($b^{(8)}$, $b^{(7)}$, ..., $b^{(0)}$) of difference matrix

Fig. 3: Illustration of the difference matrix's bit-planes creation process.

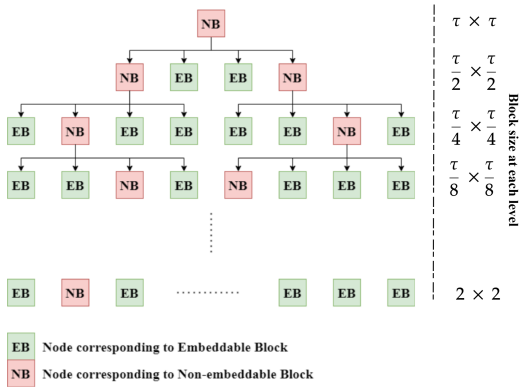


Fig. 4: Representation of full quad-tree.

Algorithm 1 Quad-tree Generation

function GENERATEQUADTREE(τ, bit-plane $b^{(k)}$)
  Create an empty $queue$ and a $root$ node representing the entire bit-plane $b^{(k)}$ of dimension $\tau \times \tau$ as a block
  $enqueue$ the $root$ into the $queue$
  while $queue$ is not empty **do**
    $dequeue$ the $front$ item from $queue$ and assign to $node$
    if the block represented by $node$ is not an EB and the size of the block is greater than 2×2 **then**
      Subdivide the block into four equal size quadrants/sub-blocks
      Create a child node for each quadrant/sub-block and $enqueue$ each child onto the $queue$
    end if
  end while
  Return the $root$ node of the quad-tree.
end function

sions $\tau \times \tau$ as a block using $queue$ as a data structure. The algorithm subdivides a block into four sub-block (i.e., top-left (top_{left}), top-right(top_{right}), bottom-left($bottom_{left}$), and bottom-right($bottom_{right}$)), if it is not an EB. This subdivision is recursively called until each sub-block becomes EB or its size reduces to 2×2. The resulting quad-tree thereby models the entire $\tau \times \tau$ bit-plane, where each EB signifies a sparse region and each NB denotes a complex one within the bit-planes. Thereafter, $q^{(k)}$ is traversed in level-order (also known as breadth-first search (BFS) order) to generate the compressed bit-stream $s^{(k)}$ for the bit-plane $b^{(k)}$, where '0' and '1' are collected for each EB and NB, respectively. The complete step-by-step algorithm to generate the compressed bit-stream ($s^{(k)}$) for $q^{(k)}$, is discussed in algorithm 2.

3) Inter Bit-plane Redundancy Mitigation: In order to further optimize the representation of bit-planes, a novel IBRM strategy is employed so that redundancy between the subsequent bit-planes' representations can be minimized. This is being done based on findings of a comprehensive experimental analysis conducted using three commonly used standard test images, namely, $Lena$, $Airplane$, and $Baboon$ with two consecutive bit-planes ($b^{(k)}$ and $b^{(k-1)}$) and their respective quad-trees ($q^{(k)}$ and $q^{(k-1)}$) along with compressed bit-streams ($s^{(k)}$ and $s^{(k-1)}$). As discussed in subsection III-A2, each bit in the bit-stream represents a block of size ranging from $\tau \times \tau$ to 2×2, where the bit '0' denotes EB and '1' represents NB.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993

Algorithm 2 Quad-Tree Encoding into Bit-Stream using Level Order Traversal

function ENCODEQUADTREE($root$, $b^{(k)}$)
  Create an empty $queue$ and empty bitstream $s^{(k)}$
  $enqueue$ the $root$ node into the $queue$
  while $queue$ is not empty **do**
    $dequeue$ the $front$ item from $queue$ and assign to $node$
    Determine the corresponding block of $b^{(k)}$
    if the block is an EB **then**
      Append '0' bit to $s^{(k)}$
    else
      Append '1' bit to $s^{(k)}$
      if the node corresponds to NB and is not a leaf node **then**
        $enqueue(node.top_{left})$ into the $queue$
        $enqueue(node.top_{right})$ into the $queue$
        $enqueue(node.bottom_{left})$ into the $queue$
        $enqueue(node.bottom_{right})$ into the $queue$
      end if
    end if
  end while
  return bitstream $s^{(k)}$
end function

TABLE II: PROBABILITY OF NB FOR $b^{(k)}$ WITH RESPECT TO SIMILARLY POSITIONED NB FOR $b^{(k-1)}$

Test image	$k=8$	$k=7$	$k=6$	$k=5$	$k=4$	$k=3$	$k=2$	$k=1$	$k=0$
Lena	1.0000	0.6436	0.7673	0.8199	0.8435	0.8975	0.9334	0.9070	-
Airplane	0.2500	0.6031	0.7018	0.7388	0.7177	0.9353	0.9157	0.8429	-
Baboon	0.0000	0.7104	0.8277	0.8742	0.9069	0.9273	0.9363	0.9435	-

From the findings calculated in terms of probability using Eq. (5) as provided in Table II, it has been observed that NB for $b^{(k)}$ can significantly impact the corresponding blocks in the subsequent bit-plane $b^{(k-1)}$.

$$p^{(k)} = \frac{\delta^{(k-1)}}{\theta^{(k)}}, \tag{5}$$

where $\theta^{(k)}$ denotes the total count of NBs present in $b^{(k)}$, and $\delta^{(k)}$ represents the total count of NBs in $b^{(k)}$ whose corresponding blocks in the subsequent bit-plane $b^{(k-1)}$ are also NBs.

Therefore, the proposed method marks all the corresponding blocks of the subordinate/subsequent planes (i.e., $b^{(k-1)}$ to $b^{(0)}$) as NBs of the NB of $b^{(k)}$, to optimize their representation. Consequently, the entries for each NB of $b^{(k)}$ are excluded from $s^{(k-1)}$. The complete aforementioned IBRM strategy is detailed in step by step manner in Algorithm 3, which can be directly applied on the bit-planes starting from $b^{(7)}$, ..., $b^{(0)}$ resulting into compressed bit-streams i.e., $\hat{s}^{(7)}$, ..., $\hat{s}^{(0)}$.

To better explain the process, the example depicted in Fig. 3 is extended and illustrated in Fig. 5 to provide a visual representation of how the bit-planes are compressed into bit-streams. More specifically, Fig. 5a shows the generation of quad-tree for $b^{(8)}$ and its encoding in compressed bit-stream using level-order tree traversal as per the process defined in subsection III-A2. Further, Fig. 5b and 5c depicts the quad-tree generation and their encoding into compressed bit-stream for the bit-planes $\hat{s}^{(7)}$ and $\hat{s}^{(6)}$, where only EB nodes of previous bit-plane are extended as per the Algorithm 3.

Algorithm 3 Inter Bit-plane Redundancy Mitigation

function IBRM($s^{(k)}$, $b^{(k-1)}$)
  Initialize an empty bitstream $\hat{s}^{(k-1)}$
  for each bit x in $s^{(k)}$ from left to right **do**
    if $x = 0$ **then**
      Determine the sub-tree for the corresponding block of $b^{(k-1)}$ using Function-GENERATEQUADTREE
      Generate sub bit-stream $s_x'^{(k-1)}$ from the sub-tree using Function-ENCODEQUADTREE
      Append $s_x'^{(k-1)}$ to $\hat{s}^{(k-1)}$
    end if
  end for
  return bitstream $\hat{s}^{(k-1)}$
end function

However, it's essential to highlight that the IBRM strategy is maintained until the size of $\hat{s}^{(k)}$ is less than the total number of embeddable bits in the current bit-plane. Natural images might have lower-order bit-planes $b^{(k)}$ with fewer embeddable bits than their corresponding bit-streams $\hat{s}^{(k)}$, leading to possible information loss. Therefore, the number of bit-planes in an image that, when included for embedding, improve the overall EC, are referred to as 'embeddable bit-planes'. The total count of these embeddable bit-planes, denoted by n, can be encoded by a maximum of 3 bits. The process for ascertaining the complete sum of embeddable bits in the k^{th} bitplane, denoted as $T_{EC}^{(k)}$, requires a thorough examination of each node within the corresponding trees. $T_{EC}^{(k)}$ is determined using Eq. (6).

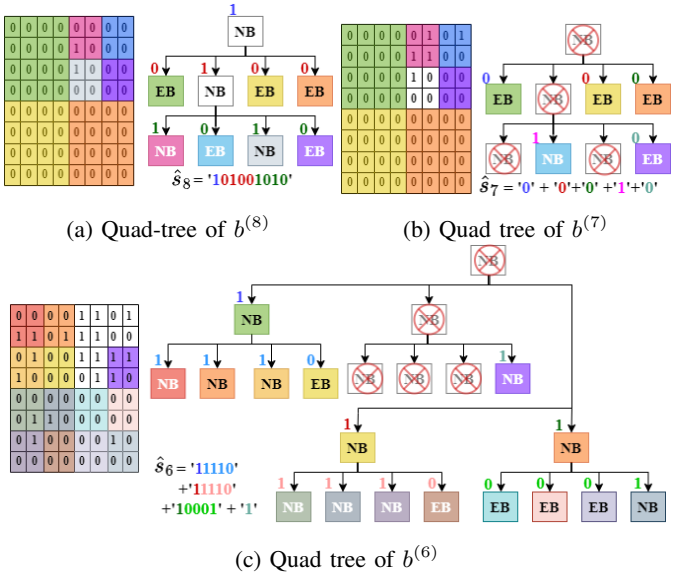
$$T_{EC}^{(k)} = \sum_{c=1}^{\alpha} (h_{(c)} + 2)^2, \quad \text{if } node_{(c)} \in EB, \tag{6}$$

where $h_{(c)}$ represents the height of the node ($node_{(c)}$) within the tree (which is zero for leaf nodes), and α symbolizes the total count of nodes within the k^{th} bit-plane.

4) Organization of Supporting Information: As discussed above, the proposed method makes use of level-order traversal while generating bit-streams for representing the bit-planes, despite knowing the fact that level-order traversal is more complex than pre-order traversal of the quad-tree. This deliberate choice is made to capture even the most minute redundancies present in the codes so that further optimization using Huffman encoding can be done. In the level-order traversal of the quad-tree, all 15 possible combinations of a node's four child nodes are represented by codes ranging from 0001, 0010, 0011, ..., 1111. Next, the generated bit-streams to represent the quad-tree can be further condensed using Huffman encoding (if needed). Therefore, the proposed MBQ-RDHEI method can transform itself into three versions namely Direct-MBQ (MBQ), Huffman-MBQ (HMBQ), and Separate Huffman-MBQ (SHMBQ), where the MBQ method does not employ Huffman encoding to compress the received bit-streams (i.e.,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993



(a) Quad-tree of $b^{(8)}$ (b) Quad tree of $b^{(7)}$

(c) Quad tree of $b^{(6)}$

Fig. 5: Illustrative process of compressed bit-stream ($s^{(8)}$, $s^{(7)}$, $s^{(6)}$) generation.

$\hat{s}^{(8)}$, $\hat{s}^{(7)}$, $\hat{s}^{(6)}$, ..., $\hat{s}^{(8-n+1)}$), instead directly utilizes them. In contrast, HMBQ first traverses the bit-streams from left to right and determines the frequencies of each code to generate the corresponding Huffman codes. Next, all the codes in the bit-streams are replaced by their corresponding Huffman codes to get the Huffman-encoded bit-streams. Likewise, SHMBQ also employs Huffman encoding but separately on each bit-stream. However, it is to be noted that each version of the method offers a different EC with varying execution time. Further, supporting information (SI) that includes a total number of embeddable bit-planes (n), and compressed bit-streams ($\hat{s}^{(8)}$ to $\hat{s}^{(8-n+1)}$) arranged in descending order of their significance, RP, non-embeddable bits of 8^{th} bit-plane denoted as $NE_{b^{(8)}}$, and/or Huffman code/rules, which is indispensable for identifying the embeddable space in the image, as well as reconstructing the image itself, is also generated, whose length is denoted by SI_{len}. The complete SI is crucial for data hiding/extraction and image reconstruction, that's why it is embedded in the image using a process detailed in the next sub-section.

B. Image Encryption and SI Embedding

Though the field of image encryption presents a range of encryption strategies for ensuring the confidentiality of cover image contents, the Advanced Encryption Standard with counter mode (AES-CTR) has been widely adopted as a secure, trustworthy, and simple solution. So, the proposed method also makes use of AES-CTR for the image encryption by first generating a pseudo-random matrix ($\mathbb{E}$) of size $R \times C$, using the image encryption key (K_I) as a seed with the help of pseudo-random number generator. This matrix $\mathbb{E}$ is initially converted into 8 bit-planes $\mathbb{E}^{(7)}$, $\mathbb{E}^{(6)}$, ... , $\mathbb{E}^{(0)}$ as follows,

$$\mathbb{E}^{(k)}_{(i,j)} = \left\lfloor \frac{(\mathbb{E}_{(i,j)})\%(2^{k+1})}{2^k} \right\rfloor, \tag{7}$$

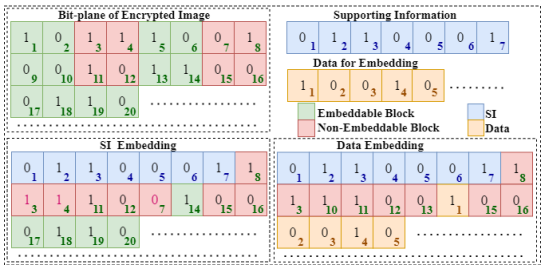


Fig. 6: SI and data embedding.

where $\%$ and $\lfloor * \rfloor$ represent the modulus operator and the floor function, respectively, and $7 \leq k \leq 0$. Next, pixels of the image are encrypted using bit-wise XOR ($\oplus$) operation defined as follows, to get the encrypted image (I_e).

$$I^{(k)}_{e(i,j)} = I^{(k)}_{(i,j)} \oplus \mathbb{E}^{(k)}_{(i,j)}. \tag{8}$$

Thus, the obtained image also has the 8 bit-planes i.e., $b_e^{(7)}$, $b_e^{(6)}$,..., $b_e^{(0)}$, like the I.

Thereafter, SI is embedded in the encrypted bit-planes as it is required for the efficient operation of data embedding, data extraction, and image reconstruction. For this, the encrypted bit-planes are considered in order of $b_e^{(7)}$, $b_e^{(6)}$, ...,$b_e^{(8-n+1)}$ and embedding in each bit-plane is done in raster scan order by replacing the bits of bit-planes by the bits of SI while collecting the bits belonging to NBs as NE_{bits}. After the SI embedding, NE_{bits} are embedded by replacing the bits belonging to onward EBs only, with NE_{bits}. Thus, the SI-enabled encrypted image I'_e is obtained, by collecting the resultant bit-planes i.e., $b_e'^{(7)}$, $b_e'^{(6)}$, ..., $b_e'^{(0)}$ using following Eq. 9. To elucidate the SI embedding process, Fig 6 offers an illustration where the first block only shows EB and NB positions in a bit-plane and the second block shows the SI of length 7 bits and data to be embedded of 5 bits. As shown in the third block, the SI replaces all the bits in sequence (that includes bits from both EB and NB), however for the sake of reversibility and lossless extraction the NE_{bits} pinpointed at positions 3, 4, and 7 are recorded and embedded at subsequent embeddable positions (shown in green color). Thus, NE_{bits} undergo a positional shift to locations 9, 10, and 13, to avoid the information loss.

$$I'_{e(i,j)} = \sum_{k=7}^{0} b'^{(k)}_{e(i,j)} \times 2^k. \tag{9}$$

C. Data Hiding

Initially, the data hider must extract the SI from I'_e in order to identify the reserved room for data embedding. This is accomplished by reading the image starting from the MSB-plane in raster scan order by first decoding the initial three bits to get n. After skipping the RP bits from SI, the next bits of SI are processed to generate the quad-trees for $b^{(8)}$ using algorithm 4. Likewise, to generate the quad-tree for bit-plane $b^{(7)}$, the quad-tree for the previous bit-plane i.e., $b^{(8)}$ (in case of $b^{(7)}$) is subjected to level-order traversal and for each EB

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993

node, the tree is extended using the steps outlined in Algorithm 4. This process is repeated for $8-n+1$ subsequent bit-planes. Next, the data hider starts traversing from the MSB-plane in raster scan order to count the NE_{bits} in the first SI_{len} positions. Next, he/she skips SI_{len} positions followed by NE_{bits} number of positions/bits belonging to EBs to start embedding the secret data bits. However, the secret data is first encrypted using a secure encryption algorithm and the data-hiding key (K_D) to enhance the level of security and then the embedding is done at each EB position by replacing its bits with the secret data bits. The process of embedding is continued till $8-n$ bit-plane and a marked encrypted image I_m is obtained. To elucidate the data hiding process, the last block of Fig 6 shows the embedding of data at positions - 14, 17, 18, 19, and so on.

Algorithm 4 Reconstruction of Quad-tree from SI

procedure RECONSTRUCTQUADTREE(SI, $block_size$, m)
  Create a new quad-tree with $root$ node having value $= SI[m]$ of $block_size \times block_size$
  Enqueue $node_block$ pair i.e., $root$ node with its $block_size$ in $queue$ and set $m \leftarrow m+1$
  while $queue$ is not empty **do**
    Dequeue a $node_block$ pair from the $queue$
    if If the node's value is 1 and the $block_size > 2$ **then**
      Partition the block into 4 new sub-blocks, each with size $block_size/2 \times block_size/2$
      for i in range 1 to 4 **do**
        Create a new node with the value $= SI[m]$
        Enqueue the new $node_block$ pair in $queue$
        set $m \leftarrow m+1$
      end for
    end if
  end while
  return quad-tree and m
end procedure

D. Data Extraction and Image Recovery

To recover the hidden data and reconstruct the original image, the receiver first extracts the n, SI and $NE_{b^{(8)}}$. Subsequently, the receiver fabricates the quad-tree, mirroring the approach of the data hider as elucidated in subsection III-C. Next, the receiver requires access to K_D and/or K_I to extract the hidden data and/or recover the original image. Thus, three cases are possible, which are defined and discussed as follows:

In a case where the recipient holds only the K_D, the recipient must retrieve the bits from each EB denoted as E_{bits}, sequentially, beginning from the MSB-plane and the second pixel, while traversing in raster scan order. Following the skipping of the initial SI_{len} bits from the collected bits, the recipient decrypts the remaining bits using K_D to recover the secret data, without any loss.

If the recipient only possesses the K_I, he/she must read the image similarly as in the previous case. Next, the NE_{bits} (where SI was embedded), are replaced by the same number of bits belonging to EBs. Afterwards, the recipient generates a pseudo-random matrix ($\mathbb{E}$) as in section III-B and converts the updated marked encrypted image (I_m) into eight bit-planes using equation (7). Next, the bit-planes of I_m are decrypted using Eq. 8 to get the decrypted bit-planes i.e., $b_d^{(7)}$, $b_d^{(6)}$, ..., $b_d^{(0)}$.

To determine the bit-plane of the difference matrix, the quad-tree of each bit-plane is used. Initially, the MSB-plane of the difference matrix is generated using the first quad-tree and $NE_{b^{(8)}}$. A bit-plane of the image size is created with all zero values, and all non-embeddable positions are replaced with their respective bits from $NE_{b^{(8)}}$. Subsequently, in order to generate the next $8-n$ MSB-planes, all the positions belong to EBs of $8-n$ MSB-planes ($b_d^{(7)}$ to $b_d^{(8-n)}$) are replaced with '0'. The remaining LSB-planes are given as $b_d^{(k)}$ for $8-n-1 \geq k \geq 0$. These 9 bit-planes are collected using Eq. 9 for $8 \geq k \geq 0$, and the difference matrix D is obtained. Finally, the lossless and identical reconstructed image I_r is obtained using the difference matrix and RP from Eq. (2).

When both the data-hiding and encryption keys are available with the receiver, he/she can extract the embedded data as in the first case and can also reconstruct the original image as in the second case, without any loss/error.

TABLE III: NET GAIN IN EC (BITS) FOR IMAGE $Lena$ BY IBRM

Bit-plane		$b^{(7)}$	$b^{(6)}$	$b^{(5)}$	$b^{(4)}$	$b^{(3)}$	$b^{(2)}$
Without IBRM	$s^{(k)}$/ EC	757/ 261740	6745/ 256952	21845/ 236184	56445/ 177804	84917 /72876	87321/ 27324
With IBRM	$\hat{s}^{(k)}$/ EC	757/ 261740	6487/ 266808	19005/ 234953	44331/ 172932	48571/ 59124	15352/ 6712
Total Difference in Pure EC (With IBRM - Without IBRM)		0	114	1609	7242	10553	-

IV. EXPERIMENT RESULT AND DISCUSSIONS

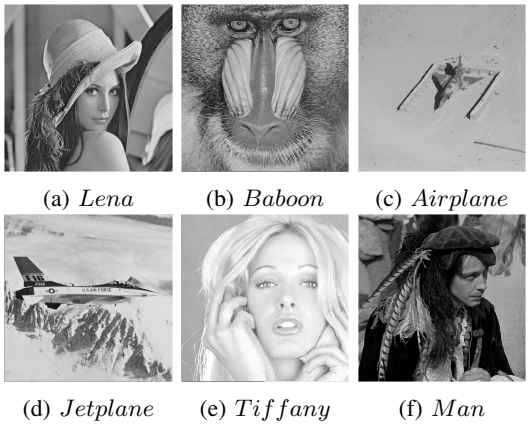
This section presents a detailed discussion of the performance of the proposed method and its comparison with the SOTA methods. The discussion is organized into four subsections. In the first subsection, the embedding performance of the proposed method is evaluated and analyzed on well-known test images of the SIPI dataset [39], including $Lena$, $Baboon$, $Airplane$, $Jetplane$ (F-16), $Tiffany$ and Man (Male) as depicted in Fig. 7, as well as on the big datasets, namely, BOSSbase [40] and BOWS-2 [41]. The second subsection presents a comparison of the proposed method with the modern SOTA methods, to highlight the effectiveness of the proposed method. The third subsection evaluates the security effectiveness of the proposed method considering the perceptual level, several statistical measures and malicious attacks. In the last subsection, a detailed discussion of the complexity analysis of the MBQ-RDHEI method is presented.

A. Embedding Performance Analysis

To experimentally demonstrate the embedding performance of the proposed method, the cover image $Lena$ is used as a test image. As discussed, the image is first transformed into a difference matrix having 9-bit-planes ($b^{(8)}$ to $b^{(0)}$), where eight bit-planes, starting from $b^{(8)}$ to $b^{(1)}$, represent the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993



(a) $Lena$ (b) $Baboon$ (c) $Airplane$

(d) $Jetplane$ (e) $Tiffany$ (f) Man

Fig. 7: Standard test images from the SIPI dataset [39].

magnitude, while the $b^{(0)}$ contains the sign bit. For each bit-plane, a quad-tree is designed using Algorithm 1, where each node corresponds to a block of the bit-plane. The quad-tree is subsequently transformed into a bit-stream with the help of Algorithm-2, while inter bit-plane redundancy is efficiently reduced through the implementation of the novel IBRM using Algorithm-3. To demonstrate the efficacy of the proposed IBRM, a detailed analysis using the test image $Lena$ is provided in Table III. Here, the length of the bit-streams with and without IBRM is presented, where it can be clearly observed that the IBRM strategy significantly minimizes the length of the bit-streams which in turn increases the pure EC. Though it may result in a few misclassifications, a reduction in SI size-ably increases the overall net gain in Pure EC. More specifically, the cumulative benefit after employing IBRM for $Lena$ is a reduction of 19518 bits SI.

Next, in Table IV, the details about the number and types of nodes and their corresponding block's size along with the total EC, for each bit-plane are provided. This table also helps us elaborate on the size of the quad-tree, which is essentially an overhead. As evident from the Table IV, the bit-plane $b^{(8)}$ does not contain any '1' bit which means a single node can represent the entire bit-plane, corresponding to a block size of 512×512. Consequently, the subsequent bit-plane i.e., $b^{(7)}$ will only have one quad-tree $q^{(7)}$ that has a total of 757 nodes which can be represented by a total of 757 bits. Similarly, for $b^{(6)}$, a total of 467 quad-trees are constructed having a total of 6487 nodes which can be represented by a total of 6487 bits. Similarly, for other remaining bit-planes, the results detailing about the number of quad-trees, total nodes and their corresponding block size, and total EC are provided in Table IV. Further, it can also be observed that the maximum EC i.e., 261780 is provided by $b^{(7)}$ followed by $b^{(6)}$ and so on. This is because of the level of sparsity present in the bit-planes, which also significantly affects the size of the overhead.

Further, Table V, which is organized into three sections where the first, second and third section details the results on bit-plane wise compressed bit-stream size, Huffman code length, SI length (SI_{Len}) and pure EC of the proposed MBQ, HMBQ and SHMBQ, respectively, is provided.

As evident from the Table IV and the first section of Table V, the total EC of the bit-planes ($b^{(2)}$ to $b^{(0)}$) is less than the size of their compressed bit-streams ($\hat{s}^{(2)}$ to $\hat{s}^{(0)}$), therefore the bit-planes $b^{(2)}$ to $b^{(0)}$) are not used for embedding. It can be further observed that the overall length of SI accounts to 119163 bits, which includes 3 bits for the number of embeddable bit-planes, 8 bits for the RP, 119152 bits for the quad-tree representation of $b^{(8)}$ to $b^{(3)}$ bit-planes, and 0 bits for $NB_{b^{(8)}}$. Hence, the pure EC is 866393 bits for the MBQ method. As far as the HMBQ method is concerned, the required Huffman code's length is 127 bits. It can be further observed that the overall length of SI accounts to 115577 bits. Hence, the pure EC of HMBQ is 869979 bits which is 3586 bits higher than the same of the MBQ. For the SHQMS approach, the cumulative length of Huffman codes is 511 bits, which incorporates a single bit for each bit-plane indicating the utilization of Huffman codes for that bit-plane. The pure EC of this method is 871820 bits, which is having a net gain of 5427 bits and 1841 bits compared to the MBQ and HMBQ methods, respectively. Therefore, it can be concluded that the SHQMB method offers the highest EC among all the three versions of the proposed method.

Furthermore, Table VI provides the results of EC, embedding rate (ER) and SI for all the test images shown in Fig. 7. The presented findings indicate a strong correlation between an image's smoothness and its EC, ER, and SI. Notably, the $Baboon$ image, characterized by a higher presence of textures, exhibits significantly lower ER and EC values compared to the $Airplane$ image, which predominantly consists of smoother regions. Additionally, the length of SI for the $Baboon$ image is comparatively higher, while it is minimal for the $Airplane$ image. More specifically, the proposed methods i.e., MBQ, HMBQ, and SHMBQ, achieve ER of 1.6638, 1.6772, and

TABLE IV: NUMBER OF DIFFERENT SIZE AND TYPE OF BLOCKS ALONG WITH TOTAL EC PER BIT-PLANE FOR IMAGE $Lena$

Block Type	Block Size	Bit-plane $b^{(8)}$	$b^{(7)}$	$b^{(6)}$	$b^{(5)}$	$b^{(4)}$	$b^{(3)}$	$b^{(2)}$	$b^{(1)}$	$b^{(0)}$
EB	512×512	1	0	0	0	0	0	0	0	0
EB	256×256	0	1	0	0	0	0	0	0	0
EB	128×128	0	3	0	0	0	0	0	0	0
EB	64×64	0	22	19	4	0	0	0	0	0
EB	32×32	0	34	76	54	1	0	0	0	0
EB	16×16	0	58	190	212	43	0	0	0	0
EB	8×8	0	77	441	722	606	14	0	0	0
EB	4×4	0	105	1017	2321	3668	517	4	0	0
EB	2×2	0	167	2006	6414	15857	12489	1662	300	0
NB	2×2	0	101	1233	5464	15505	28452	13103	1378	300
No. of total nodes		1	757	6487	19005	44331	48571	15352	1682	300
Total EC		-	261740	256808	234952	172932	59124	6712	1200	0

TABLE V: BIT-PLANE WISE PERFORMANCE ANALYSIS OF ALL THREE VERSIONS OF THE PROPOSED METHOD ON IMAGE $Lena$

Method		Bit-planes $b^{(8)}$	$b^{(7)}$	$b^{(6)}$	$b^{(5)}$	$b^{(4)}$	$b^{(3)}$	$b^{(2)}$	$b^{(1)}$	$b^{(0)}$
MBQ	$\hat{s}^{(8)}$ **to** $\hat{s}^{(0)}$	1	757	6487	19005	44331	48571	15352	1682	300
	SI Length (SI_{Len}) **:** 3 (n) + 8 (RP) + 119152 ($\hat{s}^{(8)}$ to $\hat{s}^{(3)}$) = 119163									
	Pure EC : 985556 (Total EC of $b^{(8)}$ to $b^{(3)}$) - 119163 = 866393									
HMBQ	$\hat{s}^{(8)}$ **to** $\hat{s}^{(0)}$	1	724	6224	18339	42820	47331	15250	1681	300
	Huffman code Length : 127									
	SI Length (SI_{Len}) **:** 3 (n) + 8 (RP) + 115439 ($\hat{s}^{(8)}$ to $\hat{s}^{(3)}$) + 125 = 115577									
	Pure EC : 985556 (Total EC of $b^{(8)}$ to $b^{(3)}$) - 115577 = 869979									
SHMBQ	$\hat{s}^{(8)}$ **to** $\hat{s}^{(0)}$	1	600	5904	17992	42610	46107	14668	1773	300
	Huffman code Length :	0+1	0+1	126+1	126+1	127+1	126+1	142+1	0+1	0
	SI Length (SI_{Len}) **:** 3 (n) + 8 (RP) + 113214 ($\hat{s}^{(8)}$ to $\hat{s}^{(3)}$) + 511 = 113736									
	Pure EC : 985556 (Total EC of $b^{(8)}$ to $b^{(3)}$) - 113736 = 871820									

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993

TABLE VI: PURE EC ($bits$), SI_{Len} ($bits$) AND ER (bpp) ON THE STANDARD TEST IMAGES

Methods		Lena	Airplane	Baboon	Jetplane	Tiffany	Man
MBQ	**Pure EC**	866393	1068952	436146	973997	898815	790014
	SI_{Len}	119163	93736	126298	129395	115069	126446
	ER	3.3050	4.0777	1.6638	3.7155	3.4287	3.0137
HMBQ	**Pure EC**	869979	1074747	439659	977700	902006	793149
	SI_{Len}	115577	87941	122785	125692	111878	123311
	ER	3.3187	4.0998	1.6772	3.7296	3.4409	3.0256
SHMBQ	**Pure EC**	871820	1074449	440288	979118	903635	794267
	SI_{Len}	113736	88239	122156	124274	110249	122193
	ER	3.3252	4.0987	1.6796	3.7350	3.4471	3.0299

TABLE VII: PURE ER (bpp) ON BOSSBASE AND BOWS2

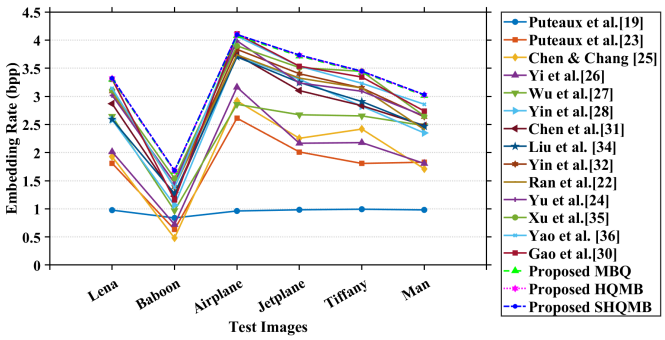
	BOSSBase			BOWS2		
	Average	Best	Worst	Average	Best	Worst
MBQ	3.9970	7.8648	0.9654	3.9161	7.2973	0.8839
HMBQ	4.0110	7.8650	0.9780	3.9299	7.3016	0.8972
SHMBQ	4.0197	7.8648	0.9825	3.9343	7.3021	0.9002

1.6796 bpp for the $Baboon$ and 4.0773, 4.0998, and 4.0987 bpp for the $Airplane$, respectively.

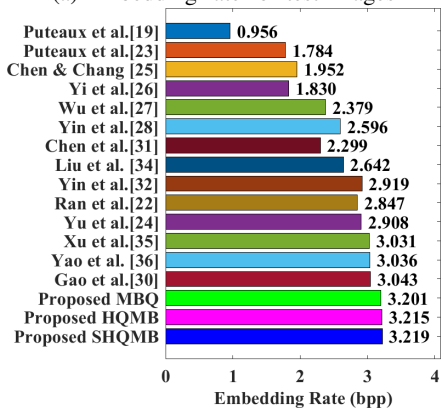
The proposed method has been evaluated on two datasets, BOSSBase and BOWS-2, each comprising 10,000 images of size 512×512 pixels. The results of the evaluation are presented in Table VII, where the PSNR and SSIM of each reconstructed image are $+\infty$ and 1, respectively. The average net ER for the proposed MBQ, HMBQ and SHMBQ methods in BOSSBase are 3.9970, 4.0110 and 4.0197 bpp, respectively, while in BOWS-2, they are 3.9161, 3.9299 and 3.9343 bpp, respectively. The best ER achieved for an image in BOSSBase is 7.8648, 7.8650 and 7.8648 bpp for proposed methods, respectively, whereas in BOWS-2, it is 7.2973, 7.3016 and 7.3021 bpp. The worst ER for an image in BOSSBase is 0.9654, 0.9780 and 0.9825 for proposed methods, respectively, while the worst ER for an image in BOWS-2 are 0.8839, 0.8972, and 0.9002 bpp, respectively. Based on these analyses, it can be concluded that the proposed MBQ, HMBQ and SHMBQ methods are efficient and can be used for the privacy and security of data.

B. Comparison with State-of-the-Art Methods

To effectively demonstrate the performance of proposed methods, an extensive comparison with fourteen SOTA methods, that include [19], [22]–[28], [30]–[32], [34]–[36], have been made. Since all three versions of the proposed method and the aforementioned SOTA methods are completely reversible, performance comparison only in terms of ER is done. It can be seen from Fig. 8 that the proposed method is able to achieve the maximum ER of 3.325 bpp for the $Lena$, whereas the highest ER attained by Gao *et al.*'s method [24] among the SOTA methods, is 3.304 bpp, for the same image. Thus, a gain of 0.001, 0.015, and 0.021 bpp is attained by all three versions of the proposed methods with respect to [30]. Further, it can be observed that the proposed method achieves the lowest ER in the case of $Baboon$, i.e., 1.680 bpp whereas among the SOTA methods, Xu *et al.*'s method [35] provides the maximum ER of 1.541 bpp for the same $Baboon$ image. Thus, a gain of 0.123, 0.126, and 0.139 bpp by the proposed method with respect to Xu *et al.*'s method [35], for the $Baboon$ image, is obtained. Thus, it can be clearly stated based on the observation of the Fig. 8 that the proposed method easily outperforms all the fourteen aforementioned SOTA methods in terms of ER, for all the six test images (except $Airplane$ where our performance is comparable with the best performing SOTA method i.e., Gao *et al.* [24]). It is worth to be noted that the proposed method has significant performance gain even for the challenging $Baboon$ image. Further, as evidenced by Fig. 8b, the proposed method unequivocally demonstrates superior performance, with the AER for all the six test images consistently exceeding the rates achieved by the other fourteen SOTA methods. This observation also confirms that the proposed method has been able to effectively exploit redundant information present in the natural images, which ultimately leads to superior performance compared to the SOTA methods.



(a) Embedding rate for test images



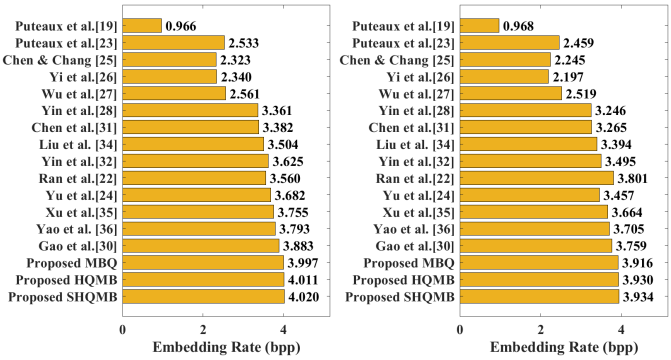
(b) Average embedding rate for test images

Fig. 8: Comparative results on six standard test images.

Further, an extensive analysis on two large and benchmark datasets, namely BOSSBase and BOWS-2, to compare the performance of our proposed method with that of the SOTA methods, has been done. As demonstrated in Fig. 9, all three versions (i.e., MBQ, HMBQ and SHMBQ) of the proposed method, achieve an impressive ER of 3.997, 4.011 and 4.0197 bpp for BOSSBase and 3.9161, 3.9299 and 3.9343 bpp for BOWS-2, respectively. As far as the SOTA methods are concerned, Yao *et al.* [36] and Gao *et al.* [30] methods achieve the highest ER of 3.793 and 3.883 bpp for the BOSSBase, and 3.705 and 3.644 bpp for BOWS-2, respectively. Thus, it can be easily stated that all the versions of the proposed method

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993



(a) $BOSSBase$ (b) $BOWS-2$

Fig. 9: Comparative results on two datasets.

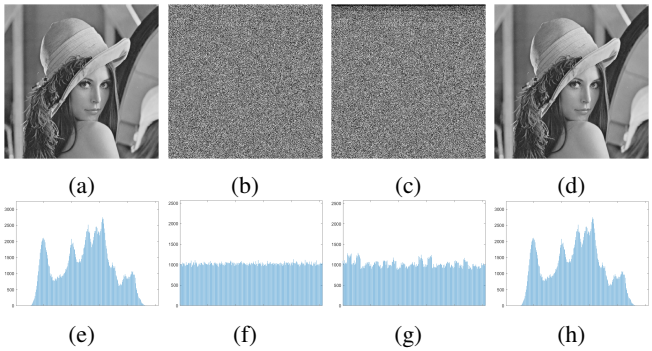


Fig. 10: Representation of resultant images and their histogram distribution of each phase of the proposed MBQ-RDHEI method: (a) Original image $Lena$, (b) Encrypted image, (c) Marked encrypted image, (d) Reconstructed Image (e-h) Histogram distribution of (a-d), respectively.

easily outperform the SOTA methods in terms of ER, for both the datasets.

C. Security Analysis

In this subsection, the robustness of the proposed method is analyzed against various security attacks such as Perceptual, Statistical and other Malicious attacks. For this, the subsection is divided into three parts as follows:

1) Robustness against Perceptual Attacks: As discussed above, the proposed MBQ-RDHEI method uses separate keys to protect the hidden data and cover images. More specifically, K_D encrypts the secret data using any secure encryption technique, making data leakage highly improbable, while K_I encrypts the cover image to anonymize its contents to ensure confidentiality. To evaluate the security resilience of the proposed method, the original $Lena$ image I, as shown in Fig. 10a, has been considered. The I is first pre-processed to reserve a room for embedding and subjected to the CTR-AES method, resulting in the encrypted image I_e as shown in Fig. 10b. The SI is subsequently embedded into I_e. Next, the data hider proceeds to embed the encrypted secret data in I_e, producing a marked encrypted image I_m, as depicted in Fig. 10c. At the receiving end, the recipient extracts the data and reconstructs the image using the available keys, yielding the reconstructed image I_r, as shown in Fig. 10d. From Fig. 10, it is apparent that both I_e and I_m exhibit significant distortion and pervasive noise, rendering the original content unidentifiable. To further analyze the characteristics of the I, I_e, I_m, and I_r, frequency distribution of the 256 gray-scale levels using histogram representation are presented in Figs. 10e, 10f, 10g, and 10h, respectively. The distributions presented in Fig. 10f and 10g are noticeably uniform, in contrast to the histograms of I and I_r. This uniformity signifies that the proposed method provides a high level of perceptual security as far as encrypted and marked encrypted images are concerned.

2) Robustness against Statistical Attacks: To evaluate the performance of the proposed MBQ-RDHEI method against various statistical attacks, different statistical measures such as horizontal and vertical correlation [42], Shannon entropy [43], number of changing pixel rate (NPCR) [44], unified average changed intensity (UACI) [44], peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) [45] have also been considered. The experimental findings for the test image $Lena$ are presented in Table VIII. As evident from the findings, the correlation in all directions i.e., horizontal, vertical, or diagonal for I, is high (or close to one), whereas the same is almost negligible (or more specifically, close to zero), for both I_e and I_m. This signifies that the proposed method completely destroys the relations among the neighbour pixels to make the resultant image fully anonymized. Further, measured entropy for both I_e and I_m is close to 8, which is near to maximum for an 8-bit gray-scale image, as given in Table VIII. The high entropy suggests that the pixels are spread more equally within the permitted range i.e., 0 to 2^8-1, which in turn signifies a high degree of unpredictability. Furthermore, it can also be observed from Table VIII that NCPR and UCID reach close to 100% for both I_e and I_m, which means almost all the pixels of the images are changed. This in turn indicates high perceptual security.

TABLE VIII: STATISTICAL ANALYSIS ON IMAGE $Lena$

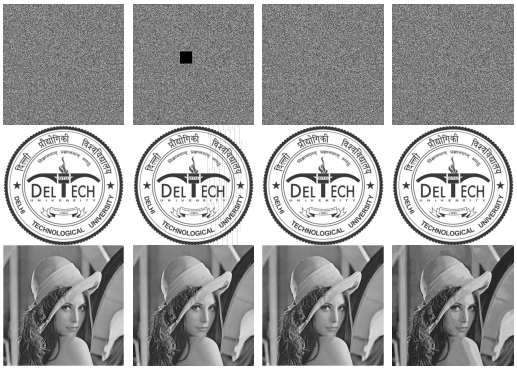
Image	Original image (I)	Encrypted image (I_e)	Marked encrypted image (I_m)
Horizontal Correlation	0.9719	-0.0001	0.0251
Vertical Correlation	0.9850	-0.0005	0.0199
Diagonal Correlation	0.9593	0.0018	0.0241
Information Entropy	7.4451	7.9993	7.9970
NPCR (I and I_e/I_m)	-	99.80	99.70
UACI (I and I_e/I_m)	-	72.98	73.42
PSNR (I and I_e/I_m)	-	9.2390	9.1902
SSIM (I and I_e/I_m)	-	0.0106	0.0108

To conduct a more in-depth assessment of the similarity of I with I_e and I_m, the PSNR and SSIM are evaluated, and corresponding results are presented in Table VIII. The experimental results show that the obtained PSNR and SSIM, are low for both the images, i.e., I_e and I_m, which in turn indicates that the I_e and I_m are significantly different and structurally dissimilar with respect to I. Thus, it can be stated that the proposed MBQ-RDHEI method demonstrates significant robustness against differential attacks.

3) Robustness against Other Malicious Attacks: To evaluate the robustness against other malicious attacks, the MBQ-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993



(a) (b) (c) (d)

Fig. 11: The experimental results obtained under various removal attack scenarios (a). Original, (b) Patch removal, (c) Fourth bit-plane removal, and (d) Fifth bit-plane removal.

RDHEI has been subjected to tampering, brute-force, and collage attacks. For this, a logo has been embedded in the encrypted $Lena$ image and various removal attacks such as Patch removal, Fourth bit-plane removal and Fifth bit-plane removal have been conducted as shown in Fig. 11. The images presented in the last two rows show that there is only a minimal distortion in the restored cover image, and the extracted logo is also visibly similar. However, there exists the potential for data loss should the bit-planes containing secret data be compromised. Similarly, the removal of bit-plane containing the SI bit-planes within the encrypted image could lead to a loss of the cover image itself. Conversely, the application of brute-force attacks becomes virtually unfeasible given the robust design of the proposed RDHEI method. This method incorporates AES for image encryption, a highly reliable and secure encryption protocol. Furthermore, the flexibility of the proposed method allows for the encryption of data through any robust algorithm selected by the data hider, which can provide a protection for the data against brute-force attacks.

Moreover, since the MBQ-RDHEI is separable and solely dependent on the bit-stream of SI for data embedding, resilience against collage attacks is enabled. Thus, it can be concluded that the MBQ-RDHEI has good resilience against various forms of malicious attacks. It is also important to note that since all three versions of the proposed method i.e., MBQ, HMBQ and SHMBQ share identical characteristics, the aforementioned findings regarding the security resilience are equally applicable to all three versions of the proposed method.

D. Complexity Analysis

The proposed MBQ-RDHEI method has also been analyzed for its time complexity, and found to have $O((R\times C)log(R\times C))$ time complexity. It is mainly because the number of nodes in the quad-tree is proportional to the size of the image, which is $R\times C$, and each node is processed only once. Further, since the method performs a logarithmic number of subdivisions on each node, the total time complexity becomes $O((R\times C)log(R\times C))$.

Further, the computational efficiency of the MBQ method has also been evaluated and compared with the SOTA methods.

TABLE IX: COMPARISON OF COMPUTATION TIME AND EC OF THE PROPOSED METHOD WITH THE SOTA METHODS

Method	Total computational time of 6-test images(seconds)	Total EC of 6-test images (bit)	Average time per embedded bit (seconds/bit)
Puteaux *et al.* [19]	98.1472	1503142	0.65294×10^{-4}
Puteaux *et al.* [23]	247.9135	2805341	0.88371×10^{-4}
Chen & Chang [25]	187.2173	3070239	0.60978×10^{-4}
Yi *et al.* [26]	115.9477	2878944	0.40274×10^{-4}
Wu *et al.* [27]	286.2726	3742053	0.76501×10^{-4}
Yin *et al.* [28]	200.8890	4083817	0.49196×10^{-4}
Chen *et al.* [31]	276.5286	3615464	0.76485×10^{-4}
Liu *et al.* [34]	302.7732	4154739	0.72874×10^{-4}
Yin *et al.* [32]	502.4172	4590510	1.09450×10^{-4}
Ren *et al.* [22]	484.6203	4478570	1.08208×10^{-4}
Yu *et al.* [24]	308.0526	4573023	0.67363×10^{-4}
Xu *et al.* [35]	481.5319	4767897	1.00994×10^{-4}
Yao *et al.* [36]	302.0317	4774954	0.63253×10^{-4}
Gao *et al.* [30]	422.1275	4786829	0.818521×10^{-4}
Proposed MBQ	310.3405	5034317	0.61645×10^{-4}
Proposed HMBQ	315.1940	5056940	0.62329×10^{-4}
Proposed SHMBQ	324.1043	5063577	0.64007×10^{-4}

For this, all the methods are implemented on MATLAB running on a computer with a 6-core AMD Fx-6300 processor, 8 GB of RAM, and the Windows 10 operating system. The results for computational time (in seconds) along with the maximum EC (in bits) and the average time per embedded bit (in seconds/bit) for six test images shown in Fig. 7 are provided in Table IX. The presented results demonstrate that the MBQ method outperforms Ren *et al.* [22], Yu *et al.* [24], Gao *et al.* [30], Yin *et al.* [32], and Xu *et al.* [35] in terms of computational time, despite having the highest EC among the SOTA methods. As far as average time per embedded bit is concerned, the proposed MBQ method is only behind Chen & Chang [25], Yi *et al.* [26] and Yin *et al.* [28], while surpassing their EC performance by a big margin.

In addition, the proposed MBQ method has also been evaluated for its space complexity. The space complexity of generating the quad-tree is proportional to the size of the image, with the algorithm creating one node for each NB in the quad-tree. The queue used by the algorithm also contributes to the space complexity, but its size is bounded by the number of nodes in the quad-tree. Thus, the overall space complexity is $O(R\times C)$.

V. CONCLUSION

In this paper, a novel high-capacity RDHEI method using multi-level blocking with quad-tree was introduced. The MBQ-RDHEI method transforms the original image into a difference matrix having high sparsity in the MSB-planes. The sparse bit-planes are then encoded using a quad-tree based compression strategy to significantly help in reserving a bigger room for embedding. The encoding exploits both inter as well and intra bit-plane redundancy to significantly reduce their size. Experimental findings indicate that the MBQ-RDHEI methods outperform the SOTA methods, by achieving higher embedding rates. More specifically, three versions i.e., MBQ, HMBQ, and SHMBQ of the proposed method have achieved impressive ER of 3.997, 4.011, and 4.0197 *bpp* for BOSSBase and 3.9161, 3.9299, and 3.9343 *bpp* for BOWS-2, respectively, while ensuring complete reversibility of the original images. Going forward, we aim to further enhance the performance of

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993

the proposed MBQ-RDHEI method by conducting research on predictors to generate even more sparse difference images.

REFERENCES

[1] M. Wu and B. Liu, "Watermarking for image authentication," in *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, vol. 2, pp. 437–441, IEEE, 1998.

[2] G. Langelaar, I. Setyawan, and R. Lagendijk, "Watermarking digital image and video data. a state-of-the-art overview," *IEEE Signal Processing Magazine*, vol. 17, no. 5, pp. 20–46, 2000.

[3] M. Celik, G. Sharma, A. Tekalp, and E. Saber, "Reversible data hiding," in *Proceedings. International Conference on Image Processing*, vol. 2, pp. II–II, 2002.

[4] N. Kumar, R. Kumar, and R. Caldelli, "Local moment driven pvo based reversible data hiding," *IEEE Signal Processing Letters*, vol. 28, pp. 1335–1339, 2021.

[5] Z. Ni, Y.-Q. Shi, N. Ansari, and W. Su, "Reversible data hiding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 3, pp. 354–362, 2006.

[6] M. Sahu, N. Padhy, S. S. Gantayat, and A. K. Sahu, "Local binary pattern-based reversible data hiding," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 4, pp. 695–709, 2022.

[7] R. Kumar, S. Chand, and S. Singh, "An optimal high capacity reversible data hiding scheme using move to front coding for lzw codes," *Multimedia Tools and Applications*, vol. 78, no. 16, pp. 22977–23001, 2019.

[8] W. Puech, M. Chaumont, and O. Strauss, "A reversible data hiding method for encrypted images," in *Security, forensics, steganography, and watermarking of multimedia contents X*, vol. 6819, pp. 534–542, SPIE, 2008.

[9] X. Zhang, "Reversible data hiding in encrypted image," *IEEE signal processing letters*, vol. 18, no. 4, pp. 255–258, 2011.

[10] X. Zhang, "Separable reversible data hiding in encrypted image," *IEEE transactions on information forensics and security*, vol. 7, no. 2, pp. 826–832, 2011.

[11] W. Hong, T.-S. Chen, and H.-Y. Wu, "An improved reversible data hiding in encrypted images using side match," *IEEE signal processing letters*, vol. 19, no. 4, pp. 199–202, 2012.

[12] Z. Qian and X. Zhang, "Reversible data hiding in encrypted images with distributed source encoding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 636–646, 2015.

[13] J. Zhou, W. Sun, L. Dong, X. Liu, O. C. Au, and Y. Y. Tang, "Secure reversible image data hiding over encrypted domain via key modulation," *IEEE transactions on circuits and systems for video technology*, vol. 26, no. 3, pp. 441–452, 2015.

[14] P. Puteaux, S. Ong, K. Wong, and W. Puech, "A survey of reversible data hiding in encrypted images–the first 12 years," *Journal of Visual Communication and Image Representation*, vol. 77, p. 103085, 2021.

[15] K. Ma, W. Zhang, X. Zhao, N. Yu, and F. Li, "Reversible data hiding in encrypted images by reserving room before encryption," *IEEE Transactions on information forensics and security*, vol. 8, no. 3, pp. 553–562, 2013.

[16] X. Zhang, J. Long, Z. Wang, and H. Cheng, "Lossless and reversible data hiding in encrypted images with public-key cryptography," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1622–1631, 2015.

[17] X. Cao, L. Du, X. Wei, D. Meng, and X. Guo, "High capacity reversible data hiding in encrypted images by patch-level sparse representation," *IEEE transactions on cybernetics*, vol. 46, no. 5, pp. 1132–1143, 2015.

[18] S. Yi and Y. Zhou, "Binary-block embedding for reversible data hiding in encrypted images," *Signal Processing*, vol. 133, pp. 40–51, 2017.

[19] P. Puteaux and W. Puech, "An efficient msb prediction-based method for high-capacity reversible data hiding in encrypted images," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1670–1681, 2018.

[20] Y. Puyang, Z. Yin, and Z. Qian, "Reversible data hiding in encrypted images with two-msb prediction," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, 2018.

[21] A. Mohammadi, M. Nakhkash, and M. A. Akhaee, "A high-capacity reversible data hiding in encrypted images employing local difference predictor," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2366–2376, 2020.

[22] F. Ren, Z. Wu, Y. Xue, and Y. Hao, "Reversible data hiding in encrypted image based on bit-plane redundancy of prediction error," *Mathematics*, vol. 11, no. 11, p. 2537, 2023.

[23] P. Puteaux and W. Puech, "A recursive reversible data hiding in encrypted images method with a very high payload," *IEEE Transactions on Multimedia*, vol. 23, pp. 636–650, 2020.

[24] C. Yu, X. Zhang, X. Zhang, G. Li, and Z. Tang, "Reversible data hiding with hierarchical embedding for encrypted images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 451–466, 2022.

[25] K. Chen and C.-C. Chang, "High-capacity reversible data hiding in encrypted images based on extended run-length coding and block-based msb plane rearrangement," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 334–344, 2019.

[26] S. Yi and Y. Zhou, "Separable and reversible data hiding in encrypted images using parametric binary tree labeling," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 51–64, 2019.

[27] Y. Wu, Y. Xiang, Y. Guo, J. Tang, and Z. Yin, "An improved reversible data hiding in encrypted images using parametric binary tree labeling," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 1929–1938, 2020.

[28] Z. Yin, Y. Xiang, and X. Zhang, "Reversible data hiding in encrypted images based on multi-msb prediction and huffman coding," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 874–884, 2020.

[29] G. Gao, L. Zhang, Y. Lin, S. Tong, and C. Yuan, "High-performance reversible data hiding in encrypted images with adaptive huffman code," *Digital Signal Processing*, vol. 133, p. 103870, 2023.

[30] H. Gao, X. Zhang, and T. Gao, "Hierarchical reversible data hiding in encrypted images based on multiple linear regressions and multiple bits prediction," *Multimedia Tools and Applications*, pp. 1–27, 2023.

[31] F. Chen, Y. Yuan, H. He, M. Tian, and H.-M. Tai, "Multi-msb compression based reversible data hiding scheme in encrypted images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 905–916, 2021.

[32] Z. Yin, Y. Peng, and Y. Xiang, "Reversible data hiding in encrypted images based on pixel prediction and bit-plane compression," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 2, pp. 992–1002, 2022.

[33] Y. Yang, H. He, F. Chen, Y. Yuan, and N. Mao, "Reversible data hiding in encrypted images based on time-varying huffman coding table," *IEEE Transactions on Multimedia*, 2023.

[34] Y. Liu, G. Feng, C. Qin, H. Lu, and C.-C. Chang, "High-capacity reversible data hiding in encrypted images based on hierarchical quadtree coding and multi-msb prediction," *Electronics*, vol. 10, no. 6, p. 664, 2021.

[35] S. Xu, J.-H. Horng, C.-C. Chang, and C.-C. Chang, "Reversible data hiding with hierarchical block variable length coding for cloud security," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[36] Y. Yao, K. Wang, Q. Chang, and S. Weng, "Reversible data hiding in encrypted images using global compression of zero-valued high bitplanes and block rearrangement," *IEEE Transactions on Multimedia*, 2023.

[37] T. Markas and J. Reif, "Quad tree structures for image compression applications," *Information Processing Management*, vol. 28, no. 6, pp. 707–721, 1992. Special Issue: Data compression for images and texts.

[38] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The loco-i lossless image compression algorithm: Principles and standardization into jpeg-ls," *IEEE Transactions on Image processing*, vol. 9, no. 8, pp. 1309–1324, 2000.

[39] A. G. Weber, "The usc-sipi image database: Version 5," *http://sipi. usc. edu/database/*, 2006.

[40] P. Bas, T. Filler, and T. Pevnỳ, "" break our steganographic system": the ins and outs of organizing boss," in *International workshop on information hiding*, pp. 59–70, Springer, 2011.

[41] P. Bas and T. Furon, "Image database of bows-2," *Accessed: Jun*, vol. 20, pp. 2016–2017, 2017.

[42] T. Chu, W. Ranson, and M. A. Sutton, "Applications of digital-image-correlation techniques to experimental mechanics," *Experimental mechanics*, vol. 25, no. 3, pp. 232–244, 1985.

[43] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[44] Y. Wu, J. P. Noonan, S. Agaian, *et al.*, "Npcr and uaci randomness tests for image encryption," *Cyber journals: multidisciplinary journals in science and technology, Journal of Selected Areas in Telecommunications (JSAT)*, vol. 1, no. 2, pp. 31–38, 2011.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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 Multimedia. 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/TMM.2023.3325993

Ankur received the M. Tech. degree from the National Institute of Technology, Jalandhar, India, in 2013. He is currently pursuing the Ph.D. degree from Delhi Technological University, Delhi, India. His research interests include image processing, steganography and data hiding in encrypted domain.

Rajeev Kumar received his Ph.D. in Computer Engineering from the University of Delhi, India, in 2017. He did Post-Doctorate from Kyungil University, South Korea. Currently, he is working as an Assistant Professor in the Department of Computer Science and Engineering, Delhi Technological University (Formerly Delhi College of Engineering), Delhi, India. He has been a two-time recipient of the prestigious Brain Pool Fellowship of the National Research Foundation (NRF) of Korea, and has worked as an invited scientist with them. His research interests include steganography, reversible data hiding, multimedia forensics, Image Processing, compression, and wireless sensor networks.

Ajay K Sharma obtained his Ph.D. in Electronics Communication and Computer Engineering from NIT, Kurukshetra, India in 1999. Currently, he is serving as the Director of NIT, Delhi and Senior Professor (presently on lien) in the Department of Computer Science and Engineering at NIT, Jalandhar, India. He is also a former Vice Chancellor, IKGPTU, Jalandhar, India and former Director, NIT Delhi, NIT Hamirpur and IIIT Una, India. He carries over more than three decades of experience in teaching, research and various leadership roles in Central Government/State Government of India and has supervised 38 Ph.D. and 46 M.Tech. thesis in the area of Electronics and Communication Engineering and Computer Science Engineering. His research interests include optical systems and networks, wireless communication systems, wireless security and data hiding. He has published around 400 research articles in International and National Journals/Conferences of repute and has also authored 12 books. He has eight R&D projects. He is a fellow member of ISTE, OSA, SPIE USA, CSI Mumbai, IETE New Delhi. He has various International research collaborations and visited several countries such as the USA, UAE, Singapore and Malaysia etc.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:06:32 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.

MAT JOURNALS

Journal of Construction and Building Materials Engineering
e-ISSN: 2581-6454
Volume 5 Issue 2

Comparative Study on Traditional Timber Bridges of Bhutan with Timber Bridges of Other Nations

[1]Raju Sarkar, [2]T. Vijaya Kumar, [3]Kiran Gurung, [3]Dipti Sharma, [3]Tashi Choki, [3]Kezang Wangmo, [3]Kezang Tshomo, [3]Sonam Penjor, [4]Abhirup Dikshit, [4]Ritesh Kurar

[1,2]Associate Professor, Department of Civil Engineering, Delhi Technological University, Bawana Road, Delhi, India

[1]Cordinator and Head, [3]Undergraduate Researcher, [4]Visiting Researcher

Center for Disaster Risk Reduction and Community Development Studies, Royal University of Bhutan, Rinchending, Bhutan

Email: *[1]sarkar_raju@yahoo.com*

Abstract

This paper has assorted multifarious studies depicting the behaviour of timber bridges to administer a literature review. It focuses on weaknesses of the timber bridges under seismic loading as well as methodologies adopted to overcome them. The behaviour of reinforced joints of timber bridges, the serviceability limits of vibration induced by earthquake loading and the preference of suitable bearings such that the seismic capacity of timber bridges can be bolstered is summarized. Additionally, the authors have studied the effect of dynamic behaviour undergone by the bridges under vehicular loading and its cost efficiency. Moreover, this paper has also included diverse experimental studies that construe the behaviour of existing bridges and the seismic behaviour of timber bridges from past earthquakes. The conclusions of the compiled studied researchers provided the repercussions of time and climate on the load-bearing capacity of the bridges. By regulating a comparative study between the behaviour of the seismically loaded former bridges and existing current bridges, the effects of the earthquakes yet to occur can be marginalized.

Keywords: *Bazam, rectification, seismic behavior, traditional architecture*

INTRODUCTION

Timber is likely the prime construction material to be used in for constructing bridges since Neolithic era. The oldest bridge, perchance, has been erected by arranging the trunks across the stream in girder fashion. Subsequently, multitude of the timber bridges that have been built across the globe consisted different adaptations of beams, cantilevers, arches and trusses [1]. Timber has been used as a raw material because of its availability in abundance [2]. Usage of timber for bridge construction provides diverse dividends because of its ease in construction under any weather conditions without any detrimental effect. The seasonal variation doesn't affect the timber bridges and are also able to bear overloads for brief timeframes because of its energy absorbing ability [2]. Despite of the benefits, timber has its disadvantages. It is vulnerable to fungi, fire accidents and insects which damage the bridges. The deeper sections reduce the hydraulic operation and flood flow capacity beneath the bridge [2].

Traditional timber bridges have always been an essential part of the Bhutanese architectural tradition. Timber has been the base material for construction of bridges in the country, principally because of its ease of construction, abundance, cost efficiency and environmental friendliness. The first cantilever timber bridge known as "Bazam" was constructed in Punakha in the later 16th century for pedestrian usage.

Page 51-62 © MAT Journals 2019. All Rights Reserved

It extended up to a maximum length of 32 meters with two stone-masonry abutments at each end. The wooden horizontal deck was supported by inclined cantilever arrangements with varied angles. The horizontal component of the bridge is covered with wooden shingles (two or three-layered depending upon the climate and purpose).

Even after the introduction of better, high strength construction materials like concrete and steel, timber remains an important component in traditional bridge construction as it holds significant cultural values. Since, Bhutan lies in seismic zone IV and V according to Indian seismic codes [3], it is of utmost important concern that bridges constructed entirely based on thumb rule without the use of standard codes and procedures needs to be monitored. Several researchers have summarized the results on timber bridges [2, 4]. This paper summarizes the use of timber bridges in Bhutan as well as in global context. The paper is organized as follows. In section 2, we explain the traditional timber bridges of Bhutan. Section 3 describes the use of timber bridges across the globe. Thereafter, (Section 4) explains the current state of Punakha Bazam in Bhutan. Finally, (Section 5) concludes the shortcomings of timber bridges in Bhutan and the need for a technological transfer.

Components of Traditional Timber Bridges of Bhutan

The timber bridges in Bhutan are a simple arrangement of cantilevers supported on either side of the river over which the horizontal beams are placed (Fig. 1a). The need for such infrastructure increased over time and more advanced bridges were developed but without the use of nails and codes. Being a Buddhist country, the Bhutanese people have generally constructed bridges with traditional components such as traditional abutment houses on the ends of the bridge and shingle roof over the deck (Fig. 1b, 1c).

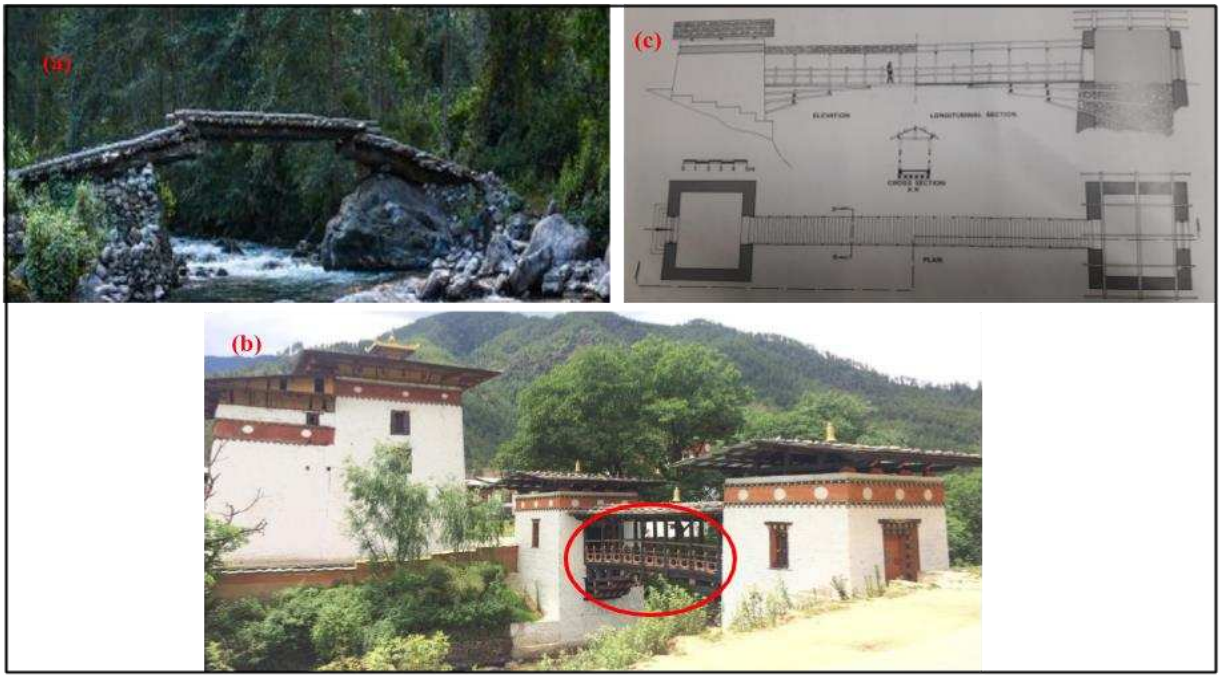


Figure 1: *(a) Simple cantilever bridge of Bhutan (Source: Getty Images), (b) Traditional timber bridge of Bhutan (Photography: Kiran Gurung), (c) Plan and structural drawing of traditional timber bridge of Bhutan (Source: Pro Bhutan).*

Page 51-62 © MAT Journals 2019. All Rights Reserved

A traditional timber bridge generally has following construction components:

WOODEN CANTILEVER SUPPORTS

The cantilever arrangement of the bridge is embedded into the stone rubble of abutment houses with an external projection of about 1.5-2 meters. The angle of the cantilever is varied in accordance with the horizontal plane of the earth surface. The arrangement functions by reducing the length of the span of the bridge and by providing supports on both ends of the bridge (Fig. 2a).

STONE AND POUNDED MUD ABUTMENT HOUSE

The abutment house is generally a single room with single storied structure constructed on each end of the traditional timber bridge wherein the cantilever arrangements are embedded. Smaller stones are used on the outer faces of the wall whereas the inner parts are filled with a random arrangement of mud and smaller pebbles. Stone filling below the deck level has been intellectually executed, mainly because the stone had been the only easily accessible heavy material, in order to balance with the load of the horizontal component.

WOODEN DECK

Deck of the bridge is placed over the cantilever arrangements upon which wooden planks are fixed perpendicularly to the beams making the floor of the deck (Fig. 2b).

Wooden Railings (Jadhang Tazi)

These are the hand railings provided along the sides of the deck of the bridge which provides safety and ease in access (Fig. 2b).

Wooden Columns

These columns are fixed and erected from the deck to support the trusses of shingle roof (Fig. 2b).

Wooden Shingle Roof (Jamthok)

The roof is elevated open to downplay the dragging effect of fast winds as well as for better and free circulation of wind. The stones placed are typical of oblate shape (Fig. 2c).

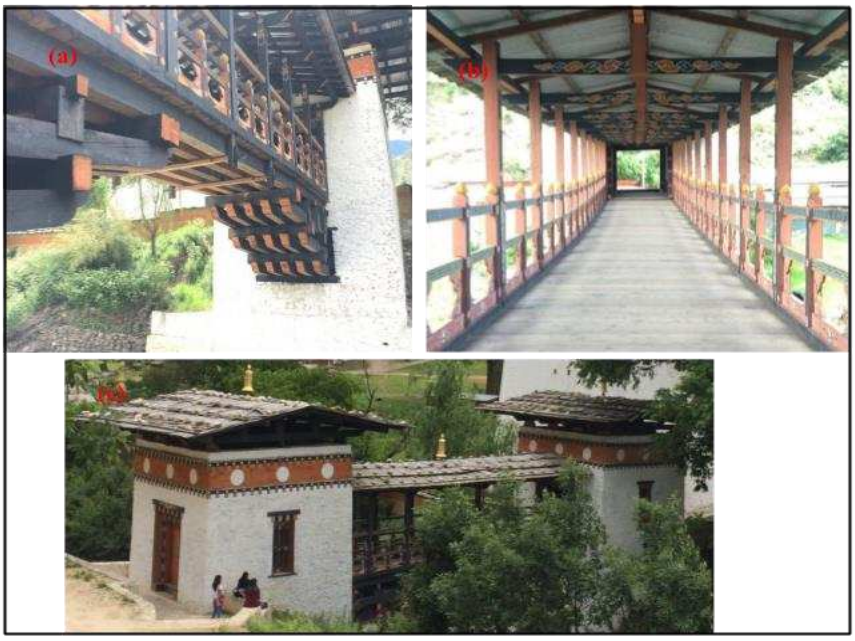


Figure 2: *(a) Wooden Cantilever Supports, (b) Wooden Deck, Railings, and Columns, (c) Shingle Roof (Photography: Kiran Gurung).*

Page 51-62 © MAT Journals 2019. All Rights Reserved

LITERATURE REVIEW ON TIMBER BRIDGES GLOBALLY

This section deals with a timeline of the prominent research conducted on timber bridges all across the globe. It also explains the various tests conducted by researchers and their conclusions. Such study would help in understanding the current state of several timber bridges across Bhutan and formulate a way to either improve the present structure or to move towards technological advancement.

Ritter et al. [5] conducted field performance of stress-laminated timber bridges on low-volume roads by monitoring 24 bridges over a period of 2 years and concluded that bridges stressed with a single jack, 3 to 6 stressing passes should be made along the bridge length. The stress level should be gradually increased to minimize the deck distortion. They observed that the attachments of the bridge superstructure and substructure should not be made until the bridge has been fully stressed twice. In addition, the average moisture content of the wood lamination should be 10%-16% and never exceed 20% during the construction period. The authors developed numerous conclusions, like for sawn lumber bridges, bar force needs to be analyzed by the year for the first 2 years of working life and every 2 years thereon. Later on, it may be extended to 2-5-year intervals once bar force stabilize. For glued laminated timber bridges, bar force is needed to be analyzed biennially for the first 4 years of working life and every 5 years thereon. Bridge live-load deflection should check the maximum of 1/360 to 1/400 of the bridge span. When oil-type wood preservatives are used, the preservative retention should fulfill American Wood Protection Association Standard (AWPA). Stressing bars enclosed in grease-filled plastic tubes should be taken into account if the bridge is subjected to corrosion.

Wipf et al. [6] prepared report on the dynamic field performance of timber bridges. The report included American Association of State Highway and Transportation Officials (AASHTO) recommendations of dynamic load allowance of 0.165 for timber bridges. Thereafter, researchers selected five glued-laminated timber girder bridges and four longitudinal glued-laminated timber panel bridges for testing. They discerned that nine tested bridges have fundamental frequencies between 5Hz and 11Hz as well as a dynamic load allowance of less than 0.25. It was also found that nearly all the timber bridges tested have natural frequencies between the body bounce and axle hop vehicle frequencies. This frequency grouping is mainly due to the bridges all having short span lengths. The report concluded that the timber bridges have Dynamic Amplification Factor (DAF) values less than steel values (Fig. 3).

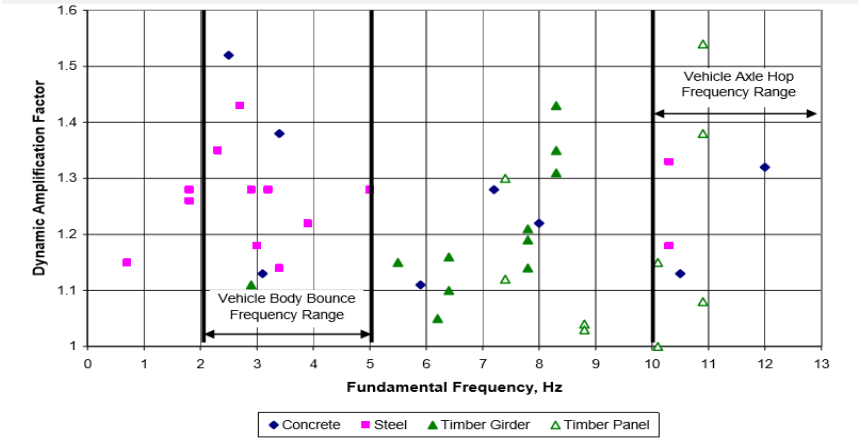


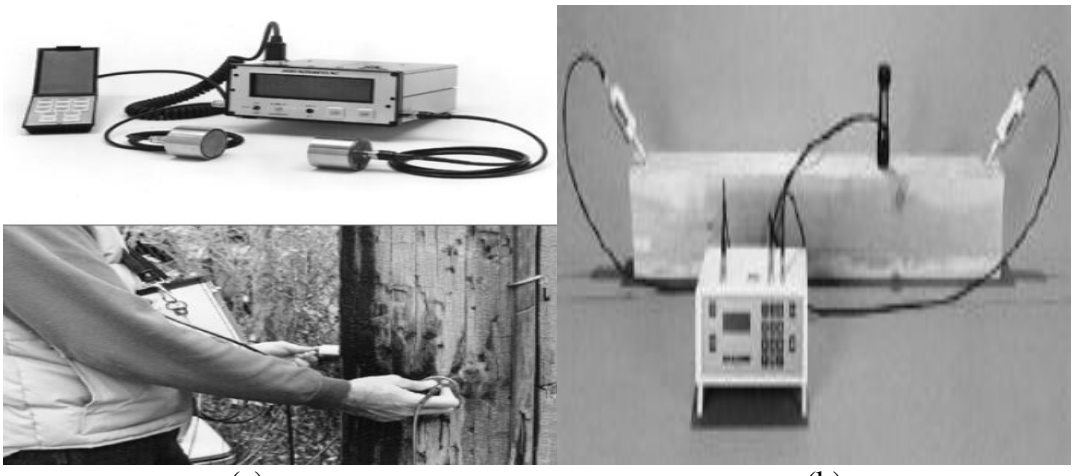
Figure 3: *DAF for bridges constructed of various materials [6].*

Page 51-62 © MAT Journals 2019. All Rights Reserved

Wood et al. [7] studied the dynamic response of three glued laminated (glulam) timber bridges wherein static and dynamic field test were performed using a heavily loaded truck. Bridge deformations (deflections and accelerations) were analyzed for diverse vehicular speeds at the bridge midspan. The deformations (deflections and accelerations) were reported using high-speed data acquisition system and DAF computations. The intent of the study was to determine the dynamic behavior of glulam timber bridges so that reliable design specifications may be established.

Spyrakos et al. [8] conducted the seismic study of historic covered bridges in West Virginia by analyzing the structural modelling of the bridges. The static analysis has been carried out with various loading to represent the 19th century loading and the modern loading as per AASHTO. For the seismic analysis the bridge should be characterized based on seismic performance category SPC=A and SPC=B for A=0.19, SPC=C for A=0.25 as per AASHTO specifications, whereas the bridge of the region has been modelled to 3D beam in accordance to the guidelines given by the authors. The authors have conducted numerous tests to attain the static analysis of the modelled bridges such as acoustic emission testing, ultrasonic field-testing and in-situ testing. The result obtained from the static analysis and model analysis included results of selected members of the bridges, such as Top-chord, Diagonal, Vertical, Arch member or Top-cross beam. The authors also listed the six lowest natural frequencies from the 3D model analysis.

Ross et al. [9] used stress wave timing nondestructive estimation tools to investigate timber bridges. It ascertained that the role of the vehicular bridges is very important for the safe and efficient working of the transportation. The bridges have an average life cycle of 40 years and thereafter it needs to be either repaired or replaced by assessing its condition. The inspection of bridges could be done by hammering, sounding, coring or drilling and another nondestructive testing such as James V-Meter (Fig. 4a), Sylva Test (Fig. 4b), and using stress wave timbers.



(a) (b)

Figure 4: *(a) James V-Meter, (b) Sylva Test [9].*

Mander et al. [10] studied seismic performances of timber bridges where parameters like strength, ductility, seismic vulnerability were analyzed experimentally. The study examined seismic resistance of bridge piles. Experiments were conducted in

Page 51-62 © MAT Journals 2019. All Rights Reserved

two phases, firstly, combined axial load and reversed cyclic flexural tests on piles, secondly; tabular study of shaking behavior of timber piers. Fig. 5 illustrated the results of the tests conducted with four different timber pile specimens.

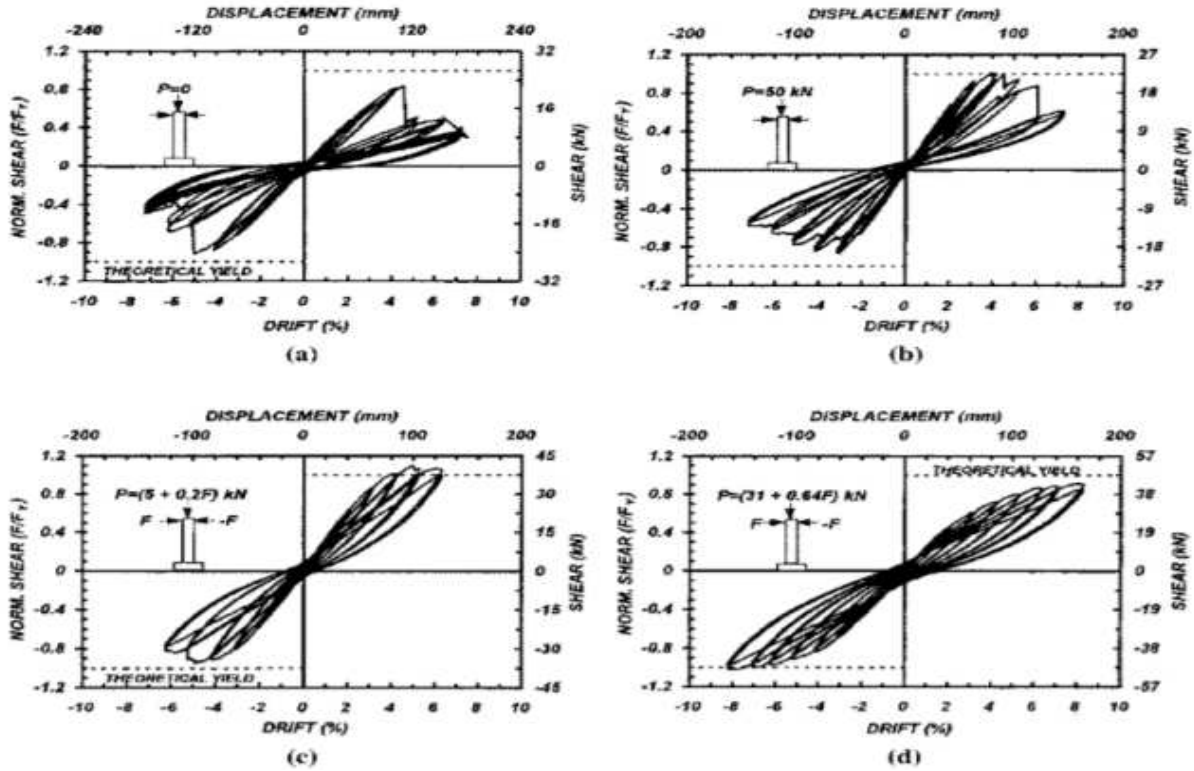
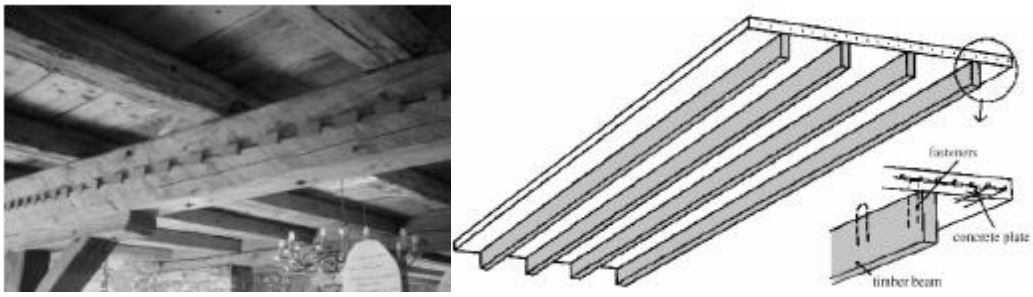


Figure 5: *Experimentally observed force-drift relationships for four timber pile specimens [10].*

Mueller et al. [11] performed an experimental and numerical study on long term performances of hybrid bridges. Three different types of connectors were experimented to understand the climatic dependency for serviceability and load bearing capacity with the shear test under long and short time and dynamic loading. The authors determined stiffness and ultimate load by a short time, fatigue behavior by dynamic loading and variation of moisture by a long time. The study showed that timber tends to creep more with the change in humidity. The increase in moisture content causes swelling whereas a decrease causes shrinkage. However, concrete creeps and shrinks more in dry conditions compared to timber. Ceccotti [12] studied composite concrete-timber structures and stated that they are better compared to archaic bridges. Fig. 6a depicts the cross-sectional dimensions of structural timber bridges which were enlarged through laminating techniques using mechanical joins to create all-wood composite members. Concrete-timber composite systems (Fig. 6b) are reinforced with concrete slab layer stressed in compression and timber ribs in tension which utilizes the strength and stiffness of both materials. By the virtue of its lower density, it is far more efficient than all reinforced concrete structures in terms of a load carried per unit self-weight. Fig. 7 shows the relationship between self-weight and span for various types of floor construction. The study also showed that such bridges have good insulation, fire resistance and better architectural insight.

Page 51-62 © MAT Journals 2019. All Rights Reserved



(a) (b)

Figure 6: *(a) Composite members in old timber constructions, (b) Possible concrete-timber solutions for floors [12].*

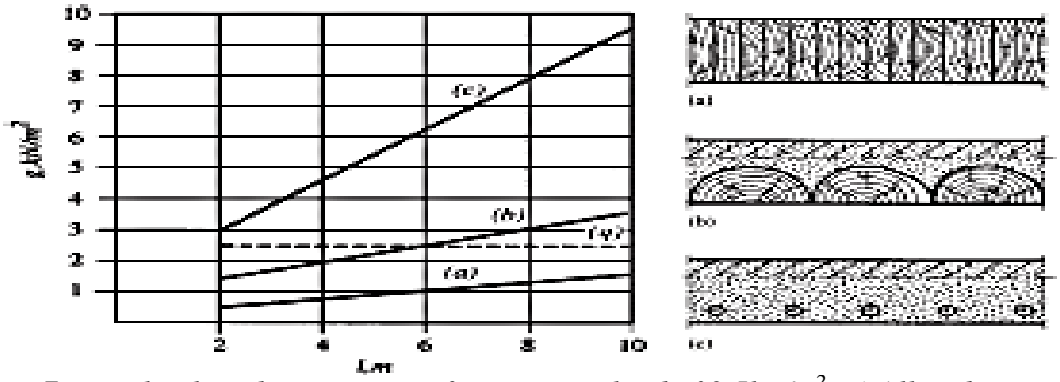


Figure 7: *Weight plotted against span for a service load of 2.5kN/m^2. a) All timber section; b) Concrete-timber composition section; c) All concrete section [12].*

Seavey and Larson [13] inspected the timber bridges to assess the lifespan of existing 19th century, and critically judge whether they should be retrofitted or rebuilt. The study stated that such bridges should undergo a thorough inspection. A system POINTIS computer is used for the evaluation of the bridges elements, as it requires accurate reliable information. For timber bridges, it requires information such as its moisture content, preservatives, span, legs support, and depth of degradation and so on.

Mckeel and Gomez [14] extensively researched the modern timber bridges in Virginia addressing the costs, structural performance, and maintenance requirements of these structures. The study deduced that the timber bridges were not economically competitive, and their life cycle cost competitiveness cannot be resolved in the moment. They concluded that most of the bridges constructed under the auspices of the initiative have performed well structurally, and only in the case of the Fredericksburg Bridge has any serious distress, possibly related to the treatment process rather than design. They also discerned that there is no particular relationship among the timber bridges between span length and cost per sq. m.

Brunner [15] studied the design of a timber bridge in Switzerland. This design work of the bridge has forms of two continuous, simply supported beams of 10.5m span (Fig. 8a). The substructures abutment and the central column were constructed with gabions from galvanized steel wires imported from France (Fig. 8b). He observed that the gabions is cheap alternative and is technically advantageous in being insensitive to erosion hazards. The central column is about 3m high with its foundation on rock in the middle of the

Page 51-62 © MAT Journals 2019. All Rights Reserved

river. It has been protected by surrounding it with a belt of timber logs filled with rock material. The foundations of the abutments have been protected with hewn rock blocks placed on the sides facing the river (Fig. 8c).

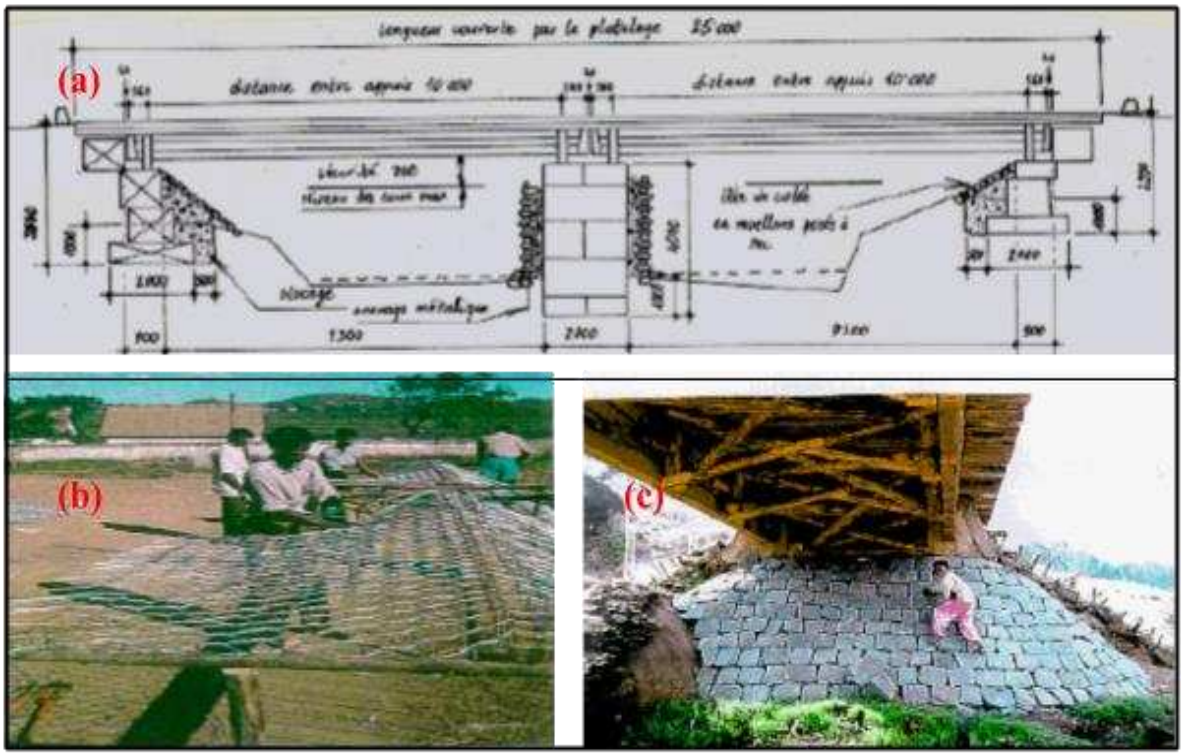


Figure 8: *(a) Structural System of the pilot bridge, (b) Weaving wires for gabions, (c) Erosion protection [15].*

Parisi and Piazza [16] undertook research on seismic strength of traditional timber structures. The authors recognized that experimental and numerical analyses of fully strengthened trusses depicted a discernable seismic response by the structure in the post elastic range. Timber faces permanent deformation when load is applied perpendicular to fibers but yields higher mechanical properties when compressed and load is applied parallel to the fibres. Different possible reinforcing solutions have been incorporated such as insertion of bolts, placing of steel stirrups on the connection sides, and steel binding strip. It also stated that for seismic upgrading of joints, metal fasteners have to be designed such that they can ensure functionality in a situation of reduced compression that may develop under seismic loading. It concluded that the timber-concrete composite structure offers the highest mechanical characteristics.

Ghosh et al. [17] conducted a study comparing thermal and seismic behavior of a bridge constructed with traditional and modern isolation bearings. The authors observed that designing of bearings is essential for the safety of bridges for earthquake displacements and resistance. The bearings have to be designed considering the long-term effects due to temperature, shrinkage and creep. It has been discerned that the modern isolation bearings have superior thermal and seismic response than traditional bearings, wherein latter requires revolutionary change in the concepts of bearing design. These bearings have been also designed to ensure that superstructure is isolated from the vibrations transmitted by the substructure. Fig. 9 illustrates the various types of arrangement of bearings were used.

Page 51-62 © MAT Journals 2019. All Rights Reserved

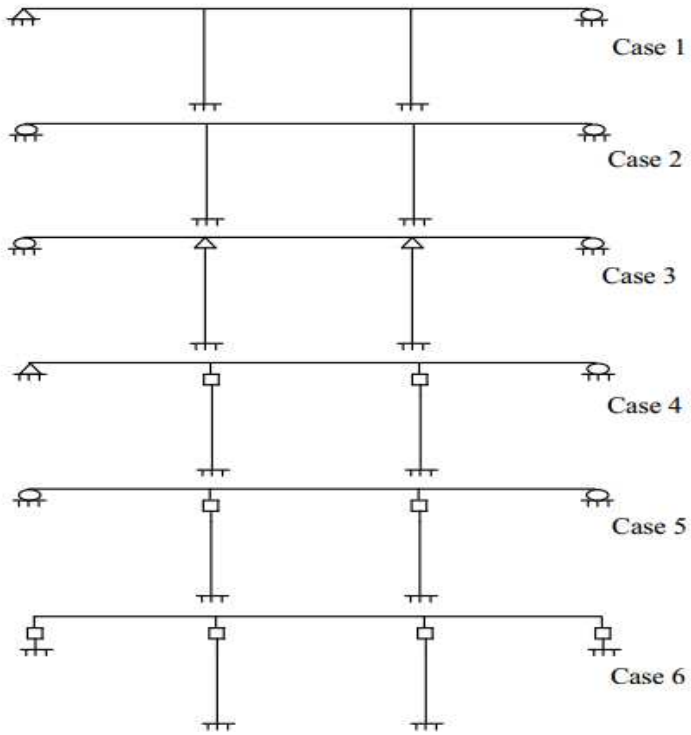


Figure 9: *Six bearing arrangements considered [17].*

Morris [18] studied several archaic wooden bridges in North America and from various parts of the world like Å set Bridge, Flisa Bridge, Sletta Bridge in Norway. The paper recommends selection of material and durability design according to Canadian Highway Bridge Design Code. It further stated that all wood in permanent structures shall be preservative treated in accordance with above design code or American Wood Protection Association. The author identified that relying entirely on preservative treatment and ignoring design is equally risky. Priory, across the globe, the longevity of many non-covered timber bridges mainly depended upon bulky application of creosote. But due to escalating environmental concerns provoked by application of creosote, numerous new timber bridges are being dealt by using the management practices tool developed by the Western Wood Preservers Institute, Wood Preservation Canada and Southern Forest Products Association.

Sandra et al. [19] mentioned that there are two techniques of preservation of the timber framed structures from earthquake. The first technique known as "Borbone system" developed in southern Italy and second technique is Greek earthquake resistant solution called "Lefkada traditional building technique". In the Borbone system, the timber frame extends all over the floors, stiffness for wall is mainly provided by filling material inserted within frames and bracing if any and the roof is supported by a truss system which doesn't transfer any thrust to the walls and structural joints and is fixed by nails. In the Lefkada system, the structures are multi storied building with masonry walls at ground floor and upper floor are held by timber frame only and stiffened by diagonal members. In this system, the timbers are provided parallel to the wall which supports the first floor slab but doesn't contribute to the seismic resistance. However, it prevents the building from total collapse and repairs can be done without interfering with the rest of the building.

Page 51-62 © MAT Journals 2019. All Rights Reserved

When compared to the significant work being done for timber bridges across the globe, the bridges in Bhutan lack a technological advancement. Several bridges have not been designed or retrofitted for seismic conditions. All the Bhutanese timber bridges are cantilever bridges with 3 to 4 cantilever beams placed depending on the span length of the bridge. Moreover, bearings are not provided for making the bridges spring supportive which could withstand seismic loadings. Bhutan has a history of earthquakes and it is very important that effective measures are needed to be taken to prevent any calamity.

CURRENT STATE OF PUNAKHA BAZAM

Bhutanese bridges with its ancient wisdom have withstood any disaster and will remain standing for upcoming 30 to 40 years. Some minor retrofication has been done on some of the bridges like replacement of wooden planks, plastering the mud walls for strength with cement and concrete, inserting L-straps in the joints for rigidity and double layering the roofs with CGI sheets. However, a major retrofication has been done on the whole structural designing of Punakha Bazam.

Cantilever Beam

The cantilever beam which lends support from the stones piled up in the abutment house is sliced vertically throughout its length. The timber used in bridge construction is seasoned but there are still rooms for shrinkage with the aging of timber (Fig. 10a). So, the timber is sliced vertically in the middle to leave room for expansion and shrinkage of timber.

Diagonal Planking

Diagonal planks are nailed to the beam for efficient force transfer (Fig. 10b). The planks add to the strength of the beam and protect the upper horizontal planks.

Reinforcement (Small Screws)

The use of small screws as reinforcement for the wooden beams, wherein, the connection of two beams make a deep beam (Fig. 10c). Not only deep beams can carry higher loads than diagonal cracking loads, moreover, their flexural behavior is no longer critical depicting negligible deflection. (Guan and Kong, 1994).

Placement of Stone at the River Bank

Stones and boulders piled at the river bank on the water front of the abutment wall to reduce the force of water acting on the wall and reduce scourge of the same (Fig. 10d).

Figure 10: *(a) Cantilever Beam, (b) Diagonal Planks (Source: Ministry of Home and Cultural Affairs), (c) Reinforcement, (d) Placement of stones at the river bank (Photography: Deepti Sharma).*

Page 51-62 © MAT Journals 2019. All Rights Reserved

CONCLUSIONS

Customary Bhutanese timber spans with all elaborative Bhutanese convention and culture have assumed a fundamental part in facilitating the development of Bhutanese native. Indeed, even with the initiation of technological advances, traditional bridges have dependably been a pride and landmark of Bhutan. Some have been reconstructed and others are timely monitored and retrofitted. Bhutan faced the highest magnitude of earthquake of 6.1 in the year 2009 and faces such disaster in the near future. The summary of the paper are as follows:

- Since, Bhutan is situated in high seismic zone it is imperative not only to retrofit existing bridges, but also to develop seismically resistant new bridges.
- The existing timber bridges need to see a technological advancement either through the reconstruction using concrete or retrofitting the existing structure keeping the aesthetic value.
- Efficient retrofit techniques can be adopted such as, heartwood of naturally durable woods having sufficient decay resistance for the anticipated and potential exposure to moisture with appropriate design are used. Modern glue-laminated timbers are fabricated in a dry condition, thenceforth; oil-type preservatives should be used.
- The approach used in Norway may also be used, wherein, a lamina which has been pressure treated with a waterborne copper-based preservative is fabricated into beams. The beams are cut to distance and all connections are ready by stress considering the bridge modules with creosote, along with woody modules spread out apart by metal connectors.
- For stronger effective stiffness bridges, concrete timber bridges can be considered. It gives pleasing aesthetic appearance and has an economic construction advantage.

REFERENCES

1. Echarri, V., Brebbia, C. A., (2018), “Heritage Architecture Studies”, ISBN: 978-1-78466-273-8.
2. Ou F. L., Weller C. (1986), “An Overview of Timber Bridges Transportation Research Record”, Volume 1053.
3. Indian Standard: (1893), Part I, “Criteria for Earthquake Resistant Design of Structures: General Provisions for all structures and specific Provisions for Buildings”, *Bureau of Indian Standards,* New Delhi, India, 2016 (Sixth Revision of IS 1893 Part 1).
4. Gutkowski, R.M., Williamson T.G. (1983), “Timber Bridges: State-of-the-art”, *Journal of Structural Engineering,* Volume 109, Issue 9. https://doi.org/10.1061/(ASCE)0733-9445(1983)109:9(2175).
5. Ritter, M. A., Wacker, J. P., Duwadi, S.R. (1995), “Field performance of stress-laminated timber bridges on low-volume roads”, *Proceedings of 6th International Conference on low-volume roads, Minneapolis,* MN. Washington, DC: National Academy Press, Volume 2, pp. 347−356.
6. Wipf, T., Phares, B., Bigelow, J., Wood, D., Ritter, M. (1996), “Dynamic field performance of timber bridges”, *United Department of Agriculture, Iowa State University*, pp. 1−24.
7. Wood, D. L., Wipf, T. J., Ritter, M. A., (1996), “Dynamic field performance of timber bridges”, *Semisequicentennial Transportation Conference Proceedings, Iowa State University*, Ames, Iowa, pp. 1−5.
8. Spyrakos, C.C., Kemp, E.L., Venkatareddy, R. (1999), “Seismic study of an historic covered bridge, Engineering Structures”, *West Virginia*

Page 51-62 © MAT Journals 2019. All Rights Reserved

University, Morgantown WV 26506-6103, USA, 1997, Volume 21, pp. 877-882.
9. Ross, R. J., Pellerin, R. F., Volny, N., Salsig, W. W., Falk, R. H. (1999), "Inspection of timber bridges using stress wave timing nondestructive evaluation tools", *United States Department of Agriculture,* pp. 1-17.
10. Mander, J. B., Allicock, D. R., Friedland, I. M. (2000), "Seismic performance of timber bridges", *Transportation Research Record 1740, Paper No. 00-1231, University at Buffalo, New Work*, pp. 75-84.
11. Mueller, J., Haedicke, W., Simon, A., Rautenstrauch, K. (2000), "Long-term performance of hybrid timber bridges-experimental and numerical investigations", *Article, University of Weimar, Bauhaus,* pp. 1-8.
12. Ceccotti, A., (2002), "Composite concrete-timber structures", *IUAV University,* Venice, Italy, pp. 264-275.
13. Seavey, R., Larson, T. (2002), "Inspection of timber bridges, Minnesota Local Road Research Board", *Department of Wood and Paper Science, University of Minnesota*, pp. 1-55.
14. Mckeel, W.T., Gomez, J.P. (2002), "The introduction modern bridges in Virginia: A final report on the Virginia timber bridge initiative", *Virginia Transportation Research Council in cooperation with the U.S. Department of Transportation;* Federal Highway Administration, Charlottesville, Virginia, pp. 1-12.
15. Brunner, M., (2003), "A timber bridge straight from the forest, XII World Forestry Congress", *Quebec City, Canada, Madagascar, Switzerland*, pp. 1-5.
16. Parisi, M. A., Piazza, M. (2004), "Seismic strength of traditional timber structures", *13th World Conference on Earthquake Engineering*, Paper No. 909, Vancouver, B.C, Canada, 1-6 August, 2004, pp. 1-14.
17. Ghosh, G., Singh, Y., Thakkar, S.K. (2008), "A comparative study of thermal and seismic behavior of a bridge with traditional and isolation bearings", *The 14th World Conference on Earthquake Engineering*, Beijing, China, Seismic Zone IV, Northern India, 12-17 October, pp. 1-9.
18. Morris, P. (2015), "Durability of wood bridges, Fujian and Zhejing provinces", *Hyde Hall Bridge, New York*, pp. 1-9.
19. Sandra T., Nicola R., Chesi C. (2016), "Comparison between two historical timber framed solutions representing both tradition and pre-modern earthquake resistance concepts", *Conference paper, 41st IAHS World Congress, Sustainability and Innovation for the future, Albuferia*, Algarve, Portugal, pp. 1-10.

Page 51-62 © MAT Journals 2019. All Rights Reserved

International Journal of Multimedia Information Retrieval (2023) 12:34
https://doi.org/10.1007/s13735-023-00301-9

REGULAR PAPER

ConvST-LSTM-Net: convolutional spatiotemporal LSTM networks for skeleton-based human action recognition

Abhilasha Sharma[1] · **Roshni Singh**[1]

Received: 11 February 2023 / Revised: 10 August 2023 / Accepted: 24 September 2023 / Published online: 27 October 2023
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

Abstract
Human action recognition (HAR) emphases on perceiving and identifying the action behavior done by humans within an image/video. The HAR activities include motion patterns and normal or abnormal activities like standing, walking, sitting, running, playing, falling, fighting, etc. Recently, it sparks the attention of researchers especially in 3D skeleton sequence. The actions of human can be represented via sequence of motions of skeletal keyjoints, although not all the skeleton keyjoints are informative in nature. Various approaches for HAR are used like LSTM, ConvLSTM, Conv-GRU, ST-LSTM, etc. Thus far, ST-LSTM approaches have shown tremendous performance in 3D skeleton sequence tasks but the detection of irrelevant keyjoints produce noise that deteriorates the performance of the model. So, the intent is to bring attention toward improving the efficacy of the model by focusing on informative keyjoint coordinates only. Therefore, the research paper introduces a new class of spatiotemporal LSTM approaches named as ConvST-LSTM-Net (convolutional spatiotemporal long short-term memory network) for skeleton-based action recognition. The prime focus of proposed model is to identify the informative keyjoints in each frame. The result of extensive experimental analysis exhibits that ConvST-LSTM-Net outperforms the state-of-the-art models on various benchmarks dataset, viz. NTU RGB + D 60, UT-Kinetics, UP-Fall Detection, UCF101, and HDMB51 for skeleton sequence data.

Keywords Activity recognition · LSTM · Spatiotemporal · Skeleton sequence · ConvLSTM

1 Introduction

Human action recognition has turn out to be a prominent & diligent research area in computer vision and image processing, which includes classification and recognition of normal & abnormal human activities of daily routine. It belongs to the automated recognition of human activity (normal & abnormal) in various application areas via analyzing the sequence of observations. Nowadays, crowded places with normal and abnormal activities are familiar due to population increase that turn toward suspicious activities. So, HAR has become an essential part in the automatic interpretation of human environment interaction in various online-offline applications such as auto-driving, intelligent surveillance [1–5], smart-gadgets analysis [6], object detection & tracking [7], video retrieval [8], and assisted daily living. Other HAR applications firmly coupled with the daily activities such as motion analysis [9–12], pose motion analysis [13, 14], health monitoring [15], classification or detection of actions or motions [16], and understanding human action behavior [17]. By recognizing and analyzing the human actions from the videos, one can clearly distinguish between normal and abnormal behaviors that can make significant improvements in public safety. Withal, HAR remains an ambitious challenge due to its clutter backgrounds, slight interclass segregation, and wide intra-class deviation. The main thing to recognize high accuracy & efficiency is to conquer both static appearances within each frame of the videos as well as temporal relationships throughout the multiple frames generated via videos. Some applications such as monitoring suspicious detection and early reporting for fall detection are also considered in human activity recognition. However, various techniques are there for the representation of human action based on motion, such as RGB-based

✉ Roshni Singh
roshnisingh1815@gmail.com

Abhilasha Sharma
abhi16.sharma@gmail.com

[1] Department of Software Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India

videos [18–20], RGBD-based videos [21–24], and skeleton-based videos [25–30]. Skeleton-based action recognition has become more prominent in recent times as it offers a focused and concise approach. The representation of human skeletons in videos typically involves a series of joint coordinates, which can be obtained through pose estimators or action prediction methods. By focusing solely on the action poses and disregarding contextual factors like background variations and lighting changes, skeleton sequences provide a compact and robust way to capture action information. On comparing these techniques, all the skeleton-based methods represent: (i) human motion via 3D coordinates positions for key-body points and (ii) are more robust to problems like variations of background clutter, observation viewpoints, illumination or intensity conditions, and so on. These advantages of skeleton motion sequences motivate researchers to develop new techniques for exploring informative features for human action recognition. These methods are gaining utmost importance in HAR since skeletons represent a compact sequence of data forms that depict dynamic motion within human body movements [31]. In respect of annotation, i.e., capturing the effective motion action representation from several unlabeled skeleton samples, manual annotation founds to be very expensive and challenging, nonetheless an unexplored area. These days, sensors are used for collecting data for their low-cost and high mobility. Some approaches are used for tracking and calculating the skeleton keyjoints with feature invariant to human key-body points, observation point, camera viewpoint, and so on.

Some approaches are used for tracking and calculating the skeleton keyjoints with feature invariant to human key-body points, observation point, camera viewpoint, and so on. The skeletal features within the human body are responsible for recognizing all the normal & abnormal activities. Besides this, they have also been used for evaluating (some activities such as falling, discriminating between jogging and running) the variation in the keyjoint coordinates between the center mass of the body, acceleration motion, velocity motion, for movement: angles between the keyjoint points within the skeleton. Some new methods like ST-LSTM and ST-GCN (spatiotemporal graph convolution network) are practices to extract these features also. The movement of body parts and the execution of various actions are made possible by the human skeletal system. When it comes to data modality, the use of skeleton-based information aligns with the structure of the human anatomy, which enhances the interpretability of ConvLSTM learning. This modality specifically focuses on 3D coordinates of keyjoints in the human body. By analyzing these skeleton sequences, the model is capable of recognizing and understanding human movements. Another advantage of the skeleton modality is its emphasis on privacy, as it is considered to be more privacy-friendly compared to other modalities.

In this work, the prime objective is to efficiently combine the important cues in CNN (convolutional neural networks), and LSTM using spatiotemporal data with skeleton-based recognition approaches call up as ST-LSTM. Here, a set of extracted skeleton features in conjunction with skeletal keyjoint is fed as input to the model. The skeletal tracking algorithms were used for detecting the keyjoints followed by the feature extraction that has been done through RGB frame data (extracted from videos) for improvising the efficiency of model. Some standard features, like angle between the keyjoint coordinates, velocity motion, acceleration motion, and human body position of the center of mass, for movement: angles between the joint key points, have been extracted from keyjoint coordinates of the human skeleton.

Once the feature extraction is done along with preprocessing thereupon, the preprocessed data are feed to the model, consisting of 17 extracted features among 25 skeleton coordinates. The overall pipeline of proposed ConvST-LSTM-Net model is illustrated in Fig. 1. The model exploits a spatiotemporal network consisting of CNNs, ST-LSTMs & fully connected dense layers. The model first detects the skeleton keyjoints of the persons using the skeleton-based recognition method. These keyjoints are fed to the CNN layers, followed by ST-LSTMs for the extraction of spatial–temporal features. Then, output from a hidden layer of ST-LSTMs is passed via FC dense layer (fully connected) for classification.

The key contributions of the research work can be summarized as follows:

1. A spatiotemporal ConvST-LSTM-Net model has been proposed that utilizes human body keyjoint coordinates from skeletal data obtained from RGB videos. The keyjoints are fed as an input to CNN layers for extracting the spatial–temporal features followed by ST-LSTM and output is passed to the time-distributed FC dense layer.
2. Motivated by the advances in CNN, ConvLSTM, and ST-LSTM, we have seamlessly combined the ideas of these models and integrated them to propose a new paradigm for skeleton-based action recognition termed as ConvST-LSTM-Net. The model brings the attention toward improving the efficacy by focusing only on informative keyjoint coordinates.
3. Among 25 keyjoints, a set of 17 extracted skeleton features along with 21 skeleton keyjoint coordinates are fed to the model as not all the skeleton keyjoints are informative in nature for recognizing the action classes.
4. The proposed ConvST-LSTM-Net model shows better performance in comparison to the existing models by using different modalities over various benchmarks, viz. NTU RGB + D dataset [32], UT-Kinect dataset [33], UP-Fall Detection [34], UCF101 [35], and HDMB51 dataset [36].

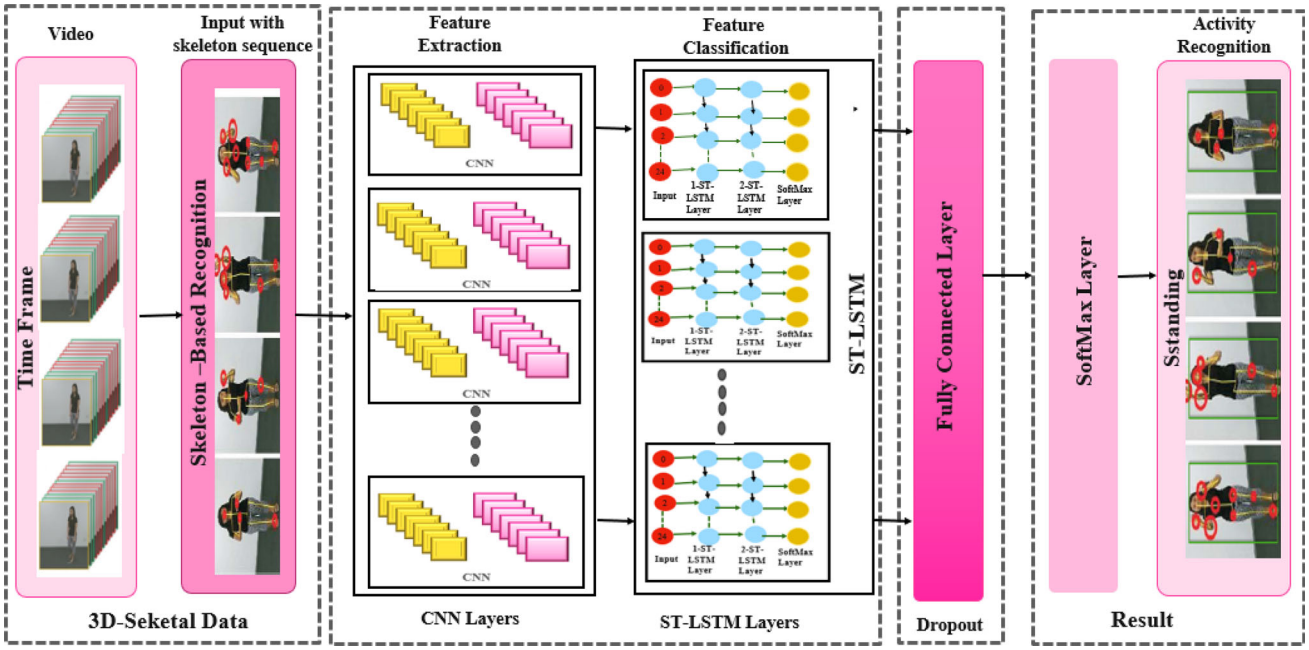


Fig. 1 Process informatics pipeline of ConvST-LSTM-Net. At first, the video frames are passed through the skeleton-based recognition feature method to extract the skeleton keyjoint coordinates. Then, the obtained keyjoint coordinates are fed to a modified ST-LSTM cell followed by ST-LSTM layers to evaluate the spatiotemporal feature. Further, the outputs are passed to FC dense layers. Ultimately, SoftMax shows the framewise prediction scores of human action behaviors

The rest of this paper is systematized as follows: Sect. 2 describes an overview of relevant studies available in literature. Section 3 briefly introduces the key terms and techniques along with the proposed methodology. Section 4 shows the experimental results and analysis. At last, the Sect. 5 discusses about the conclusion and future perspective.

2 Related work

This section briefly discusses the work related to skeleton-based RNN, LSTM & ST-LSTM for human action recognition.

2.1 Skeleton-based action recognition

Earlier, conventional skeleton-based recognition in the HRA modal aimed at the handcrafted features [37, 38], a well-known method for describing and classifying image features. However, it fails to classify the adequate semantic-labeling information of the human body. Deep learning approaches such as convolutional neural networks (CNNs) [39–43], recurrent neural networks (RNNs) [44–49], and graph convolutional networks (GCNs) [50] have achieved the exotic performance for learning more informative features about skeleton sequence, which helps in HAR learning. Many works have been introduced to achieve high performance of the skeleton sequence model. Li et al. [41] introduced a framework on convolutional co-occurrence feature learning that gradually works on hierarchical methods to aggregate contextual information on various levels. Vemulapalli et al. [37] designed a rolling map based on relative 3D rotations among different human body parts. Liu et al. [44] elongate RNN-based technique into the spatial–temporal model for revisiting the result based on the action-related performance of human action. Zhu et al. [43] introduced a cuboid-CNN in skeleton actions, ultimately concluding a human's normal keyjoint movements. Zhang et al. [51] implemented a view-adaptive modal for auto-regulate angle viewpoints during any motion action & obtaining different viewpoint observations of human actions. However, for skeleton sequences, these models fail to extract the temporal–spatial correlation configurations & even fail to explore the graphical aspects of human body structure. Due to the popularity of graph-based techniques, Yan et al. [52] introduced an approach based on GCN for the skeleton-based HAR, then introduced the ST-GCN method for featuring the spatial & temporal dynamics configuration of keyjoint skeletons of humans synchronously. Song et al. [53] worked on solving the occlusion issues and implemented multi-stream GCN for extracting qualified features for activated skeleton keyjoints in human action. Furthermore, they proposed a non-local technique [54] by using

2-stream GCN approach: 2 s-AGCN for improving recognition accuracy. Also, Shi et al. [55] worked on GCN fusion feature and proposed the multi-stream architecture at the output layer. Cheng et al. [56] worked on shift operation based on graph and used the point-wise convolutions connected layer for lowering its computational complexity. Ye et al. [57] introduced novel work on DGCN (dynamic graph convolutional network), an approach used for skeleton-based action recognition under 2-stream-AGCN, which features global dependency via achieving preeminent accuracy. Zhang et al. [58] also work on GCN in the spatial attentive–temporal dilated network for feature extraction in skeleton frame sequences using distant spatial attention weights and temporal scales. In 2-Stream network, Shi et al. [59] confiscated on bones, i.e., bone stream and joint stream, but entirely independent of each other. Furthermore, directed graph neural networks (DGNNs) [60], graph edge convolutional neural networks (GECNNs) [61] were introduced, which depict the relation among joints-bones in terms of action, but they fail to represent the various methods to combine features in the motion action transmission field.

2.2 Skeleton-based action recognition using LSTM's and RNN's approaches

Currently, the deep learning area mainly focuses on recurrent neural network (RNN)-based techniques, since it manifests its growth in skeleton-based action detection. In ConvST-LSTM network, the basic principle of this proposed model is familiarized from the ST-LSTM approach, i.e., a sequential fusion of CNN followed by the spatiotemporal method with LSTM is also known for the extension version of RNNs. In this subsection, a brief survey is provided on RNN approaches and LSTM approaches since they are the basic building blocks of the proposed methodology. Veeriah et al. [62] worked on LSTM and introduced a differential gating method to affirm the rate of information change. Du et al. [63] work on the HRNN network (hierarchical recurrent neural network) approach for depicting skeleton structure of human body along with its temporal dynamics coordinates for keyjoints in 2D. In LSTM network, Zhu et al. [64] implemented a mixed regularization technique for normalization toward learning the co-occurrence of skeletal joint features. Meanwhile, they introduced a network for trained termed as 'in-depth-dropout mechanism.' Shahroudy et al. [32] worked in LSTM to learn long-term contextual representations of different body parts individually termed as part-aware LSTM model. Liu et al. [44, 65] intended a framework network based on 2D spatiotemporal LSTM for both temporal and spatial domains to explore the hidden input layer's information-related context in the human body. For 3D coordinates of skeletal joints, they proposed a 'trust-gate-mechanism' [44] to trade on imprecise 3D-coordinates inputted via depth sensors devices. Nowadays, skeleton-based action RNN and LSTM approaches also adapt toward action forecasting and detection [66, 67].

3 ConvST-LSTM-Net: the proposed methodology

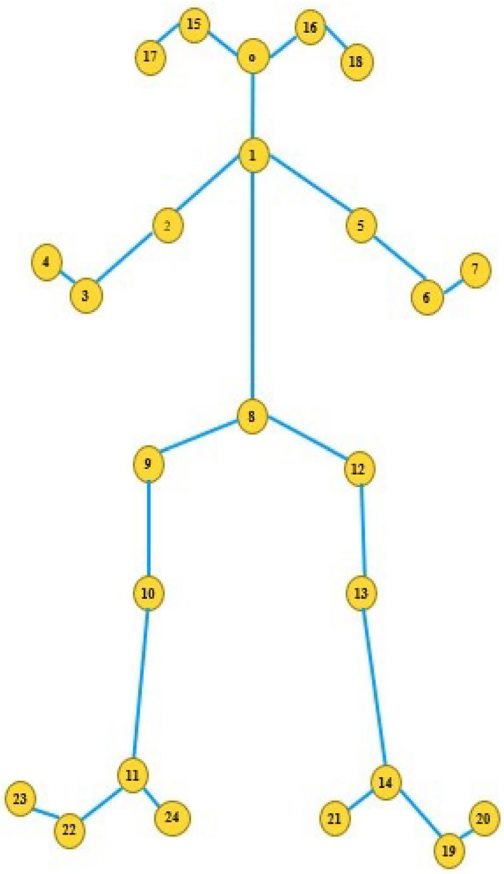
This section briefly canvasses crucial terms and approaches used in proposed ConvST-LSTM model that is divided into three models, namely CNN, ST-LSTM, and ConvST-LSTM-Net. The proposed methodology has been executed with the review of CNN along with the construction of skeleton body with coordinates and ST-LSTM, respectively.

Initially for dataset preparation, the human body frames from raw RGB videos are used to train the network and meanwhile track the 3D skeleton jointkey coordinates. For preprocessing, the 3D joint normalization method is applied that is helpful in making a bounding box above the tracked human body. Some features have been extracted for determining different activities, features such as velocity motion (v), acceleration motion (a), weight (w), depth (), height (H), angle (θ) within the consecutive skeleton joints, etc. After training, the feature extraction is done for input human activity behavior. Following subsections elucidates the whole network architecture of the ConvST-LSTM-Net model.

3.1 Keypoint detection and preprocessing

In preprocessing technique for RGB videos, the frames are inputted into the ST-LSTM network to evaluate the keyjoint locations of human body frames from skeleton coordinates. Figure 2 represents the 25-skeleton keyjoint coordinates, which have been tracked at each joint. Only 17 skeletal keyjoints are covered (since they are the informative skeletal keyjoints to specify the normal and abnormal human activities) and these are the right knee, right hip, left knee, left hip, left foot, left ankle, right foot, right ankle, head, spine mid, left wrist, spine base, right shoulder, left shoulder, right wrist, right elbow, and left elbow. Each frame tracked the human skeleton comprising X, Y, Z coordinates of human body keyjoints. After getting the 3D skeletal coordinates, normalization technique has been applied on 3D keyjoints to generate bounding boxes over the tracked human skeleton, which may vary as per the person's movement in the video.

Afterward, a feedforward network has been used based on a multi-CNN layer followed by ST-LSTM that takes input in the form of keyjoints coordinates from video frames using skeleton-based recognition. It learns the affiliation among the body parts of individuals within the frames. Table 1 presents



S.No.	Key-Joint Coordinates
0	Nose
1	Neck
2	Right Shoulder
3	Right Elbow
4	Right Wrist
5	Left Shoulder
6	Left Elbow
7	Left Wrist
8	Mid Hip
9	Right Hip
10	Right Knee
11	Right Ankle
12	Left Hip
13	Left Knee
14	Left Ankle
15	Right Eye
16	Left Eye
17	Right Ear
18	Left Ear
19	Left Thumb-Toe
20	Left Little-Toe
21	Left Heel
22	Right Thumb-toe
23	Right Little-Toe
24	Right Heel

Fig. 2 The 25-skeleton keyjoints for the human body track detection and preprocessing

Table 1 Detail of tracked skeletal keyjoint coordinates, derived features, and action class

Label	Description
Skeleton keyjoints	Right Hip, Left Hip, Left Foot, Right Foot, Right Knee, Left Knee, Right Ankle, Left Ankle, Head, Right Wrist, Left Wrist, Right Shoulder, Left Shoulder, Left Elbow, Right Elbow, Spine mid, Spine Base
Derived features	$\angle\theta$,v,α, h_d, $\not{d}$,w,H(Geometric & Kinematic features)
Action class	Sit, Stand, Walk, Run, Stand

the details of tracked skeletal keyjoints, a set of derived features, and action class. In this work, for normalizing the convergence of loss function, minimum–maximum normalization technique (i.e., min–max norm) has been used. Here, X indicates the training dataset, then normalization can be achieved as:

$$X_{\text{noramlize}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

3.2 Construction & evaluation of feature vector: geometric & kinematic features

The skeleton keyjoint coordinates are used for constructing and calculating features vectors. The keyjoints coordinates of the human body that are tracked for different activities of humans are actually decided by using feature vectors. For particular activities, different features are utilized. These features and their evaluation are as follows:

$\angle\boldsymbol{\theta}$ *(Angle between keyjoints of skeleton coordinates)*:

Among 25 keyjoints coordinates, we consider those coordinates which are connected via straight line and then a skeleton structure of tracked human body is drawn using these coordinates as shown in Fig. 3. Accordingly, only 10 keyjoint comes out to be the most relevant ones, viz. left

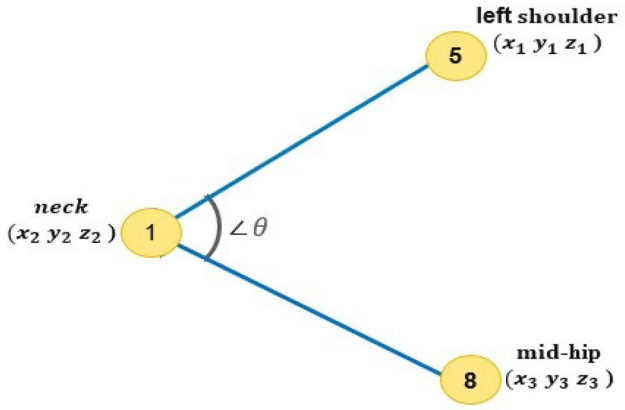


Fig. 3 Illustration for calculation of angle from the left side of the skeleton between left shoulder, neck, and mid-hip

shoulder, spine mid, right shoulder, spine base, left knee, right knee, left hip, right hip, left ankle, and right ankle are used for calculating the value of angle θ. It is the illustration for evaluating the keyjoint angles between the left shoulder, neck, and mid hip. If *A, B, C* are considered as the distance between the coordinate, then values are formulated as $A_1 = x_1 - y_1$, $B_1 = x_2 - y_2$, and $C_1 = x_3 - y_3$, then θ can be evaluated as follow:

$$\boldsymbol{\theta} = \frac{\boldsymbol{ABC}}{\boldsymbol{AB} * \boldsymbol{BC}} \tag{2}$$

where $ABC = (A1 * A2 + B1 * B2 + C1 * C2)$

$$\boldsymbol{AB} = \sqrt{\boldsymbol{A}_1^2 + \boldsymbol{B}_1^2 + \boldsymbol{C}_1^2} \text{ and } \boldsymbol{BC} = \sqrt{\boldsymbol{A}_2^2 + \boldsymbol{B}_2^2 + \boldsymbol{C}_2^2} \tag{3}$$

Velocity motion estimation ($\boldsymbol{v}$):

Velocity motion is calculated by taking the distance of positions of humans at time frames t and $t + 1$ in *x, y, z*-dimension. So, velocity of the tracked person is given by:

$$\boldsymbol{v} = \frac{d}{\boldsymbol{t}} \tag{4}$$

where $\mathscr{d}$ indicates distance of the tracked person between frames and t indicates the frame time.

acceleration motion estimation (α):

The rate of changes of velocity between consecutive frames in *x, y, z*-directions at time frame t. It is given by:

$$\alpha = \frac{v}{t} \tag{5}$$

where α is the acceleration of motion of the person. v indicates the velocity motion of tracked person between the frames. t indicates the frame time.

Head-floor distance ($\mathrm{h}_{\boldsymbol{d}}$):

It measures the distance between the head keyjoint coordinate & the floor where tracked person's location found.

Head-depth distance():

The distance measured from first camera view to the adjacent object is termed as depth. So, head-depth is calculated via the head keyjoint's coordinates in the z-dimension of a tracked person within the frame.

Width (w):

Width is defined as the difference between maximum of right-left keyjoint ($\boldsymbol{R}_{\mathbf{jMax}} - \boldsymbol{L}_{\mathbf{jMax}}$) coordinates of tracked person. The extreme left keyjoint width is estimated using a left elbow, left hip, left knee, left shoulder, left ankle, left foot, and head keyjoint values. In the same way, the right extreme keyjoints can be calculated by using all the right-side keyjoint coordinates. It is calculated as follows:

$$\boldsymbol{W} = \left|\boldsymbol{R}_{\mathbf{jMax}} - \boldsymbol{L}_{\mathbf{jMax}}\right| \tag{6}$$

where R_j indicates the right keyjoint coordinates, L_j indicates the left keyjoint coordinates.

Height (H):

Height is the measure between utmost bottom keyjoints and utmost top keyjoints of body coordinates. In extreme bottom, it includes keyjoint coordinates like left ankle, right knee, left knee, right ankle, right foot, left ankle, left foot, and right ankle and in utmost top, it includes keyjoint coordinates like head, right ankle, right elbow, left ankle, left elbow, right knee, and left knee keyjoints coordinates.

$$\boldsymbol{H} = \left|\mathbf{T}_j - \mathbf{B}_j\right| \tag{7}$$

where T_j indicates the top keyjoint coordinates, B_j indicates the bottom keyjoint coordinates.

3.3 ConvST-LSTM: the proposed model

In this section, the final preprocessed 3D keyjoint coordinates are inputted into the proposed deep learning network. We have used the sequential fusion of CNNs, Conv-LSTM & ST-LSTM to propose the ConvST-LSTM network.

3.3.1 Convolutional neural network architecture

Initially, the human action recognition has been executed by applying CNNs approach [68]. Consider, $X_t^0 = [X_1, X_2, X_3, \ldots\ldots\ldots\ldots, X_n]$ as the input vector, where n indicates the input samples and output of convolutional layers can be defined as follows:

$$\boldsymbol{C}_i^{l,j} = \boldsymbol{\sigma}\left(\boldsymbol{B}_{\mathbf{j}} + \sum_{\boldsymbol{m}=1}^{\boldsymbol{M}} \boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{j}} * \boldsymbol{X}_{\boldsymbol{i}+\boldsymbol{m}-1}^{0,\boldsymbol{j}}\right) \tag{8}$$

here l corresponds to an index of convolutional layer; σ depicts the nonlinear sigmoid-activation function; whereas B represents the bias vector corresponds to jth feature-map; filter size of CNN is indicated by M; indicates the weight metrics for the jth feature map is indicated by W_m^j; mth is the filter index.

The input frames in the proposed model consist of three input channels, namely sequences, keyjoint, and coordinates, which resemble to the x, y, and z directions, respectively. Each input frame has a resolution of 125 × 25 × 3 pixels and contains information about the movement sequences, keyjoint positions, and spatial coordinates. In convolution layer, 6 filters are passed together with configured size of kernel, padding, and SoftMax functions in the hidden layer in order to avoid the vanishing gradient problem. Max-pooling is used as a pooling operation to estimate the maximum value for feature map, and diminish the processing time by reducing the dimensionality of the frame. Then, output from the hidden layer has passed to FC dense layers. Finally, SoftMax function shows the prediction score of the action classes.

3.3.2 Spatiotemporal LSTM

Before moving to ST-LSTM, let's recap LSTM [69], which consists of 3 memory cells (gates) and escape the vanishing gradient issue. These are: (a) forget cell: a binary gate that decides how much information to pass through. (b) Input cell: decides whether the current information can be stored in the unit cell and (c) Output cell: contains sigmoid-activation gate, which decides which information to show as output. Lastly, the *tanh* layer is used to pass the cell state and further multiply it with the final output obtained from the output cell.

The equations which define the activity of each cell can be formulated as follows:

$$\boldsymbol{i_t} = \boldsymbol{\sigma}(\boldsymbol{W_{X_i}X_t} + \boldsymbol{W_{H_i}H_{t-1}} + \boldsymbol{W_{C_i}C_{t-1}} + \boldsymbol{B_i}) \tag{9}$$

$$\boldsymbol{f_t} = \boldsymbol{\sigma}(\boldsymbol{W_{X_f}X_t} + \boldsymbol{W_{\mathrm{H}_f}H_{t-1}} + \boldsymbol{W_{\mathrm{C}_f}C_{t-1}} + \boldsymbol{B_f}) \tag{10}$$

$$\boldsymbol{o_t} = \boldsymbol{\sigma}\left(\boldsymbol{W_{X_o}X_t} + \boldsymbol{W_{H_o}H_{t-1}} + \boldsymbol{W_{\mathrm{C}_f}C_t} + \boldsymbol{B_o}\right) \tag{11}$$

$$\boldsymbol{C_t} = \boldsymbol{f_t}\mathrm{C}_{t-1} + \boldsymbol{i_t}\mathbf{tanh}\left(\boldsymbol{W_{X_c}X_t} + \boldsymbol{W_{\mathrm{H_c}}H_{t-1}} + \boldsymbol{B_c}\right) \tag{12}$$

$$\boldsymbol{H_t} = \boldsymbol{o_t}\mathbf{tanh}(\boldsymbol{C_t}) \tag{13}$$

here, W_i, W_f, W_o indicates weight matrices of forget (f), input (i) and output (o) gates, respectively; $X_t \in$ input fed to LSTM cells unit at t time; σ depicts the sigmoid-activation function, whereas *tanh* depicts the hyperbolic-tangent function (both nonlinear functions); C_t indicates memory cell state within the LSTM. B_i, B_f, B_o, and B_c indicates the bias

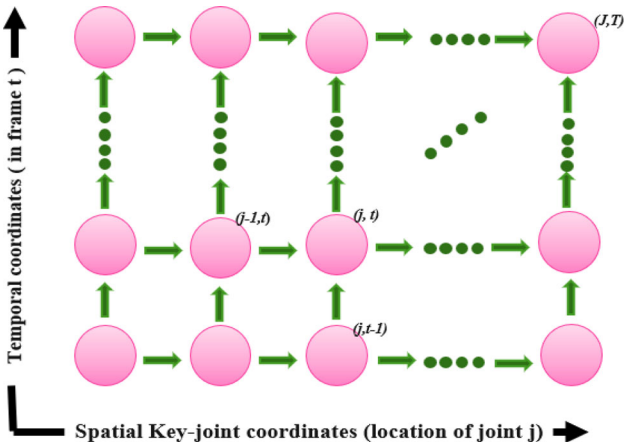


Fig. 4 Illustration of ST-LSTM cell [44]. For spatial domain, the skeletal keyjoints in each frame are aligned & feed sequentially. For temporal domain, the skeletal keyjoints are feed sequentially over the frames

vectors on forget, input & output gates, and memory cell c, respectively. Internal frame input keyjoint coordinates of each cell in the ST-LSTM model are represented in Fig. 4. The skeletal keyjoints are arranged in spatial direction and input as a chain whereas the corresponding keyjoints are inputted over various frames for temporal direction sequentially. Especially, each ST-LSTM cell is feed for a new input $(x_{j,t})$, where $x \in$ new input feed for 3D position of body keyjoint j in frame time t, the hidden layer $(h_{j,t-1})$ of the same keyjoint$j$ and the hidden layer $(h_{j-1,t})$ for the previous keyjoint *j-1* in same frame t, here j indicates the indices of keyjoint, i.e., $j \in \{1..j..J\}$ and t indicates the indices of frames, i.e.,$t \in \{1,..t..,T\}$,. An ST-LSTM unit cell consists of an input cell $(i_{j,t})$, 2-forget cells correspond to the sources of context information, i.e., temporal dimension $(f_{j,t}^T)$ & spatial domain $(f_{j,t}^S)$, in conjunction with an output gate $(o_{j,t})$.

The equations for ST-LSTM are formulated as introduced in [44]:

$$\begin{pmatrix} i_{j,t} \\ f_{j,t}^{(S)} \\ f_{j,t}^{(T)} \\ o_{j,t} \\ u_{j,t} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(W \begin{pmatrix} x_{j,t} \\ h_{j-1,t} \\ h_{j,t-1} \end{pmatrix} \right) \tag{14}$$

$$\boldsymbol{C}_{j,t} = \boldsymbol{i}_{j,t} \odot \boldsymbol{u}_{j,t} + \boldsymbol{f}_{j,t}^{(S)} \odot \boldsymbol{c}_{j-1,t} + \boldsymbol{f}_{j,t}^{(T)} \odot \boldsymbol{c}_{j,t-1} \tag{15}$$

$$\boldsymbol{h}_{j,t} = \boldsymbol{o}_{j,t} \odot \mathbf{tanh}(\boldsymbol{c}_{j,t}) \tag{16}$$

where $c_{j,t}$ indicates the cell state; $h_{j,t}$ indicates the hidden input layer in ST-LSTM unit at the spatiotemporal steps for

keyjoint j and frame time t; the modulated input frame is indicated by $u_{j,t}$; $\odot$ represents framewise product for each unit and W indicates an affine transformation within the weight.

3.3.3 ConvST-LSTM-Net architecture

Several works [53, 54] have demonstrated that each action sequence has a subset of informative keyjoints. In contrast, some keyjoints may be irrelevant in order to recognize the action classes with proper information. Therefore, for obtaining high accuracy in human action recognition the informative skeletal keyjoints have been identified while focusing on their features vector. At the same time in order to recognize human behavior, we must preferentially concentrate on the informative keyjoints (coordinates for feed), ignoring the features of the irrelevant keyjoints.

This model has been executed by taking a sequential fusion of CNN, ST-LSTM (combination of LSTM and spatiotemporal based recognition), and FC layers. Here, CNNs are pre-owned for feature extraction, ST-LSTMs are used in sequence prediction for spatiotemporal feature extraction, and the features dense layers are used for mapping. For classification, the outputs from CNN's hidden layer are fed to the ST-LSTM layers and then GAP (Global Pooling Layer) is used to flatten the data followed by FC layers within the model.

The transformation equations for ConvST-LSTM-Net can be given as:

$$\mathcal{F}_{j,t}^{(T)} = \sigma\left(W_{X_{\mathcal{F}}} X_{j,t} + W_{H_{\mathcal{F}}} H_{j,t-1} + B_{\mathcal{F}}\right) \tag{17}$$

$$\mathcal{F}_{j,t}^{(S)} = \sigma\left(W_{X_{\mathcal{F}}} X_{j,t} + W_{H_{\mathcal{F}}} H_{j,t-1} + B_{\mathcal{F}}\right) \tag{18}$$

$$\tilde{\mathrm{I}}_{j,t} = \sigma\left(W_{X_{\tilde{\mathrm{I}}}} Xj_{,\,t} + W_{\mathrm{H}_{\tilde{\mathrm{I}}}} H_{j,t-1} + B_{\tilde{\mathrm{I}}}\right) \tag{19}$$

$$\tilde{\mathrm{O}}_{j,t} = \sigma\left(W_{X\ O} X_t + W_{H\ O} H_{t-1} + B_o\right) \tag{20}$$

$$C_{j,t} = f_{j,t} C_{t-1} + i_{j,t} \tanh(W_{X_c} X_t + W_{H_c} H_{j,t-1} + B_c \tag{21}$$

$$u_{j,t} = \tanh(W_{X_u} * X_t + W_{H_u} H_{t-1} + B_u \tag{22}$$

$$H_t = o_t \odot \tanh(C_t) \tag{23}$$

where $X_{j,t}$, $Cj,_t$, $H_{j,t}$, $F_{j,t}$, $I_{j,t}$ indicates inputs states, cells states, hidden states, forget cells, input cells for keyjoint j in frame time t; u_t input modulation gates and $\tilde{O}_t$ is the output cells; C_t is the memory cell used for aggregating the states information controlled by the cells. Figure 5 depicts about the ST-LSTM layer for each unit cell.

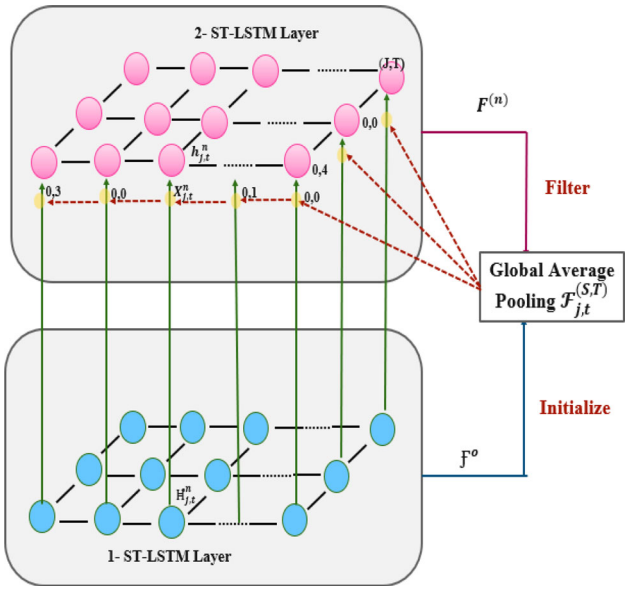


Fig. 5 The ConvST-LSTM network for ST-LSTM layer in each unit cell

Model training ConvST-LSTM-Net was trained on the frame samples obtained from the videos, keypoint recognition, followed by the fusion of spatiotemporal model consisting of ConvLSTMs. The model was trained on 150 epochs on a machine having AMDRyzen7 5800H processor, 8 GB RAM, and Graphics: NVIDIA GeForce RTX 3050 M, GPU, having a learning rate of 0.001 after repeated hyperparameter tuning.

For setup, Keras API version 2.3 of Python along with TensorFlow version 2.3.0 has been used in the backend to build the spatiotemporal model. To increase the code's reusability and readability, some helper functions are initially defined from the python libraries. Along with an optimum value has been set for the user-defined hyperparameters like size, no. of layers, iteration, epochs, no. of batch sizes, and learning rate. The training sample data with various batch size is feed to the model and get trained over 150 epochs. In first time-distributed CNN layer, we use 32 filters with kernel size 3 and its output is then regularized to attain faster convergence. Then, max-pooling is added to reduce computational costs. Dropout layer benefited to avoid the overfitting where 50% of weights are dropped randomly. For next time-distributed CNN layers, different size of filters is practices after performing feature extraction, we apply an additional dropout layer with a rate of 40%. At step 3, we use GAP layer through which the output of CNN layer is flattened to 1*56 dimension.

Further, ST-LSTM is used to handle the sequential action data of the tracked person's keyjoints coordinates. The ST-LSTM layer's output is passed to the time-distributed FC dense layer. At last, SoftMax layer gives the framewise probabilities for each action classes. The architecture of the

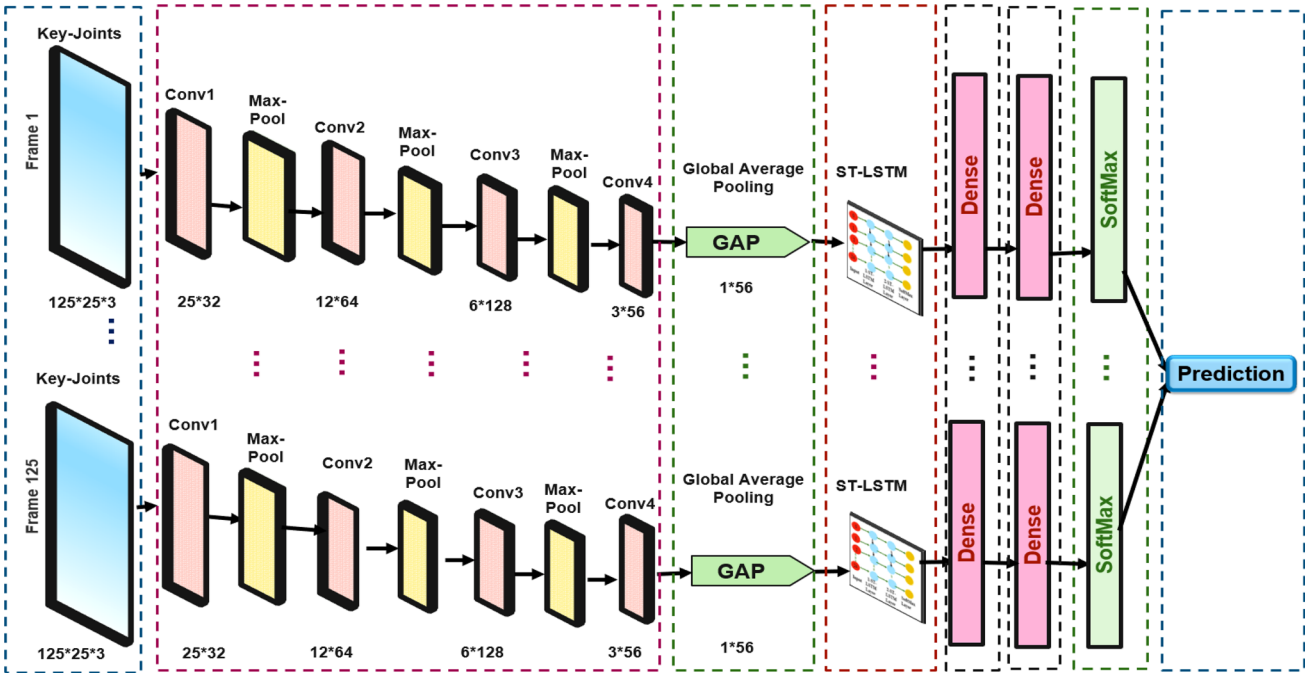


Fig. 6 Block architectural diagram of ConvST-LSTM-Net model for human action recognition. Starting from left side, the input frames clipped from videos; time-distributed convolutional layers including max-pooling, GAP, ST-LSTMs, FC dense layer followed by SoftMax function layer that results as a prediction of action

proposed ConvST-LSTM model is illustrated in Fig. 6. Further, Adam optimizer is help in optimizing the cost function and uses gradient clipping within the code. The hyperparameters such as checkpoint-path, saver-function, epochs, iterations, filter size, kernel size, and test & train data have been set for training purpose. Moreover, the proposed model utilized the stopping criteria with the value of 50. That means the training will be terminated only if there is no improvement in the monitor performance measure for 50 epochs or iterations in a row. This helps to prevent the entire model fall in to local optima. It has been found that the performance is excessively upgraded by using a sequential way.

4 Experimental results and analysis

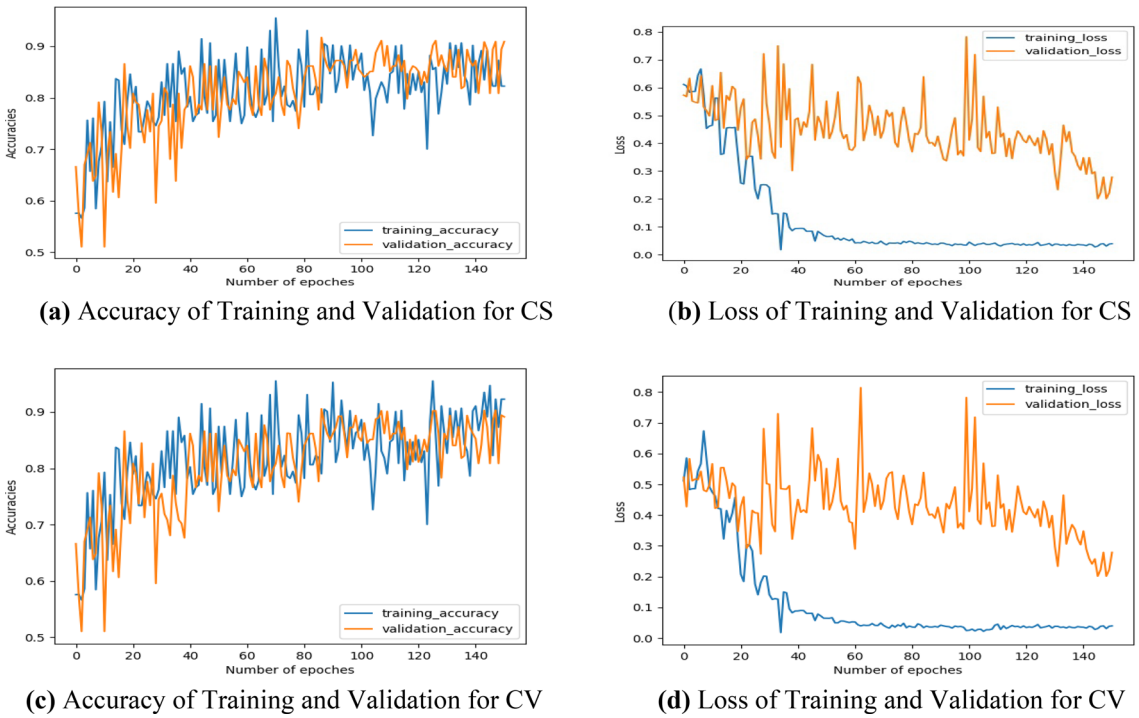
This section discusses about the implementation trait of the proposed model on various benchmarks with its training hyperparameters. We has evaluate the performance of ConvST-LSTM-Net model on three publicly available benchmarks, i.e., the NTU RGB + D 60 dataset [32], UT-Kinect dataset [33], UP-Fall-Detection Dataset [34], UCF101 [35], and HDMB51 dataset [36] for skeleton-based data.

4.1 Experiments on NTU RGB + D 60 dataset

The NTU RGB + D60 [32] is publicly available dataset used for human action recognition consisting of total 56,880 samples having 60 activity classes collected over 40 subjects in it. In this dataset, activities are classified into three categories having 40 daily living activities (drinking, standing, reading, happing, etc.), 9 medical conditions-related activities (sneezing, staggering, falling, vomiting, etc.), and 11 common activities (punching, kicking, hugging, etc.) based on multimodal information of the daily action characterization, along with 3D skeletal keyjoint, RGB-videos, masked-depth maps, full-depth maps, and infrared sequences data. The annotations provide the 3D location in x, y, z-dimension of each keyjoint in the camera coordinate system. It has total 25 key points per subject and each clip has 2 subjects. The evaluation has done on two protocols: Cross-Subject (CS) and Cross-View (CV).

For performing experiments, we choose 5 action classes (Stand, Sit, Run, Walk, Fall) contains 150 clips in each class. The two benchmarks for evaluation are set as: (1) Cross-subject (CS) contains 400 clips from 5 subjects, used for training; and the 100 clips for validation. (2) Cross-view (CV) contains 450 from 5 subjects used for training and 150 clips for validation. The proposed ConvST-LSTM-Net model surpasses the ConvLSTM network in [68] by 4.3% with the CS evaluation protocol and 3.1% with the CV evaluation protocol. This demonstrates that spatiotemporal skeleton-based recognition approaches in LSTM networks bring significant improvement. The comparative analysis for the results of the proposed ConvST-LSTM-Net model with state-of-the-art approaches is enumerated in Table 2.

The trade-off curves for training accuracy & loss and validation accuracy & loss on the benchmark of NTU RGB +



(a) Accuracy of Training and Validation for CS

(b) Loss of Training and Validation for CS

(c) Accuracy of Training and Validation for CV

(d) Loss of Training and Validation for CV

Fig. 7 Trade-off curves for model's Training and Validation Accuracy versus Training and Validation Loss on the NTU RGB + D 60 benchmark dataset

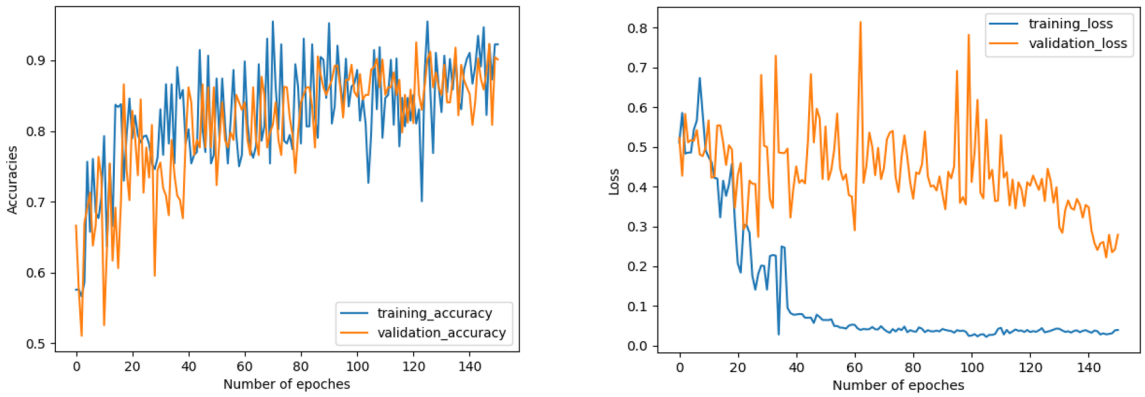
Table 2 Experimental Results on NTU RGB + D 60 Dataset for skeletal sequence data

Methods	CS (%)	CV (%)
Deep-LSTM [70]	56.3	64.1
ST-LSTM [44]	69.2	77.7
ST-LSTM + Global(1) [71]	70.5	79.5
ST-LSTM + Global(2) [71]	70.7	79.4
Conv-LSTM [68]	76.2	83.2
Conv-GRU [72]	88.9	90.1
LA-GCN [73]	90.9	89.28
TD-GCN [74]	91.82	94.2
SkeletonGCL [75]	89.2	90.3
ConvST-LSTM-Net	91.72	90.5

D 60 dataset for its two-evaluation protocol, i.e., CS and CV is illustrated in Fig. 7. Training and validation accuracy increases with time as depicted in Fig. 7a, c and finally, the growth rate reaches a steady-state value. Figure 7b and d depict the loss curve, illustrating the gradual decrease of validation loss with increasing epochs. To evaluate the model's performance, the weights are saved from the epochs that achieve the highest validation accuracy. The loss curve is shown in Fig. 7b and d, which demonstrates how the validation loss gradually decreases by increasing epoch. For testing, the weights from the epochs with the maximum validation accuracy are saved.

4.2 Experiments on the UT-Kinect dataset

The UT-Kinect dataset [34] is publicly available and was taken through a single stationary Kinect comprised of total 10 subjects that took total 10 action types (walking, stand up, pick up, carry, sit down, throw, push, pull, wave hands, clap hands). Each subject performs each action twice. 3-channels were captured for (i) RGB, (ii) depth, and (iii) skeleton keyjoint locations. We have only recorded the frames when the skeleton of human body was tracked. To assess the proposed method on this dataset, the standard leave-one-out-cross-validation protocol has been followed. Table 3 provides the comparative result of the proposed ConvST-LSTM-Net model with state-of-the-art approaches. The trade-off curves for the model accuracy and loss on the UT-Kinect dataset has been illustrated in Fig. 8. It is observed from these curves that the proposed methodology offers exceptional accuracy during training and moderate accuracy in the validation process. For training process, the model causes low and for validation process it causes moderate loss.



(a) Accuracy of Training and Validation

(b) Loss of Training and Validation

Fig. 8 Trade-off curves for Model Training and Validation Accuracy & Model Training and Validation Loss on the UT-Kinect dataset

Table 3 Experimental Results on UT-Kinect

Method	Accuracy (%)
Histogram of 3d joints [33]	88.9
ST-LSTM [44]	87.0
ST-LSTM + Global(1) [71]	91.9
ST-LSTM + Global(2) [71]	90.8
Conv-LSTM [68]	90.2
Conv-GRU [72]	89.99
ConvST-LSTM-NET	92.0

Table 4 Experimental Results on UP-Fall Detection Dataset

Method	Accuracy (%)
GCA-LSTM [71]	88.5
Conv-LSTM [68]	87.6
Conv-GRU [72]	88.8
ConvST-LSTM	89.0

Table 5 Experimental Results of ConvST-LSTM on the UCF101 Dataset

Method	Accuracy (%)
GCA-LSTM [71]	84.2
Conv-LSTM [68]	83.3
Conv-GRU [72]	86.28
PYSKL [76]	88.89
ConvST-LSTM	92.8

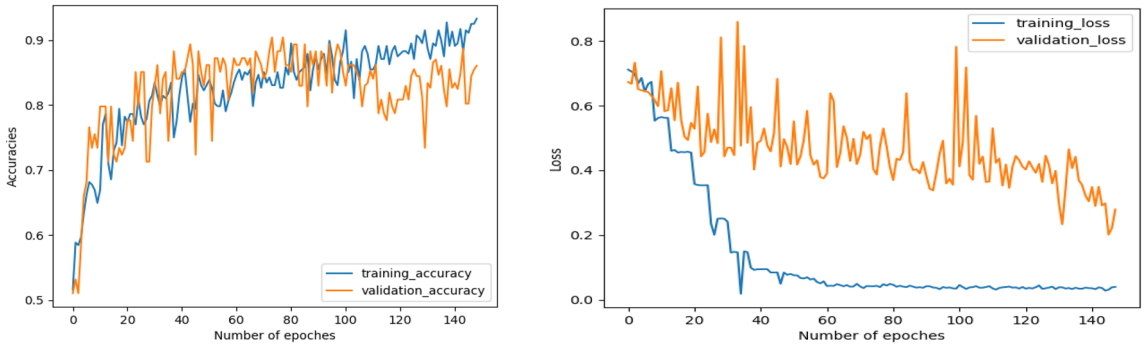
4.3 Experiments on the UP-fall detection dataset

The UP-fall detection [34] is the large-scale multimodal dataset collected by using vision-wearable, and ambient sensors. It includes Activity for Daily Livings (ADLs-850 GB), collected by 17 healthy persons including 9 male, 8 females individuals. It has total 11 actions, i.e., 6 basic actions for daily living: walk, sit, stand, picking-up an item, laying, jump and 5 fall-actions: fall-forward via knees, fall-forward via hands, fall-sitting in an empty chair, fall backward and fall-sideward). Two cameras were set up to capture the subject's front views as well as its side views. A total of 589,418 sample image frames are there taken from both cameras. Total size of this vision dataset was 277 GB. For performing experiments, we choose 5 action classes (i.e., Stand, Sit, Run, Walk, Fall) contains 1000 clips in each class in which 800 clips used for training, and the 200 clips for validation. Table 4 gives the comparative results of the proposed ConvST-LSTM-Net with various state-of-the-art methods.

Figure 9 illustrates the trade-off curves for (a) accuracy of training and validation vs. (b) loss of training and validation. It is observed from these curves that the proposed methodology offers exceptional accuracy during training and moderate accuracy in the validation process. For training process, the model causes low and for validation process it causes moderate loss.

4.4 Experiments on the UCF101 dataset

The UCF101 [35] is a popular action recognition dataset that contains 13,320 video clips from 101 action categories. The action videos are clustered in 25 groups, where each group contains 4–7 videos of an action. The action categories can be classified into five distinct types, i.e., (a) Human-Object Interaction (b) Body-Motion (c) Human–Human Interaction (d) Playing Musical Instruments (e) Sports. For performing experiments, we choose 5 action classes from body motion categories contains total 17 body motion clips. Table 5 gives



(a) Accuracy of Training and Validation

(b) Loss of Training and Validation

Fig. 9 Trade-off curves for Model Training and Validation Accuracy versus Model Training and Validation Loss on UP-Fall Detection dataset

Table 6 Experimental Results on the HMDB51 Dataset

Method	Accuracy (%)
GCA-LSTM [71]	82.3
Conv-LSTM [68]	81.2
Conv-GRU [72]	80.8
PYSKL [76]	69.4
ConvST-LSTM	91.86

the comparative results of the proposed ConvST-LSTM-Net with various state-of-the-art methods.

The trade-off curves for the model accuracy and loss on the UCF101 dataset has been illustrated in Fig. 10, i.e., (a) accuracy of training and validation versus (b) loss of training and validation. For training process, the model causes low and for validation process it causes moderate loss.

4.5 Experiments on the HMDB51 dataset

The HMDB51 [36] dataset is a commonly used benchmark dataset for action recognition in videos, which consists of video clips from various sources like movies, YouTube. From 2 GB, total 7000 clips distributed in 51 action classes. The actions categories can be divided into five types: (a) General facial actions. (b) Facial actions with object manipulation. (c) General body movements. (d) Body movements with object interaction. (e) Body movements for human interaction. The video clips have varying durations and resolutions. For performing experiments, we select the general body movements action classes in which 5 action clips are taken (i.e., Stand up, Sit down, Run, Walk, Fall). Each action classes contains minimum of 101 clips. Among them 80% are used of training and 20% are used for validation. Table 6 gives the comparative results of the proposed ConvST-LSTM-Net with various methods.

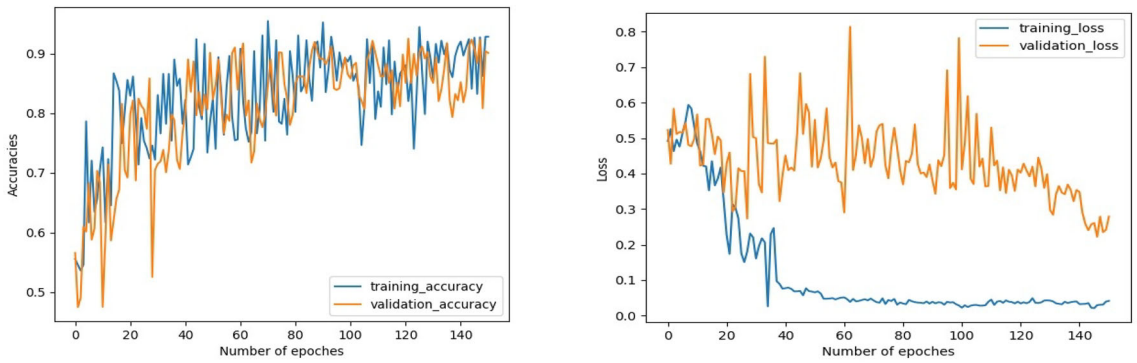
Figure 11 illustrates the trade-off curves for (a) accuracy of training and validation versus (b) loss of training and validation on HDMB51 dataset. Form this, it is observed that the proposed methodology achieves outstanding accuracy during the training process and moderate accuracy during the validation process. The model exhibits low loss during training and moderate loss during validation.

4.6 Multimodal analysis over standard performance measures

This section discusses the results analysis gained on the proposed ConvST-LSTM-Net. The performance of the model has been measured on different performance metrics, i.e., Precision, Recall, F1-score, and Accuracy. Figure 12 displays the accuracy, precision, recall, and F1-score on various benchmarks. The accuracies and losses are plotted for 150 epochs. The proposed ConvST-LSTM-Net results in a better accuracy. The effectiveness of the proposed model is verified on various benchmarks, i.e., NTU RGB + D 60, UT-Kinect, UP-Fall Detection, UCF101, and HDMB51 datasets, where the model outperforms state-of-the-art methods. Figure 13 illustrates the human action recognition results obtained in different benchmarks datasets with framing the bounding box over the tracked human. We observed that the performance of the model is sufficiently high.

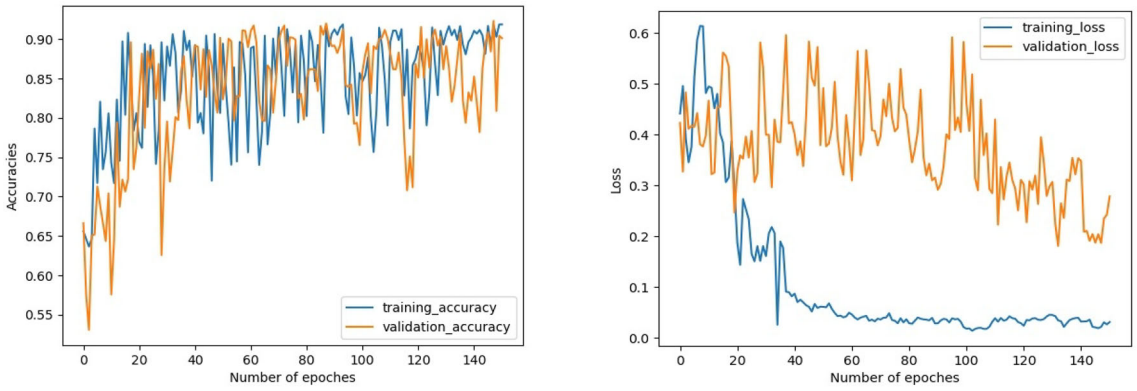
5 Conclusions and future prospective

Human action recognition has gained a large prominence in today's era, but few limitations are there in their application areas despite having networks that could achieve good results. In this paper, we improved the internal cell structure of the ST-LSTM unit and successfully proposed a ConvST-LSTM having high accuracy & reliability. The model is based on a spatiotemporal LSTM module, uses video frames



(a) Accuracy of Training and Validation **(b)** Loss of Training and Validation

Fig. 10 Trade-off curves for Model Training and Validation Accuracy versus Model Training and Validation Loss on UCF101 Dataset



(a) Accuracy of Training and Validation **(b)** Loss of Training and Validation

Fig. 11 Trade-off curves for Model Training and Validation Accuracy versus Model Training and Validation Loss on HDMB51 Dataset

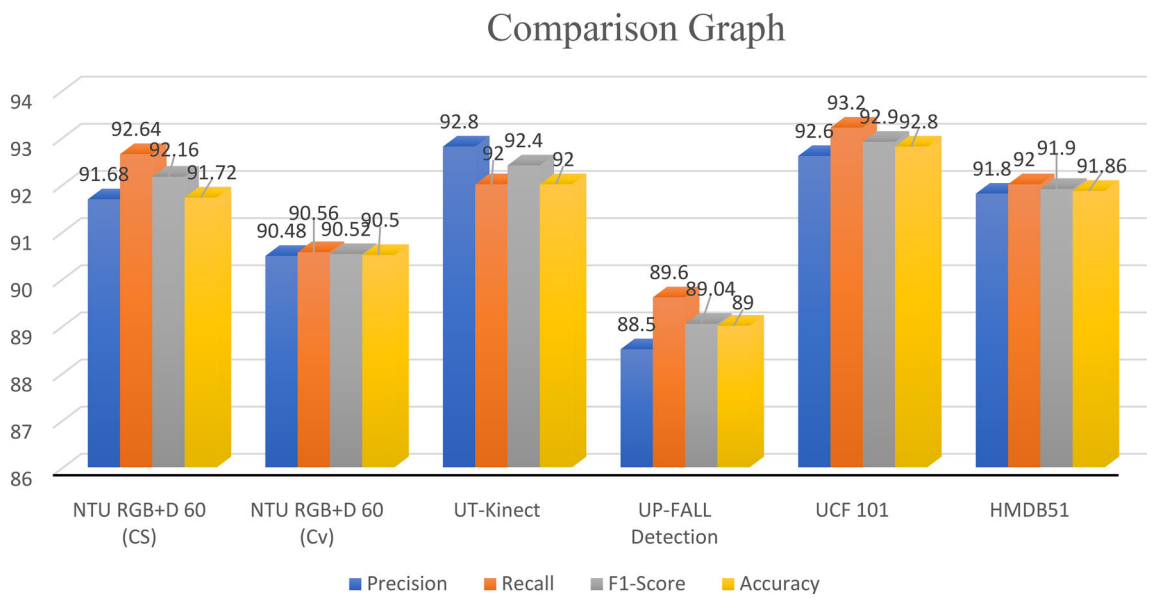


Fig. 12 Comparative Stats of Standard Performance Measure over different datasets

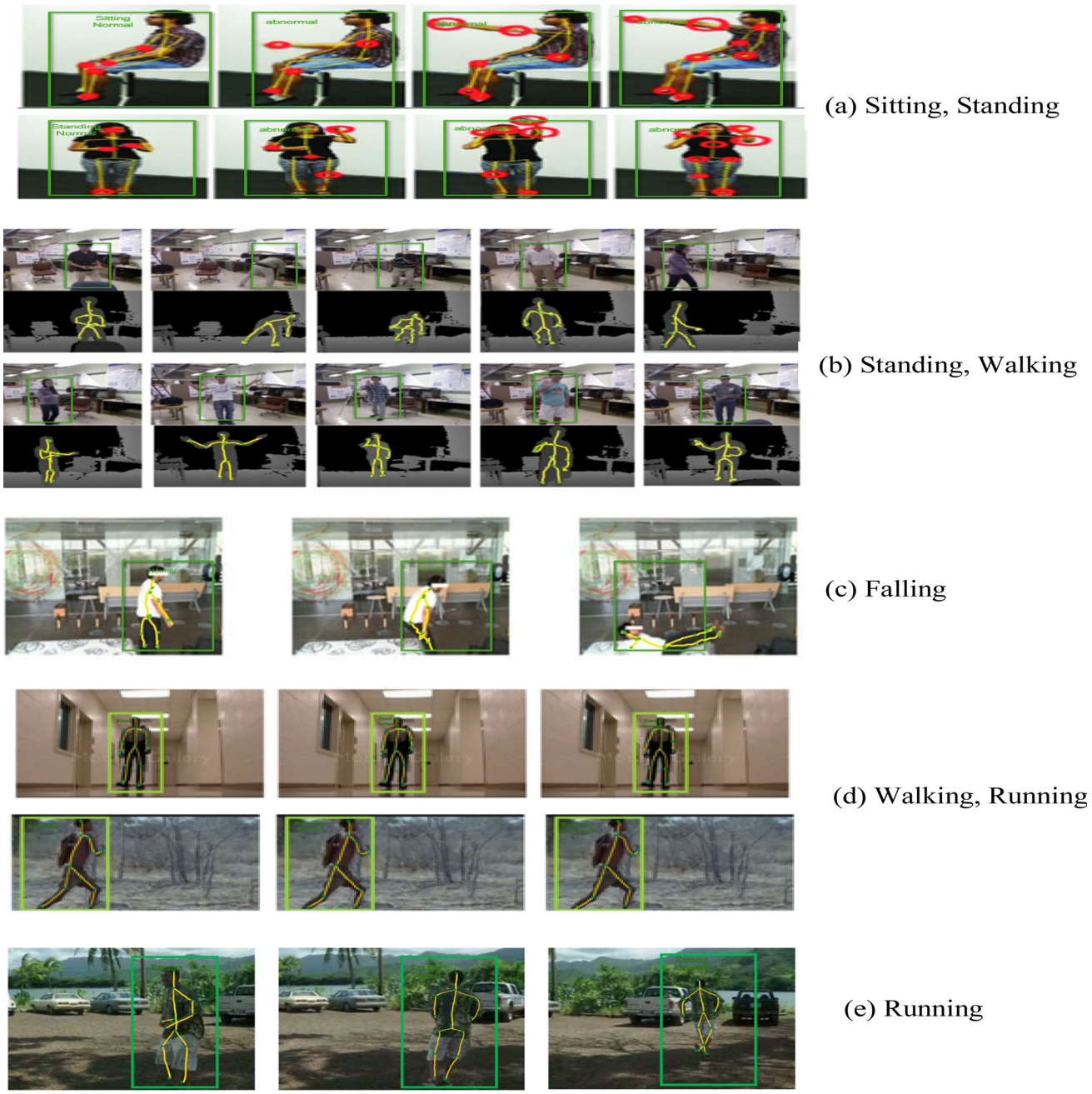


Fig. 13 Illustration of the Human Action Recognition on various benchmarks. Starting from left–right **a** NTU RGB + D 60 Dataset: Sitting, Standing **b** UT-Kinect Dataset: Standing, Walking **c** UP-Fall Detection Dataset: Fall **d** UCF101 Dataset: Walking, Running **e** HMDB51 Dataset: Running

and skeleton-based features and has the robust capability for selecting the informative keyjoints in each frame while ignoring the irrelevant keyjoints of the skeleton sequence. The model is independent of the camera orientation, clothing, background noise, etc., which can effectively recognize suspicious actions related to human activity. Finally, the experimental results show better performance and achieve good accuracy for skeleton-based anomaly activity recognition. However, the consequences of growing population and rise in ever-challenging activities fosters the need to introduce a more promising predictive methodology for recognizing human behavior that proffers a practical alternative solution for the security and protection of people from daily risks in life. With the future perspective, we can use a graph oriented spatiotemporal base data to represent humans and

objects. Moreover, GCN can also be used for the classification and detection of unsuspicious activity of human behaviors.

Acknowledgements Not applicable.

Author contributions Roshni Singh, carried out the related studies, participated in the sequence alignment and drafted the manuscript along with performances and statistical analysis. Dr. Abhilasha Sharma, conceived of the study and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Funding Not applicable.

Availability of data and materials Not applicable.

Declarations

Conflict of interest The authors have no relevant financial or nonfinancial interest to disclose. The authors have no competing interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or nonfinancial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interest in any material discussed in this article.

References

1. Chang Y, Tu Z, Xie W, and Yuan J (2020). Clustering driven deep autoencoder for video anomaly detection. In European Conference on Computer Vision (pp. 329–345). Springer, Cham.
2. Zhang D, He L, Tu Z, Zhang S, Han F, Yang B (2020) Learning motion representation for real-time spatio-temporal action localization. Pattern Recogn 103:107312
3. Niu W, Long J, Han D and Wang Y-F, Human activity detection and recognition for video surveillance, in 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), vol. 1, June 2004, pp. 719–722 Vol.1.
4. Valera M, Velastin SA (2005) Intelligent distributed surveillance systems: a review. IEE Proc–Vision, Image Signal Process 152(2):192–204
5. Lin W, Sun MT, Poovandran R, and Zhang Z (2008), Human activity recognition for video surveillance, in 2008 IEEE International Symposium on Circuits and Systems, pp. 2737–2740.
6. Kalimuthu S, Perumal T, Yaakob R, Marlisah E, and Babangida L (2021), Human activity recognition based on smart home environment and their applications, challenges. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 815–819). IEEE.
7. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
8. Chaquet JM, Carmona EJ, Fernández-Caballero A (2013) A survey of video datasets for human action and activity recognition. Comput Vis Image Underst 117(6):633–659
9. Patrona F, Chatzitofis A, Zarpalas D, Daras P (2018) Motion analysis: Action detection, recognition, and evaluation based on motion capture data. Pattern Recogn 76:612–622
10. Vishwakarma DK, Dhiman A, Maheshwari R, Kapoor R (2015) Human motion analysis by fusion of silhouette orientation and shape features. Procedia Comput Sci 57:438–447
11. Yao H, Hu X (2023) A survey of video violence detection. Cyber-Phys Syst 9(1):1–24
12. Yang Y, Liu G, Gao X (2022) Motion guided attention learning for self-supervised 3D human action recognition. IEEE Trans Circuits Syst Video Technol 32(12):8623–8634
13. Duan H, Wang J, Chen K, and Lin D (2022), DG-STGCN: dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv preprint arXiv:2210.05895.
14. Liu S, Bai X, Fang M, Li L, Hung CC (2022) Mixed graph convolution and residual transformation network for skeleton-based action recognition. Appl Intell 52(2):1544–1555
15. Abdulhussein AA, Hassen OA, Gupta C, Virmani D, Nair A, and Rani P (2022), Health monitoring catalogue based on human activity classification using machine learning. Int J Electrical Comput Eng, 12(4): (2088–8708).
16. Andrade-Ambriz YA, Ledesma S, Ibarra-Manzano MA, Oros-Flores MI, Almanza-Ojeda DL (2022) Human activity recognition using temporal convolutional neural network architecture. Expert Syst Appl 191:116287
17. Qiu S, Zhao H, Jiang N, Wang Z, Liu L, An Y, Fortino G (2022) Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. Inform Fusion 80:241–265
18. Wu L, Zhang C, Zou Y (2023) SpatioTemporal focus for skeleton-based action recognition. Pattern Recogn 136:109231
19. Mahdikhanlou K, Ebrahimnezhad H (2023) 3D hand pose estimation from a single RGB image by weighting the occlusion and classification. Pattern Recogn 136:109217
20. Dallel M, Havard V, Dupuis Y, and Baudry D (2022), A sliding window based approach with majority voting for online human action recognition using spatial temporal graph convolutional neural networks. In 2022 7th International Conference on Machine Learning Technologies (ICMLT) (pp. 155–163).
21. Sánchez-Caballero A, Fuentes-Jiménez D, and Losada-Gutiérrez C (2022) Real-time human action recognition using raw depth video-based recurrent neural networks. Multimedia Tools Appl, 1–23.
22. Yue R, Tian Z, and Du S (2022) Action recognition based on RGB and skeleton data sets: A survey. Neurocomputing.
23. Khaire P, Kumar P (2022) Deep learning and RGB-D based human action, human–human and human–object interaction recognition: a survey. J Vis Commun Image Represent 86:103531
24. Ding C, Wen S, Ding W, Liu K, Belyaev E (2022) Temporal segment graph convolutional networks for skeleton-based action recognition. Eng Appl Artif Intell 110:104675
25. Setiawan F, Yahya BN, Chun SJ, Lee SL (2022) Sequential inter-hop graph convolution neural network (SIhGCN) for skeleton-based human action recognition. Expert Syst Appl 195:116566
26. Khowaja SA, & Lee SL (2022) Skeleton-based human action recognition with sequential convolutional-LSTM networks and fusion strategies. Journal of Ambient Intelligence and Humanized Computing, 1–18.
27. Hou R, Wang Z, Ren R, Cao Y, and Wang Z (2022). Multi-channel network: constructing efficient GCN baselines for skeleton-based action recognition. Comput Gr.
28. Gao BK, Dong L, Bi HB, Bi YZ (2022) Focus on temporal graph convolutional networks with unified attention for skeleton-based action recognition. Appl Intell 52(5):5608–5616
29. Xu W, Wu M, Zhu J, Zhao M (2021) Multi-scale skeleton adaptive weighted GCN for skeleton-based human action recognition in IoT. Appl Soft Comput 104:107236
30. Song YF, Zhang Z, Shan C, and Wang L (2020), Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In proceedings of the 28th ACM international conference on multimedia (pp. 1625–1633).
31. Wang L, Suter D (2007) Learning and matching of dynamic shape manifolds for human action recognition. IEEE Trans Image Process 16(6):1646–1661

32. Shahroudy A, Liu J, Ng T-T, and Wang G (2016), Ntu rgb+d: a large scale dataset for 3d human activity analysis, in CVPR, 2016.
33. Xia L, Chen C-C, and Aggarwal JK, View invariant human action recognition using histograms of 3D joints, in Proc. CVPR, 2012, pp. 20–27 (2012)
34. Martínez-Villaseñor L, Ponce H, Brieva J, Moya-Albor E, Núñez-Martínez J, Peñafort-Asturiano C (2019) UP-fall detection dataset: a multimodal approach. Sensors 19(9):1988
35. Soomro K, Zamir AR, and Shah M (2012) UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402
36. Kuehne HH, Jhuang E, Garrote T, Poggio and Serre T (2011) HMDB: a large video database for human motion recognition, 2011 International Conference on Computer Vision, 2011, 2556–2563, https://doi.org/10.1109/ICCV.2011.6126543.
37. Vemulapalli R, Arrate F, and Chellappa R (2014), Human action recognition by representing 3d skeletons as points in a lie group, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 588–595.
38. Vemulapalli R and Chellapa R, Rolling rotations for recognizing human actions from 3d skeletal data, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4471– 4479.
39. Ke Q, Bennamoun M, An S, Sohel F, and Boussaid F (2017), 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.
40. Li, B, Dai, Y, Cheng X, Chen H, Lin Y, and He M (2017), Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn, in 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2017, pp. 601–604.
41. Li C, Zhong Q, Xie D and Pu S (2017), Skeleton-based action recognition with convolutional neural networks, in 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2017, pp. 597–600.
42. Liu M, Liu H, Chen C (2017) Enhanced skeleton visualization for view invariant human action recognition. Pattern Recogn 68:346–362
43. Zhu K, Wang R, Zhao Q, Cheng J, Tao D (2019) A cuboid cnn model with an attention mechanism for skeleton-based action recognition. IEEE Trans Multimedia 22(11):2977–2989
44. Liu J, Shahroudy A, Xu D, and Wang G (2016), Spatio-temporal lstm with trust gates for 3d human action recognition," in European Conference on Computer Vision. Springer, 2016, pp. 816–833.
45. Cao C, Lan C, Zhang Y, Zeng W, Lu H, Zhang Y (2018) Skeleton-based action recognition with gated convolutional neural networks. IEEE Trans Circuits Syst Video Technol 29(11):3247–3257
46. Zhao R, Wang K, Su H, and Ji Q (2019), Bayesian graph convolution lstm for skeleton based action recognition, in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6882–6892.
47. Song S, Lan C, Xing J, Zeng W, and Liu J (2017), An end-to-end spatiotemporal attention model for human action recognition from skeleton data, in Thirty-first AAAI Conference on Artificial Intelligence, 2017.
48. Zhang S, Yang Y, Xiao J, Liu X, Yang Y, Xie D, Zhuang Y (2018) Fusing geometric features for skeleton-based action recognition using multilayer lstm networks. IEEE Trans Multimedia 20(9):2330–2343
49. Fan Z, Zhao X, Lin T, Su H (2018) Attention-based multiview reobservation fusion network for skeletal action recognition. IEEE Trans Multimedia 21(2):363–374
50. Xie J, Miao Q, Liu R, Xin W, Tang L, Zhong S, Gao X (2021) Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition. Neurocomputing 440:230–239
51. Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N (2019) View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Trans Pattern Anal Mach Intell 41(8):1963–1978
52. Yan S, Xiong Y, and Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In Thirty-second AAAI conference on artificial intelligence.
53. Song YF, Zhang Z, Shan C, and Wang L (2020). Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition. In proceedings of the 28th ACM international conference on multimedia (pp. 1625–1633).
54. Song YF, Zhang Z, Shan C, Wang L (2022) Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans Pattern Anal Mach Intell 45(2):1474–1488
55. Shi L, Zhang Y, Cheng J, Lu H (2020) Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Trans Image Process 29:9532–9545
56. Cheng K, Zhang Y, He X, Chen W, Cheng J, & Lu H (2020). Skeleton-based action recognition with shift graph convolutional network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 183–192).
57. Ye, L., & Ye, S. (2021, April). Deep learning for skeleton-based action recognition. In Journal of Physics: Conference Series (Vol. 1883, No. 1, p. 012174). IOP Publishing.
58. Zhang J, Ye G, Tu Z, Qin Y, Zhang J, Liu X, and Luo S, A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition, CAAI Transactions on Intelligence Technology, (2020).
59. Shi L, Zhang Y, Cheng J, and Lu H (2019), Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 026–12 035
60. Shi L, Zhang Y, Cheng J, and Lu H (2019), Skeleton-based action recognition with directed graph neural networks, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7912–7921.
61. Zhang X, Xu C, Tian X, Tao D (2019) Graph edge convolutional neural networks for skeleton-based action recognition. IEEE Transact Neural Netw Learn Syst 31(8):3047–3060
62. Veeriah V, Zhuang N, and Qi G-J (2015), Differential recurrent neural networks for action recognition, in ICCV, 2015.
63. Du Y, Wang W and Wang L (2015), Hierarchical recurrent neural network for skeleton based action recognition, in CVPR, 2015.
64. Zhu W, Lan, C, Xing J, Zeng W, Li Y, Shen L, and Xie X (2016), Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks, in AAAI, 2016.
65. Liu J, Shahroudy A, Xu D, Kot AC, and Wang G (2017), Skeleton-based action recognition using spatio-temporal lstm network with trust gates, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
66. Jain A, Zamir AR, Savarese S, and Saxena A (2016), Structural-rnn: deep learning on spatio-temporal graphs, in CVPR, 2016.
67. Li Y, Lan C, Xing J, Zeng W, Yuan C, and Liu J (2016) Online human action detection using joint classification-regression recurrent neural networks, in ECCV, 2016.
68. Yadav SK, Tiwari K, Pandey HM, & Akbar SA (2022), Skeleton-based human activity recognition using ConvLSTM and guided feature learning. Soft Comput, 1–14.
69. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166
70. Hu JF, Zheng W-S, Lai, J and Zhang J (2015), "Jointly learning heterogeneous features for RGB-D activity recognition," in Proc. CVPR, 2015, pp. 5344–5352.

71. Liu J, Wang G, Duan LY, Abdiyeva K, Kot AC (2017) Skeleton-based human action recognition with global context-aware attention LSTM networks. IEEE Trans Image Process 27(4):1586–1599
72. Yadav SK, Luthra A, Tiwari K, Pandey HM, Akbar SA (2022) ARFDNet: an efficient activity recognition & fall detection system using latent feature pooling. Knowl-Based Syst 239:107948
73. Xu, H, Gao Y, Hui Z, Li J, and Gao X (2023), Language knowledge-assisted representation learning for skeleton-based action recognition. arXiv preprint arXiv:2305.12398.
74. Liu J, Wang X, Wang C, Gao Y, and Liu M (2023) Temporal decoupling graph convolutional network for skeleton-based gesture recognition. IEEE Transactions on Multimedia
75. Huang, X., Zhou H, Feng B, Wang X, Liu W, Wang J, Feng H, Han J, Ding E, and Wang J (2023) Graph contrastive learning for skeleton-based action recognition. arXiv preprint arXiv:2301.10900 (2023).
76. Duan, H, Wang J, Chen K, and Lin D (2022) Pyskl: towards good practices for skeleton action recognition. In Proceedings of the 30th ACM International Conference on Multimedia, pp. 7351–7354. 2022.
77. Duan, H, Zhao Y, Chen K, Lin D, and Dai B (2022) Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2969–2978.

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.

Clean Energy, 2023, Vol. 7, No. 5, 1110–1124
https://doi.org/10.1093/ce/zkad051
Advance Access Publication 24 October 2023
Research Article

Development and assessment of a novel integrated system powered by parabolic trough collectors for combined power, heating and freshwater production

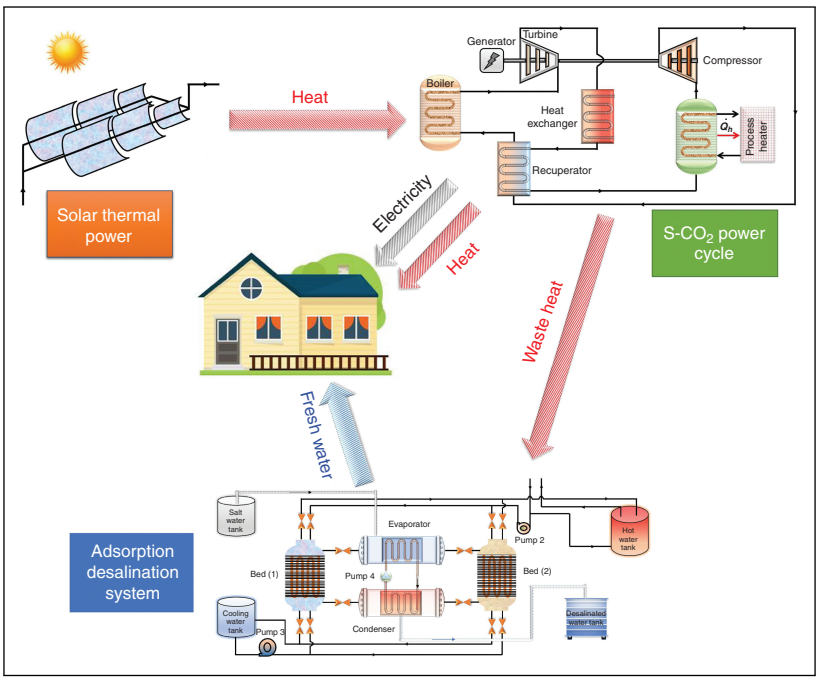
Mohd Asjad Siddiqui[1,2,3,*] and Eydhah Almatrafi[2,3]

[1]Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulatpur, Bawana Road, Delhi-110042, India
[2]Department of Mechanical Engineering, Faculty of Engineering at Rabigh, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[3]K.A. CARE Energy Research and Innovation Center, King Abdulaziz University, Jeddah 21589, Saudi Arabia, Center of Excellence in Desalination Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
*Correspondence to be sent to: Mohd Asjad Siddiqui, Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulatpur, Bawana Road, Delhi-110042, India; E-mail: siddiquimohd.asjad@gmail.com

Abstract

Hybrid solar-based integrated systems represent a viable solution for countries with abundant solar radiation, as they provide energy needs in an environmentally friendly way, offering a sustainable and economically advantageous energy solution that utilizes a free source of energy. Therefore, this research offers a thermodynamic evaluation of a novel integrated system driven by solar energy that aims to produce power, heating and freshwater. The integrated system consists of a parabolic trough collector that uses CO_2 as its working fluid and implements the supercritical carbon dioxide cycle to generate power and heating. The integrated system also includes an adsorption desalination system with heat recovery between the condenser and evaporator, which employs a cutting-edge material called an aluminium fumarate metal–organic framework to produce fresh water. For the modelling of a novel system, an engineering equation solver, which is considered a reliable tool for thermodynamic investigations, is employed. The effectiveness of an integrated system is evaluated using a mathematical model and different varying parameters are examined to ascertain their influence on thermal and exergy efficiency, specific daily water production and gained output ratio. The results revealed that the parabolic trough collector achieved a thermal efficiency of 67.2% and an exergy efficiency of 41.2% under certain conditions. Additionally, the thermal efficiencies for electrical and heating were obtained 24.68% and 9.85%, respectively. Finally, the specific daily water production was calculated, showing promising results and an increase from 7.1 to 12.5 m^3/ton/day, while the gain output ratio increased from 0.395 to 0.62 when the temperature of hot water increased from 65°C to 85°C, under the selected conditions.

Graphical Abstract



Keywords: PTC; thermodynamic analysis; supercritical CO_2 power cycle; adsorption desalination cycle; aluminium fumarate

Received: 25 April 2023; **Accepted:** 20 June 2023
© The Author(s) 2023. Published by Oxford University Press on behalf of National Institute of Clean-and-Low-Carbon Energy
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

Introduction

Over the past several decades, energy demand has grown exponentially in response to the growing population and the increasingly higher standards of living. The International Energy Agency predicts that the amount of energy that will be consumed will increase by 50% between 2018 and 2050 [1]. Combustion of non-renewable energy sources, including coal, oil and natural gas, currently serves as the dominant source of energy despite the fact that these resources are finite and their exhaustion is inevitable [2]. The use of fossil fuels has numerous negative environmental effects. Almost 87% of all human-caused carbon emissions come from carbon-containing fossil fuels [3]. In addition, fossil fuels contribute to global greenhouse gas emissions, which have detrimental effects on public health and climate change. The issues that have arisen from the use of fossil fuels can be resolved by switching to renewable forms of energy, such as solar energy. The use of solar energy has the potential to reduce various environmental risks while also providing cost-effective sources of thermal energy and electricity [4].

Concentrated solar power (CSP) is essential to address the current and anticipated power generation in tropical countries, utilizing the abundant solar heat along with other solar technologies [5]. CSP uses mirrors to magnify the solar radiation and then transmits the heat to a working fluid, which can then be subsequently deployed for various purposes, including the operation of power cycles and cooling systems. CSP systems are highly efficient and cost-effective, allowing us to take advantage of abundant and clean solar energy available [6, 7]. Various types of solar collectors can be implemented in solar power plants to harness the energy from the Sun. The parabolic trough collector (PTC) is the most popular choice, while there are other advanced options such as solar dish collectors, Fresnel collectors and solar towers [8, 9]. Because parabolic trough collectors are employed in 90% of all CSP plant installations worldwide [10], considerable attention has been paid to the development of new technological advances for use with this type of solar device. With the increasing demand for solar-thermal systems, solar collectors are becoming an incredibly reliable and efficient solution. Their design process is highly effective and has been very successful in satisfying customer needs. PTCs have been the most reliable CSP technology for a long time. They are used to generate heat at both medium and high temperatures, and are highly popular among applications [11]. It is generally agreed that PTC technology is the most well-established form of solar-thermal technology for the generation of power. This renewable-energy technology has been used in California's large power plants since the 1980s and its results have been encouraging for the future. As a result, this method of technology was chosen for the investigation.

Most applications that need heat at medium and high temperatures use PTCs to get the heat they need. Many arid regions of the world can solve their water and energy problems if desalination technology is incorporated into CSP plants [12]. Identifying technological synergies that have the potential to minimize the cost of combined power and desalination production compared with separate plants is one of the benefits of combining CSP and desalination facilities. Collaboration between CSP and desalination facilities has numerous advantages. It can help to resolve water- and energy-related conflicts by allowing the export of electricity to other countries while meeting the needs of the local population by increasing the water supply, which is partially paid for by selling power [13].

Using the supercritical carbon dioxide (S-CO_2) Brayton cycle is an ideal method to enhance the power cycle of any solar power plant because the S-CO_2 Brayton cycle is a thermodynamic cycle that uses CO_2 as the working fluid, resulting in increased efficiency and power output from solar power plants. It also requires less space and has lower capital costs than traditional cycles, making it a more cost-effective option. Researchers paid more attention to this technology because it was simple and compact. Khan and Mishra [14] conducted a study that examined the thermal performance of an integrated system composed of a pre-compression supercritical CO_2 cycle and an organic Rankine cycle (ORC), powered by a solar power tower. They found that the net power output and thermal efficiency of the pre-compression cycle were improved by 4.51% and 4.52%, respectively, using an ORC. In [15], an evaluation of the performance of various S-CO_2 Brayton cycles connected to a solar power tower was analysed. It has been demonstrated in [16] that an S-CO_2 unit is capable of providing an efficiency that is equivalent to or higher than that produced by supercritical or superheated steam cycles, while also being more cost-effective than the Rankine cycle in terms of development expenses. According to the relevant literature [17], the S-CO_2 Brayton cycle power generation is the most promising technology for future concentrating solar-thermal power systems. This is because it can convert energy more efficiently than solar power plants that use the steam Rankine cycle.

The desalination process has evolved into a well-known and widely utilized technology globally. It is the process of turning seawater into drinking water by removing salts and other minerals with a total dissolved solids (TDS) concentration of 35 000 p.p.m. There are three main categories of desalination technology: distillation methods such as solar distillation, vacuum compression (VC) [18], multi-effect distillation (MED) [19] and multi-stage flash (MSF) [20]; membrane technologies such as electro-dialysis (ED) [18] and reverse osmosis (RO) [20]; and chemical methods such as ion exchange desalination and gas hydrate [21]. One of the main issues with these types is that they use a large amount of power, making them expensive and causing a large amount of CO_2 to be released [22]. Adsorption desalination (AD) technology has been continuously developed over the last few decades and it is currently considered one of the promising alternatives to conventional desalination technologies that can help solve the issues of increased power consumption and environmental degradation that were discussed earlier [23, 24, 25]. AD technology has been shown to consume less energy than traditional methods of desalination [26] due to its ability to utilize low-temperature energy sources such as solar power [27] or waste heat [28] and its lack of moving parts, which reduces maintenance costs. In addition, an environmentally friendly refrigerant like water is used in the process [29] and cooling is achieved during freshwater generation, reducing the need for external cooling. The TDS content of this system output is 10 p.p.m., indicating a high quality of the water produced [30].

In recent years, much research has been conducted on developing solar-powered desalination technology and this has been made possible by the qualities of CSP systems. There are two types of solar-based desalination systems: pure solar-based desalination systems that only produce fresh water [31, 32, 33] and hybrid solar-powered power and desalination systems [34, 35, 36]. Olwig *et al.* studied the performance of a CSP–RO and CSP–thermal vapour compressor (TVC)–MED system powered by a PTC was studied by Olwig *et al.* [37]. This system incorporated a regenerative steam Rankine cycle. This cycle of the proposed system had the capacity to generate 42 MWe of electricity, while the potable water production capacity of the desalination system was 24000 m^3 per day. Demir and Dincer [38] developed an integrated

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

system based on a solar tower, employing a steam Rankine cycle, gas turbine cycle and a flash distillation system for the production of electricity and potable water. Their findings indicated that the combined power and desalination system could produce 19.15 MW of electricity from a gas turbine Rankine cycle, 2.63 MW from a Rankine cycle and 0.0329 MW from a thermoelectric generator (TEG), while the flash desalination system could produce 290 m^3/day of fresh water. Wang *et al.* [39] conducted a study that outlined the development and evaluation of a solar-powered S-CO_2 Brayton cycle and MSF desalination system, with the goal of achieving the co-production of power and fresh water. The efficiency of the S-CO_2 Brayton cycle was determined to be 36.6% based on the results of operational behaviour estimation for the proposed system. In addition, it was discovered that the proposed system had a daily output of 50.1 MW of electricity and 4050.8 t of freshwater.

Although several investigations on solar-based power, heating and desalination systems have been conducted, to the authors' knowledge, the literature has not yet investigated the feasibility and usability of a hybrid system that employs an S-CO_2 cycle and an AD cycle that are both directly driven by solar-thermal systems with PTC. To address the lack of research in this area, this study introduces an innovative PTC-based power, heating and desalination method that incorporates both the S-CO_2 cycle and the AD cycle with heat recovery (it shows that the condenser is cooled using chilled water by making use of the cooling effect) using aluminium fumarate material. As a result, the primary objective of this study is the effective utilization of solar irradiations to satisfy the requirements for power, heating and pure water employing the novel system mentioned above. The development of such systems appears to be capable of meeting the energy requirements of commercial units and industries. To summarize, the main novelties and objectives of this paper are:

- a novel solar-based integrated system is proposed and modelled, utilizing the S-CO_2 cycle for the production of power and heating, where the waste heat of the S-CO_2 cycle is recovered by a low-temperature AD cycle with heat recovery to produce potable water;
- a parametric investigation is conducted to explore the effect of several parameters on the performance of a solar-based integrated system;
- energy- and exergy-based assessment of the proposed solar-based integrated system for the power, heating and potable water;
- specific daily water production (SDWP) and gained output ratio (GOR) are the primary evaluation indices used to study the performance of the AD system;
- a comparative analysis is being performed to evaluate the SDWP of freshwater produced by the current integrated system and other integrated AD systems using various adsorbent materials, as reported in the existing literature.

1 System description

In this study, a thermodynamically evaluated integrated system based on PTC is designed. Fig. 1 shows the schematic of the novel

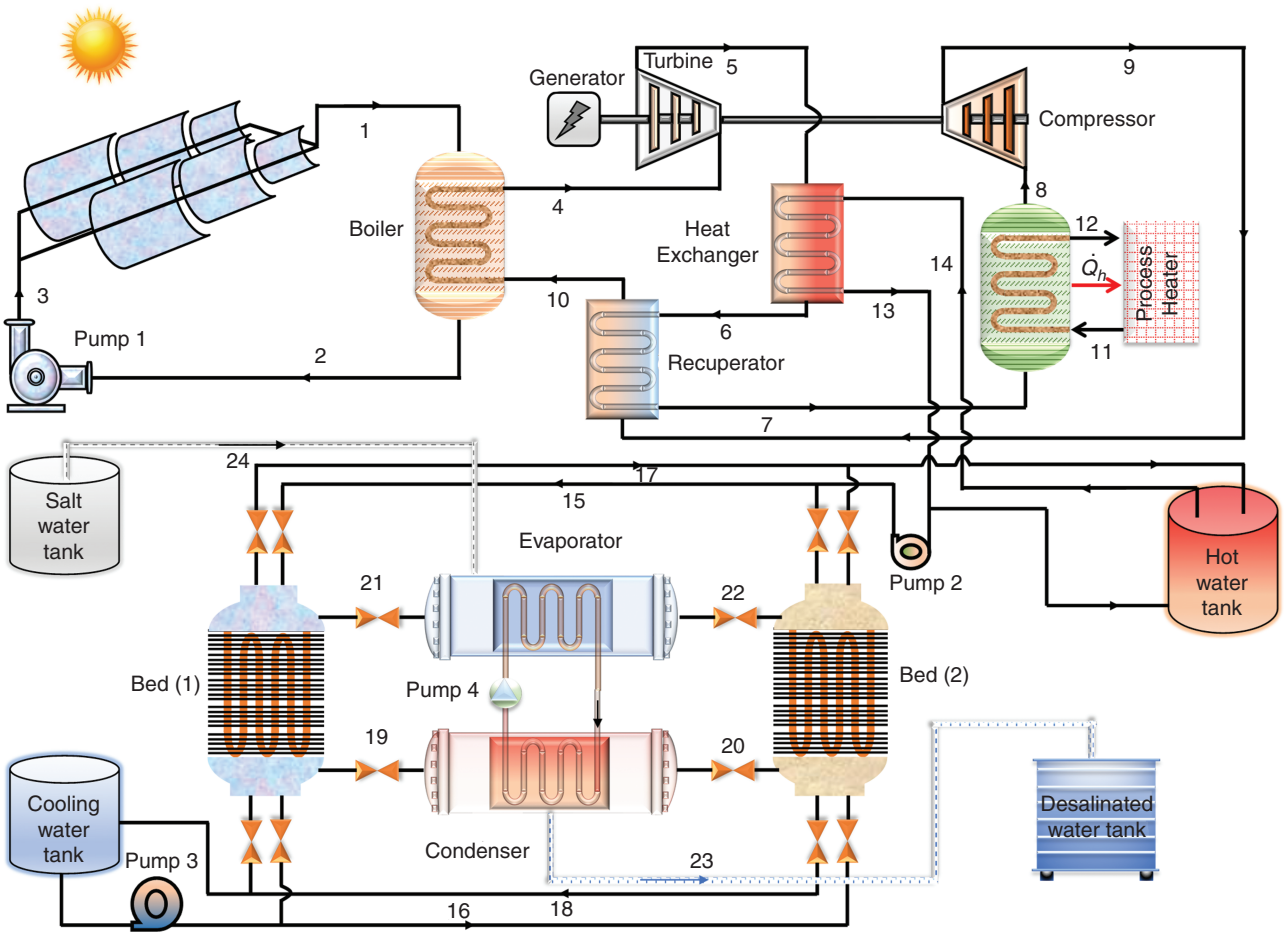


Fig. 1: Schematic diagram of the novel PTC-driven integrated system for power, heating and freshwater.

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

system. The system may be categorized into three primary parts: the PTC thermal subsystem, the S-CO_2 power cycle and the AD cycle with heat recovery.

The most efficient method of collecting solar energy is through the use of PTCs. In this system, CO_2 is used as a heat transfer fluid (HTF) in the PTC field because it has a higher heat transfer coefficient, a higher density and a lower viscosity than other fluids. The high-temperature HTF enters the boiler of the S-CO_2 power cycle to heat the high-pressure CO_2 at the exit of the compressor, which then flows through the recuperator and boiler to absorb heat before entering the S-CO_2 turbine and producing electricity. The HTF leaves the boiler and travels to the PTC field through a solar pump. The high-temperature CO_2 stream that exits the turbine then goes into the heat exchanger to act as a heat source for AD. The hot exit stream from the heat exchanger goes through the recuperator, where it exchanges heat with the high-pressure cold flow. The CO_2 stream is then cooled to State Point 8 after passing through the pre-cooler to release heat. The pre-cooler is made in a creative way to replace a condenser and provide heating capacity. At State Point 9, the stream is compressed by a compressor and then the heat from the recuperator is absorbed. The preheated S-CO_2 is then added to the boiler to finish the power cycle.

The AD system consists of components such as the evaporator, two beds (adsorber/desorber), a condenser and pumps for recirculating fluids. In the case of a two-bed AD system, the adsorption process in the first bed is running while the desorption process in the second bed is ongoing. After the completion of a half-cycle, they are exchanged and so on. The first bed will be connected to the evaporator and will function as an adsorption bed, while the second bed will be connected to the condenser and will function as a desorption bed. The adsorbent in the specified adsorption processes is responsible for maintaining the process of vapour uptake and the heat generated is dissipated by the flowing cool water in these processes. During the exothermic adsorption process, seawater is fed into an evaporator, where it changes phase from liquid to vapour. After that, the water vapour is absorbed by the dry sorbents that are placed inside the adsorber. Saturated aluminium fumarate can be regenerated by a low-temperature heat exchanger that recovers waste heat from the exit of the S-CO_2 turbine, known as the desorption process, and it involves the use of hot water with temperatures varying from 85°C to 65°C. The water vapour that is desorbed is then condensed within a condenser, resulting in pure water.

2 Thermodynamic modelling and assumptions

2.1 Assumptions

The following are the model assumptions [24, 40, 41]:

(i) The integrated system has a well-sealed connection and is operated in a steady state.
(ii) Alterations in the kinetic and gravitational potential energies are not considered.
(iii) The pressure drops and heat losses that occur across the various pipes and components are not taken into account.
(iv) An isenthalpic flow was taken into consideration for each of the expansion valves.
(v) The adsorption beds have consistent pressure, temperature and vapour throughout the entire bed.
(vi) Heat loss is not taken into account by the parts of an adsorption system.
(vii) The thermal resistance between the adsorbent material and the adsorbent tube is ignored.
(viii) The flow resistance caused by the water flow in the pipeline is ignored.
(ix) The properties of the adsorbate, the metal tube and the fluids remain constant.

2.2 Energy and exergy analysis

Taking into account the steady-state condition of the control volume, the mass rate balances and the energy and exergy balance relations for each component can be described as [38, 41]:

$$\sum_i \dot{m}_i - \sum_o \dot{m}_o = \frac{dm_{cv}}{dt} \tag{1}$$

The energy balance equation in steady state is given by:

$$\dot{Q} - \dot{W} + \sum_i \dot{m}_i \left(h_i + \frac{C^2}{2} + gz_i\right) - \dot{m}_o \left(h_o + \frac{C^2}{2} + gz_o\right) = \frac{dE_{cv}}{dt} \tag{2}$$

The exergy balance equation can be presented as:

$$\dot{E}_{x,Q} + \sum_i \dot{m}_i e_{x_i} = \dot{E}_{x,W} + \sum_o \dot{m}_o e_{x_o} + \dot{E}_{x,D} \tag{3}$$

Each point exergy of a flowing stream is expressed as:

$$\dot{E}_x = \dot{E}_x^{ph} + \dot{E}_x^{ch} \tag{4}$$

It is possible to ignore the chemical exergy in the proposed cycle because no chemical reactions take place. As a result, at each moment in time during a state, the exergy $\dot{E}_i$ is equal to the physical exergy $\dot{E}_x^{ph}$ that can be obtained by:

$$\dot{E}_i = \dot{E}_x^{ph} = \dot{m}_i[(h_i - h_0) - T_0\,(s_i - s_0)] \tag{5}$$

where s_o and h_o are the ambient entropy and enthalpy, respectively.

The mass/energy balance equation for the individual element is established and given in Table 1 by implementing the above equations in the system components.

2.3 The examined PTC

The Eurotrough ET-150 [42], a well-known commercial PTC, was selected for analysis in this study due to its widespread use in previous research and applications [43, 44]. This collector has a concentration ratio of ∼26.37, which is a typical value among commercial parabolic trough collectors. The evacuated-tube receiver and reflector are the most important components of the collector. The reflector is typically manufactured by bending a reflective material into a parabolic shape, while the evacuated tube consists of a glass cover and selective absorber separated by an extremely low-pressure environment to minimize convection losses and increase thermal efficiency [4]. Table 4 provides detailed information on the geometrical parameters of the model, as well as other crucial properties of the collector.

2.4 PTC field model

The solar energy on the collector aperture is calculated as [4]:

$$\dot{Q}_{Solar} = G_b A_c \tag{6}$$

where G_b denotes the direct beam solar irradiation and A_c is the collector aperture area.

The useful heat output (Q_u) is determined as [45]:

$$\dot{Q}_u = \dot{m} c_p\,(T_{out} - T_{in}) \tag{7}$$

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

Table 1: The equations describing the energy and exergy balances for components of a solar-based S-CO_2 system

Component		Balance equations
Boiler	Energy balance	$\dot{m}_B\,(h_1 - h_2) = \dot{m}_3\,(h_4 - h_9)$
	Exergy balance	$\dot{E}_{x,D,B} = \dot{m}_1(e_{x,1} - e_{x,2}) + \dot{m}_3(e_{x,9}$ -- $e_{x,4})$
Solar pump	Energy balance	$\dot{W}_P = \dot{m}_2(h_3 - h_2)$
	Exergy balance	$\dot{E}_{x,D,P} = \dot{m}_2\,(e_{x,2} - e_{x,3}) + \dot{W}_P$
S-CO_2 turbine	Energy balance	$\dot{W}_T = \dot{m}_4\,(h_4 - h_5)$
	Exergy balance	$\dot{E}_{x,D,T} = \dot{m}_4 e_{x,4} - \dot{m}_5 e_{x,5} - \dot{W}_T$
Heat exchanger	Energy balance	$\dot{Q}_{HE} = \dot{m}_{13}\,(h_{13} - h_{14}) = \dot{m}_5\,(h_5 - h_6)$
	Exergy balance	$\dot{E}_{x,D,HE} = \dot{m}_5\,(e_{x,5} - e_{x,6}) + \dot{m}_{13}\,(e_{x,14} - e_{x,13})$
Recuperator	Energy balance	$\dot{m}_9\,(h_{10} - h_9) = \dot{m}_6\,(h_6 - h_7)$
	Exergy balance	$\dot{E}_{x,D,rec} = \dot{m}_9\,(e_{x,9} - e_{x,10}) + \dot{m}_6(e_{x,6} - e_{x,7})$
Process heater	Energy balance	$\dot{Q}_{ph} = \dot{m}_7(h_7 - h_8)$
	Exergy balance	$\dot{E}_{x,D,ph} = \dot{m}_7 e_{x,7} - \dot{m}_8 e_{x,8}$
Compressor	Energy balance	$\dot{W}_C = \dot{m}_8\,(h_9 - h_8)$
	Exergy balance	$\dot{E}_{x,D,C} = \dot{W}_C + \dot{m}_8 e_{x,8} - \dot{m}_9 e_{x,9}$

The thermal efficiency of the PTC is expressed as:

$$\eta_{th,\ PTC} = \frac{\dot{Q}_u}{\dot{Q}_{Solar}} \tag{8}$$

The rate of exergy destruction during the process of solar heat absorption from the Sun to the receiver is determined as [46]:

$$\dot{E}_{x,D,s-r} = \eta_{opt}\dot{E}x_{Solar} - Q_{abs}\left(1 - \frac{T_{am}}{T_r}\right) \tag{9}$$

where $\dot{E}x_{Solar}$ is the exergy flow of the undiluted solar beam irradiation and is evaluated by using the Petela model, as it is given below [47]:

$$\dot{E}x_{Solar} = \dot{Q}_{Solar}\left[1 - \frac{4}{3}\left(\frac{T_{am}}{T_{sun}}\right) + \frac{1}{3}\left(\frac{T_{am}}{T_{sun}}\right)^4\right] \tag{10}$$

where T_{sun} is the Sun temperature, which can be taken to be equal to 5770 K.

The absorbed solar energy ($\dot{Q}_{abs}$) is expressed as:

$$\dot{Q}_{abs} = \eta_{opt}\dot{Q}_{Solar} \tag{11}$$

The rate of exergy destruction during heat transfer from the receiver to the fluid is expressed as [46]:

$$\dot{E}_{x,D,r-f} = \dot{Q}_u\left(1 - \frac{T_{am}}{T_r}\right) - \dot{E}_{x,u} \tag{12}$$

The useful exergy output is calculated as [45]:

$$\dot{E}_{x,u} = \dot{Q}_u - \dot{m}\ c_p.T_{am}\ ln\left[\frac{T_{out}}{T_{in}}\right] - T_{am}\dot{m}\frac{\Delta P}{\rho_{fm}T_{fm}} \tag{13}$$

The following formula is used to compute the total exergy destruction ($\dot{E}_{x,D,\ total}$):

$$\dot{E}_{x,D,\ total} = \dot{E}_{x,D,s-r} + \dot{E}_{x,D,r-f} \tag{14}$$

The exergy efficiency of the PTC is expressed as [45]:

$$\eta_{ex,\ PTC} = \frac{\dot{E}_{x,u}}{\dot{E}x_{Solar}} \tag{15}$$

2.5 AD process

2.5.1 Adsorption properties

Adsorption properties can be evaluated using the following equations [48, 49]:

$$W^* = W_o exp\left[-\left\{\frac{RT}{E}ln\left(\frac{P_s}{P}\right)\right\}^n\right] \tag{16}$$

where W^* and W_o denote the anticipated and maximum capacity for adsorption, while R and n refer to constants related to universal gas and exponent fitting. Additionally, E represents the characteristic energy.

$$\frac{dW}{dt} = \frac{FD_{so}exp\left\{\frac{-E_a}{RT_a}\right\}}{R_p^2} * (W^* - W) \tag{17}$$

where F, D_{so}, E_a and R_p refer to the particle constant, the pre-exponential coefficient, the activation energy and the particle radius, respectively.

$$H_{ads} = h_{fg} + E\left[ln\left(\frac{W_o}{W}\right)^{\frac{1}{n}}\right] + ET\alpha\left[ln\left(\frac{W_o}{W}\right)^{\frac{1-n}{n}}\right] \tag{18}$$

where H_{ads}, h_{fg} and α are the isosteric heat of adsorption, the heat of evaporation and the thermal expansion, respectively.

2.5.2 Adsorption energy balance

The energy balance equations for the adsorption/desorption beds, the condenser and the evaporator are depicted in Equations (19–22) [26, 27].

Adsorption bed:

$$\frac{dT_{ads}}{dt}\left[\left(M_{am}C_{p,am}\right) + \left(M_{cu}C_{p,cu}\right) + \left(M_{am}C_{p,Mixture}\frac{dw_{ads}}{dt}\right)\right] = \left[\left(qH_{ads}M_{am}\frac{dw_{ads}}{dt}\right) - \dot{m}_w C_{p,cwa}\left(T_{cwa,out} - T_{cwa,in}\right)\right] \tag{19}$$

Desorption bed:

$$\frac{dT_{des}}{dt}\left[\left(M_{am}C_{p,am}\right) + \left(M_{cu}C_{p,wd,cu}\right) + \left(M_{am}C_v\frac{dw_{des}}{dt}\right)\right] = \left[\left(qH_{ads}M_{am}\frac{dw_{des}}{dt}\right) - \dot{m}_{hw,d}C_{p,hw,d}\left(T_{hw,d,out} - T_{hw,d,in}\right)\right] \tag{20}$$

Condenser:

$$\frac{dT_{cond}}{dt}\left[\left(M_{cu,cond}C_{p,cu}\right) + \left(M_vC_v\right)\right] = \left[\left(-M_{am}h_{fg}\frac{dw_{des}}{dt}\right) + \dot{m}_{cwc}C_{p,w}\left(T_{chill,in} - T_{cwc,out}\right)\right] \tag{21}$$

Evaporator:

$$\frac{dT_{cond}}{dt}\left[\left(M_{cu,evap}C_{p,cu,evap}\right) + \left(M_vC_v\right)\right] = \left[\left(-M_{am}h_{fg}\frac{dw_{ads}}{dt}\right) + \dot{m}_{chill}C_{p,chill}\left(T_{chill,in} - T_{chill,out}\right)\right] \tag{22}$$

The temperature of the water that exits each heat exchanger can be determined using the heat transfer equation by employing the logarithmic mean temperature difference approach as [27]:

$$T_{out} = T_o + (T_{in} - T_o)exp\left(\frac{-UA}{\dot{m}c_p(T_o)}\right) \tag{23}$$

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

2.6 Performance parameters

The net generated power of the S-CO_2 power cycle is given by:

$$\dot{W}_{net} = \dot{W}_T - \dot{W}_C \tag{24}$$

The thermal and exergy efficiencies are expressed as follows:

$$\eta_{th,\ S-CO2} = \frac{\dot{W}_{net} + \dot{Q}_h}{\dot{Q}_{Solar}} \tag{25}$$

$$\eta_{ex,\ S-CO2} = \frac{\dot{W}_{net} + \dot{E}_{x,h}}{\dot{E}x_{Solar}} \tag{26}$$

The performance of the AD cycle was estimated by calculating the GOR and SDWP using Equations (29) and (30), respectively [50]:

$$Q_{chill}^{cycle} = \int_0^{t_{cycle}} \left\{ \left(\dot{m}_{c,w} C_{p,w} \right)_{chill} \left(T_{chill,in} - T_{chill,out} \right) \right\} dt \tag{27}$$

$$Q_{des}^{cycle} = \int_0^{t_{cycle}} \left\{ \left(\dot{m}_{hw} C_{pw} \right)_{des} \left(T_{hw,in} - T_{hw,out} \right) \right\} dt \tag{28}$$

$$GOR = \frac{Q_{chill}^{cycle}}{Q_{des}^{cycle}} \tag{29}$$

$$SDWP = \int_0^{t_{cycle}} \frac{\left\{ \left(\dot{m}_{c,w} C_{p,w} \right) \left(T_{cw,out} - T_{cw,in} \right) \right\} \tau dt}{M_{am} h_{fg}} \tag{30}$$

3 Validation of system

A comparison of the data generated by the simulation and the data reported in the research literature [51, 52] was conducted to validate the accuracy of the PTC and S-CO_2 power cycle, as demonstrated in Tables 2 and 3. Fig. 2 shows how the results of the current work and earlier research for SDWP compare. It is evident from the figure that the results of this study are superior to those of previous studies involving adsorbents such as silica gel [26, 40, 53, 54], aluminium fumarate [29], activated clay [55] and

Table 2: Validation results related to the heat transfer fluid (CO_2) of the PTC [51]

Input parameter					Reference [51]	Present work	Deviation (%)
Heat transfer fluid	Inlet pressure (kPa)	Solar irradiation intensity (W/m^2)	Mass flow rate (kg/s)	Inlet temperature (°C)	Outlet temperature (°C)	Outlet temperature (°C)	Outlet temperature (°C)
CO_2	5000	900	1.1	225	500	505.3	1.048
	7500	900	1.1	225	495	499.7	0.941

Table 3: Validation of thermodynamic simulation of the S-CO_2 power cycle [52]

Input parameter				Reference [52]	Present work	Deviation (%)
Turbine inlet temperature (°C)	Compressor inlet temperature (°C)	Compressor inlet pressure (kPa)	Compressor outlet pressure (kPa)	Thermal efficiency (%)	Thermal efficiency (%)	Thermal efficiency (%)
550	32	7630	20 000	36.87	37.42	1.492

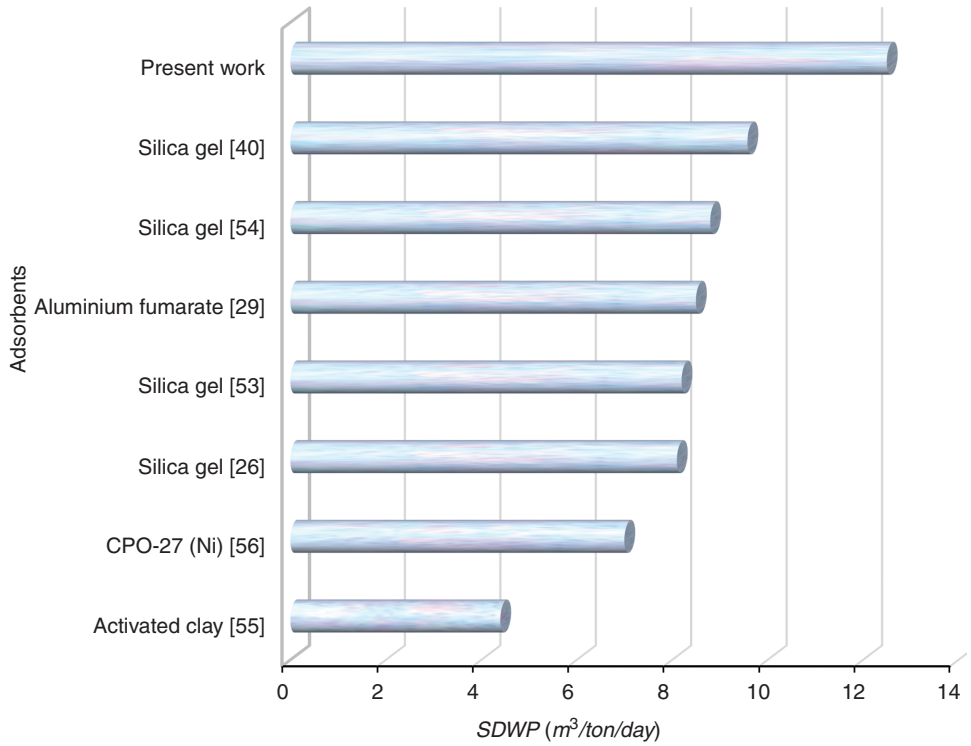


Fig. 2: Comparison of the present and previous SDWP studies [26, 29, 40, 53, 54, 55, 56].

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

CPO-27 (Ni) [56]. The simulation results demonstrate strong correspondence with the findings in the literature, thus validating the accuracy of the simulation model.

4 Result and discussion

This investigation provides a thermodynamic evaluation of a new solar-powered supercritical CO_2 cycle and low-temperature AD cycle for electricity, heating and potable water. A mathematical model is formulated to perform a simulation for examining parameters. The properties related to the suggested system for baseline operative conditions are specified in Tables 4 and 5.

Figs 3 and 4 show the influence of the inlet temperature of the HTF on the thermal and exergy efficiencies of the PTC when it is running with CO_2 for the five different mass flow rates that were taken into consideration. The thermal performance of the PTC is observed to decline when the inlet temperature increases from 127°C to 927°C. According to Fig. 3, it is abundantly obvious that the lowest mass flow rates result in the lowest thermal performance compared with the other mass flow rates where there is a substantial difference. The thermal efficiency is highest for a mass flow rate of 0.3 kg/s and ranges from 69.5% to 45.12% as the CO_2 inlet temperature increases. The other four curves, which represent different mass flow rates, are close to each other, which shows that they can be used as operating mass flow rates. On the other hand, as shown in Fig. 4, the highest exergetic performance can be achieved with the lowest mass flow rates. This result can be explained as a consequence of the fact that the exergy efficiency of the PTC is closely associated with its pressure losses, which increase with increasing mass flow rates.

Fig. 5 shows the variation of thermal and exergetic performance of the PTC with increasing solar irradiation intensity. The thermal performance is reduced with increasing solar irradiation from 600 to 1000 W/m^2. The variation in the thermal efficiency is low and ranges from 67.47% to 66.87%. A reduction in the thermal performance is observed due to the higher temperature of the PTC for a higher pressure ratio. The lower thermal efficiency of the PTC is caused by the higher temperature present in the solar collector. The performance of the PTC is heavily affected by these two parameters. On the other hand, when the intensity of the solar irradiation changes from 600 to 1000 W/m^2, the exergy efficiency of the PTC varies very significantly and ranges from 40.19% to 41.86%. This tendency is primarily observed because receiving more solar irradiation results in a more efficient usage of the PTC field. Consequently, as indicated in the figure, the exergetic efficiency of the system can be enhanced when the intensity of the solar irradiation increases.

The influence of solar irradiance on the outlet temperature of the HTF and the mass flow rate of the S-CO_2 power cycle has been demonstrated in Fig. 6. The data in the figure illustrate that as solar irradiation intensity (600–1000 W/m^2) increases, the outlet temperature of the HTF increases proportionally. This is due to the fact that when the thermal conductivity of CO_2 increases, the convection coefficient also increases, leading to a higher HTF outlet temperature. However, the mass flow of the S-CO_2 cycle increases with the change in intensity of solar irradiation. This result can be explained due to the combined effect of the specific heat and density of the CO_2 HTF, which is responsible for the increase in the mass flow rate of the S-CO_2 cycle.

Fig. 7 shows the influence of pressure at the turbine inlet on the thermal efficiency of the S-CO_2 power cycle for different solar irradiation intensities. The influence of pressure at the turbine inlet on the thermal efficiency of the S-CO_2 power cycle of the system is observed for a wide range of solar irradiations (600–1000 W/m^2). It is clear from the figure that the thermal efficiency of the S-CO_2 power cycle increases with increasing pressure at the turbine inlet from 11 000 to 15 000 kPa. It is found that with the rise in intensity of solar irradiation, the thermal efficiency increases because an increase in solar irradiation intensity improves the collector efficiency, which results in a higher mass flow rate and higher outlet temperature of the HTF from the PTC. An increase in the outlet temperature of the HTF improves the quality of the operating fluid at the turbine entry, which results in the improved power output of the S-CO_2 cycle. The variation in thermal efficiency of the S-CO_2 power cycle is higher and ranges from 26.12% to 44.31% at 1000 W/m^2 with the variation in turbine inlet pressure from 11 000 to 15 000 kPa. It is understood that the enthalpy difference throughout the turbine improves as the inlet pressure increases, resulting in an improvement in the power output of the turbine.

The influence of the turbine inlet pressure on the electrical and heating efficiencies of the S-CO_2 power cycle is illustrated in Fig. 8. It indicates that both the electrical and heating efficiencies of the S-CO_2 cycle are increased when the inlet pressure of the turbine is raised from 11 000 to 15 000 kPa. This can be attributed to an increase in the enthalpy difference across the turbine, resulting in an improved power and heating output as a consequence of the

Table 4: Input data for modelling of the developed sCO_2 power cycle [4, 45]

Reference temperature	25°C
Reference pressure	101.32 kPa
HTF inlet temperature	327°C
Solar irradiation	0.8 kW/m^2
Sun temperature	5770 K
Collector aperture area	69.6 m^2
Concentration ratio of collector	26.37
Tube length	12 m
Tube width	5.8 m
Optical efficiency	80%
Diameter of inner receiver	66 mm
Diameter of outer receiver	70 mm
Diameter of inner cover	120 mm
Diameter of outer cover	125 mm
Ambient air velocity	1 m/s
Isentropic efficiency of S-CO_2 turbine	95%
S-CO_2 turbine inlet pressure	13 000 kPa
Effectiveness of heat exchangers	0.9

Table 5: Operational and geometric parameters for AD cycle [21, 29, 57]

Half-cycle time	100–1000 s
Switching time	70 s
Adsorbent mass	0.30 kg
Water flow rate	5--19 L/min
Desorption hot water temperature	65--85°C
Adsorption cooling water temperature	25--35°C
Length	350×10^{-3} m
Fin width	173×10^{-3} m
Fin thickness	0.2×10^{-3} m
Fin height	28×10^{-3} m
Bed tube thickness	0.8×10^{-3} m
Tube outlet diameter	10×10^{-3} m
Number of heat exchangers/bed	2
Specific heat	970 J/(kg.K)
Thermal conductivity	0.13 W/(m.K)
Thermal diffusivity	257.7×10^{-9} m^2/s

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

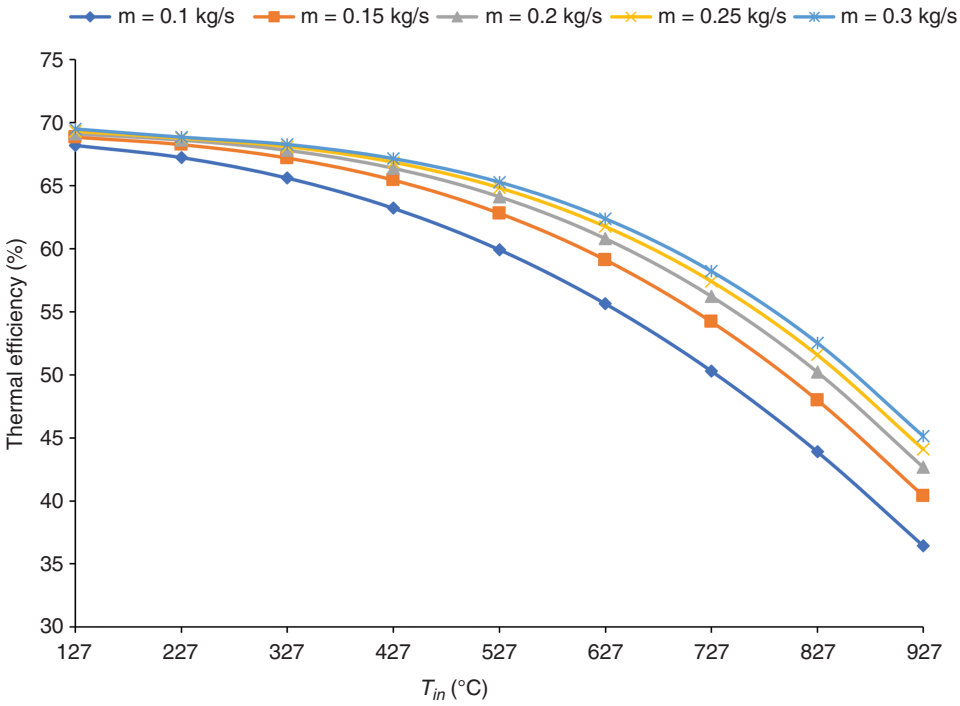


Fig. 3: Influence of the inlet temperature of the heat transfer fluid on the thermal efficiency of the PTC.

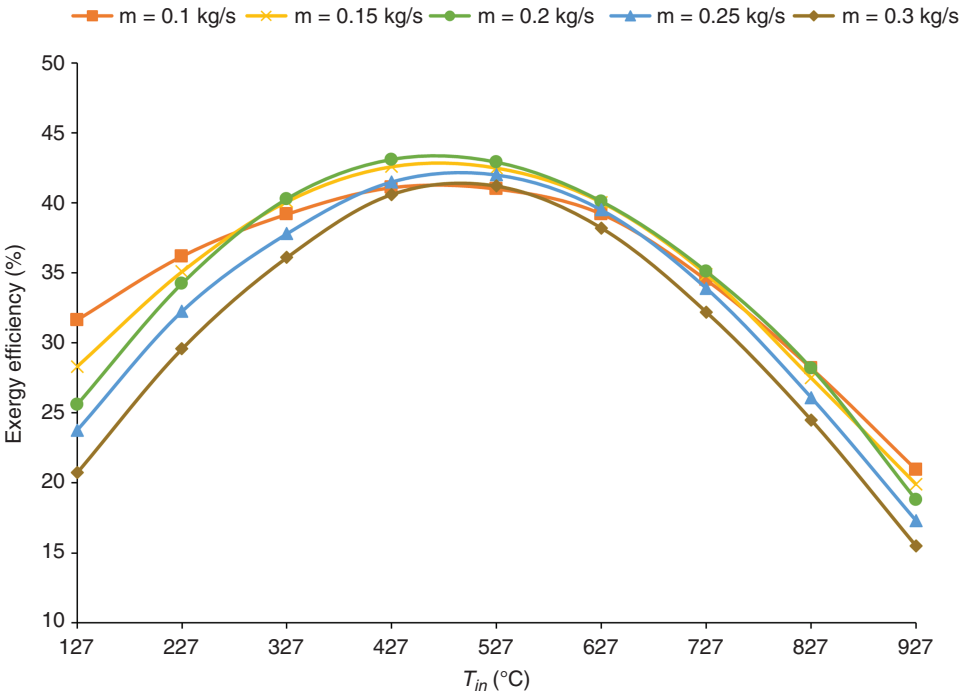


Fig. 4: Influence of the inlet temperature of the heat transfer fluid on the exergy efficiency of the PTC.

higher inlet pressure. As observed in Fig. 8, when the turbine inlet pressure varies from 11 000 to 15 000 kPa, there is an increase in the electrical and heating efficiency from 17.29% to 29.89% and 6.89% to 11.93%, respectively.

Fig. 9 represents the SDWP and GOR of the AD cycle with heat recovery for assorted cooling/heating water flow rates (passing through the adsorption beds) for three different cooling temperatures, namely 25°C, 30°C and 35°C. It is observed that as the water flow rate increases to 12 L/min, the SDWP and GOR increase dramatically and then the improvement is slight. This is due to the amount of heat needed to heat and cool down the adsorption material as a function of the water flow rate. It has also

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

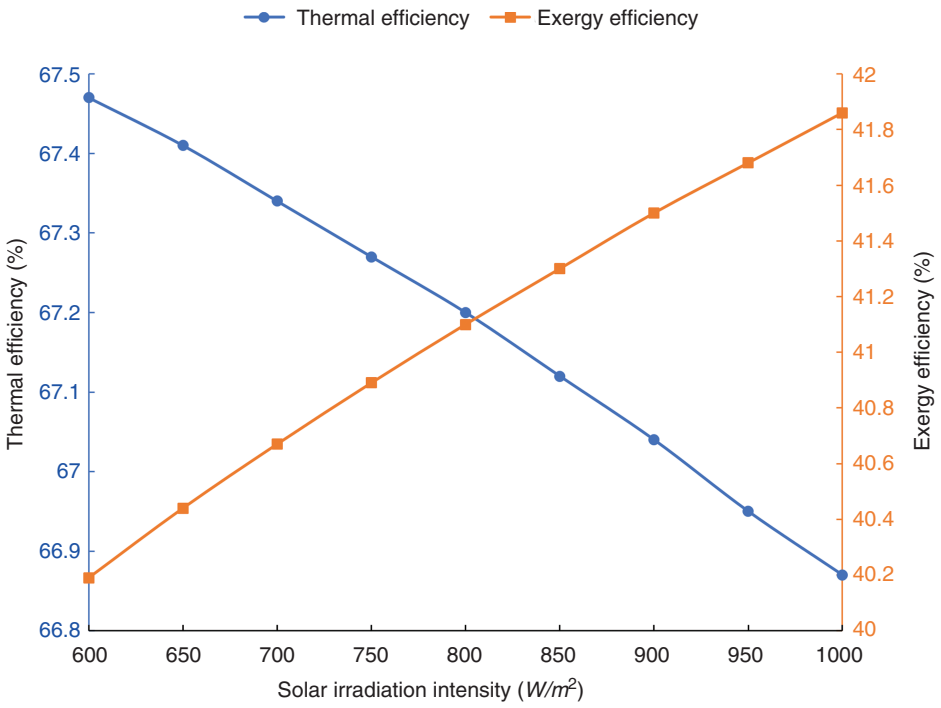


Fig. 5: Effect of the solar irradiation intensity on the thermal and exergy efficiencies of the PTC.

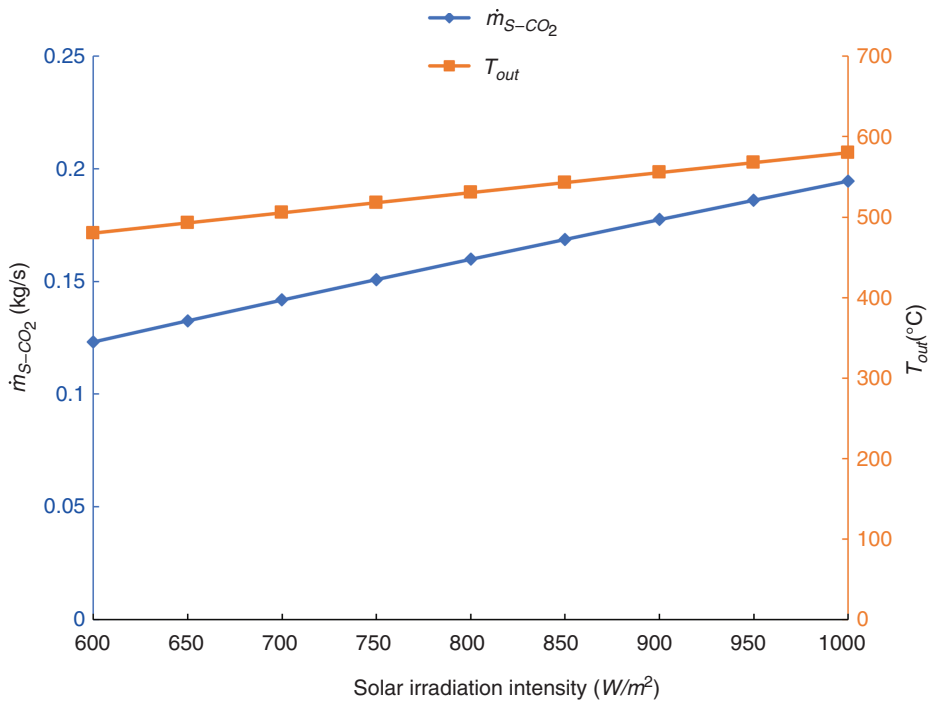


Fig. 6: Influence of the solar irradiation intensity on outlet temperature of the HTF and mass flow rate of the S-CO_2 power cycle.

been observed that the performance parameters of the AD cycle decrease with a rise in the cooling water temperature. This tendency is primarily observed due to the isotherm behaviour of the adsorption material, in which the uptake of the adsorption material improves substantially as the working relative pressure increases [20] upon lowering the cooling water temperature. Furthermore, the effectiveness of the adsorption material is highly dependent on the relative working pressure. Therefore, it is found from Fig. 9 that the flow rate of 12 L/min achieves the required amount of heat and then the performance improvement is slight.

Fig. 10 depicts the impact of half-cycle time on the SDWP and GOR of the AD cycle at a fixed water flow rate (12 L/min) and cooling water temperature (30°C) for three different hot water

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

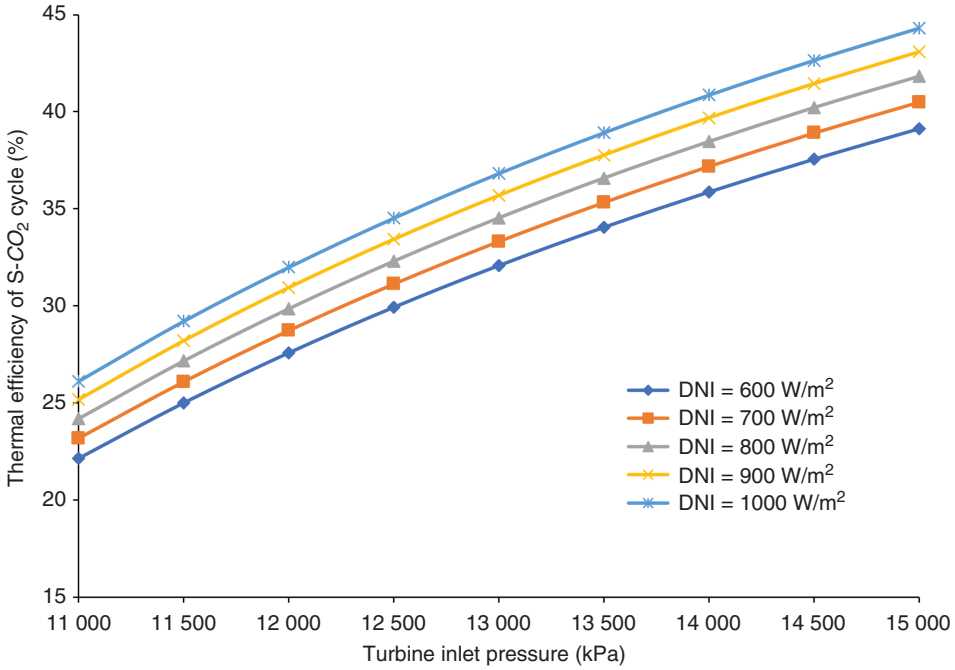


Fig. 7: Effect of the turbine inlet pressure on the thermal efficiency of the S-CO_2 power cycle.

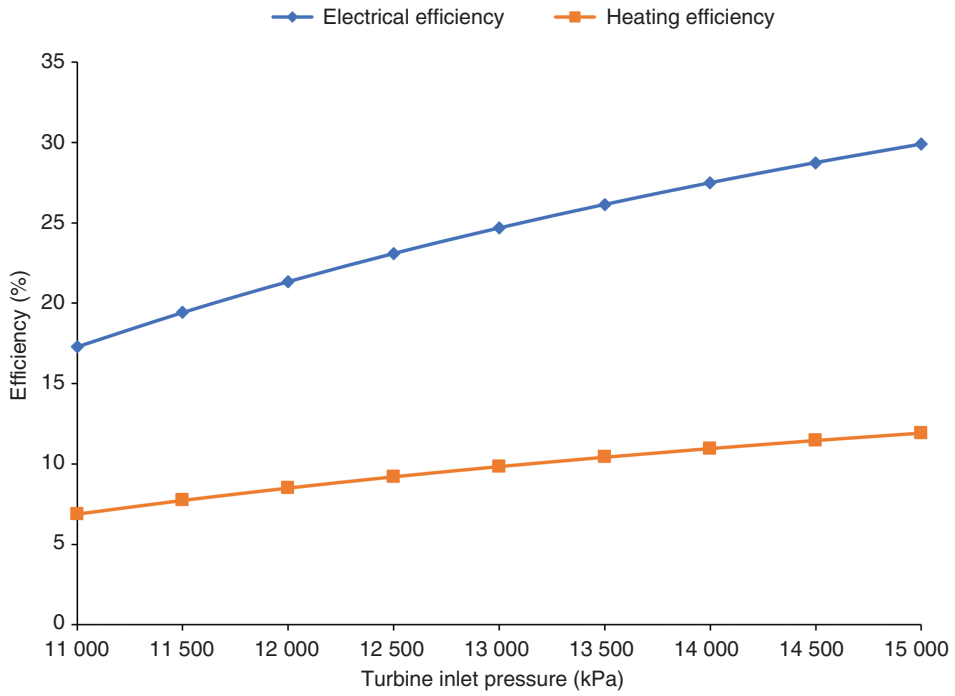


Fig. 8: Variation in the turbine inlet pressure on the electrical and heating efficiencies of the S-CO_2 power cycle.

temperatures, namely 75°C, 80°C and 85°C. It is clear from the figure that SDWP rises as the half-cycle time rises; however, the curve reaches its maximum at between 300 and 500 s, and the GOR rises linearly as the half-cycle time rises. It is noted that the peak value of the SDWP reaches a maximum at a half-cycle time of 400 s. In addition, the SDWP and GOR rise with an increase in the hot water temperature. This result can be explained by the fact that the regeneration process works better when the temperature of the hot water is higher. The highest SDWP of 12.5 m^3/ton/day was achieved at a half-cycle time of 400 s, with 0.62 GOR under 85°C of hot water temperature.

Fig. 11 illustrates the influence of increasing the temperature of the hot water on the performance parameters of the AD cycle. It is evident from the figure that the SDWP and GOR rise proportionally when the temperature rises from 65°C to 85°C. This is because the regeneration process works more efficiently at higher temperatures of the hot water. Consequently, the adsorption material uptake increases substantially with an increase in the

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

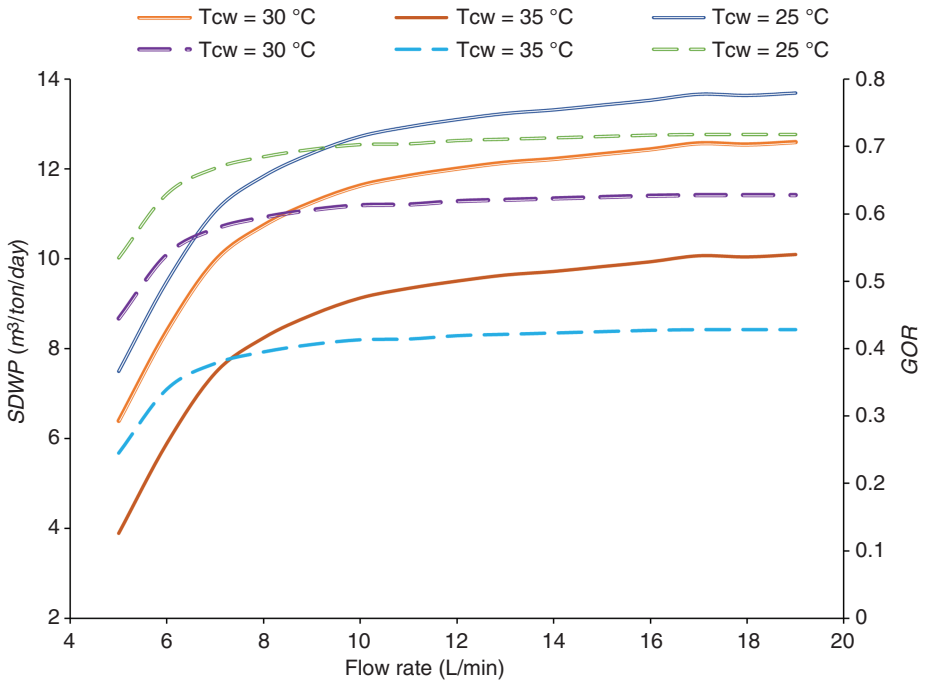


Fig. 9: Variation in the flow rate with different cooling water temperatures in the SDWP and GOR of the AD cycle.

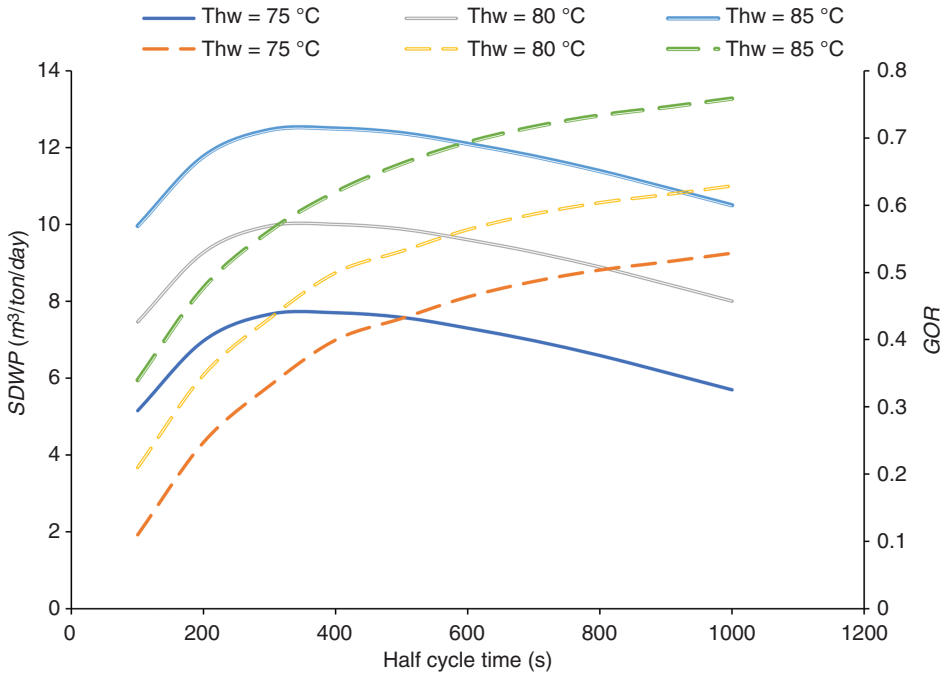


Fig. 10: Effect of the half-cycle time with different hot water temperatures on the SDWP and GOR of the AD cycle.

temperature of the hot water, as a result of its effect on the relative pressure and temperature. It is worth noting that when the temperature of the hot water increases from 65°C to 85°C, the SDWP increases substantially from 7.1 to 12.5 m^3/ton/day while the GOR increases from 0.39 to 0.62 under fixed parameters such as the half-cycle time (400 s), cooling water temperature (30°C) and water flow rate (12 L/min).

In Fig. 12, the shares of the exergy destruction rates per the different components of the S-CO_2 cycle and the AD cycle are shown. It can be clearly noticed in Fig. 12a that most of the exergy destruction occurs in the recuperator, which accounts for 54.7% of the total exergy destroyed within the S-CO_2 power cycle. The heat exchanger and boiler come next in terms of exergy destruction. On the other hand, Fig. 12b shows that the major exergy destruction rates (~91% of total) are observed at the evaporator and condenser. One apparent reason for this behaviour is the phase change resulting in higher irreversibility losses. These losses contribute to increased exergy destruction.

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

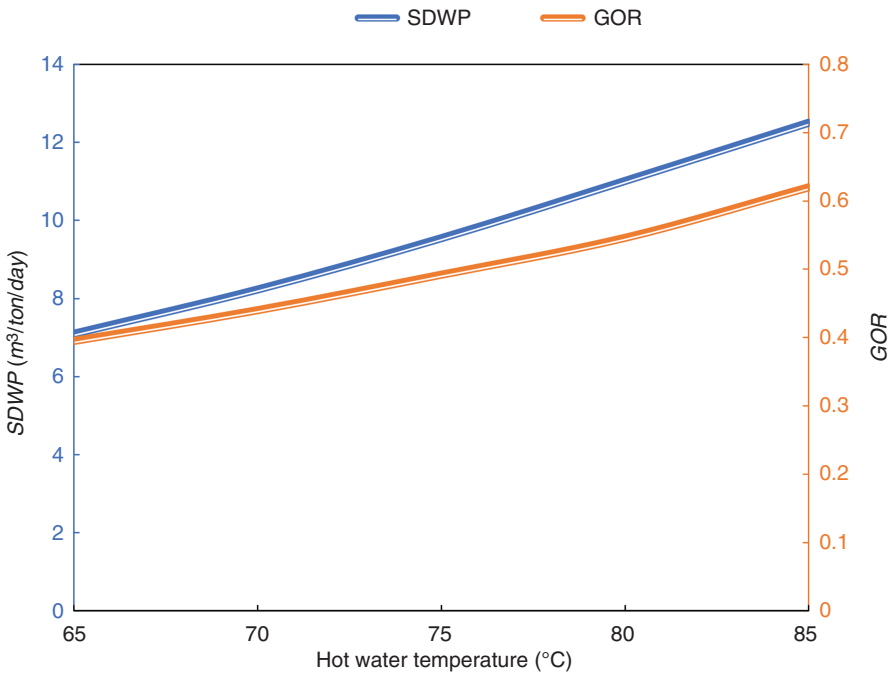


Fig. 11: Effect of hot water temperatures on the SDWP and GOR of the AD cycle.

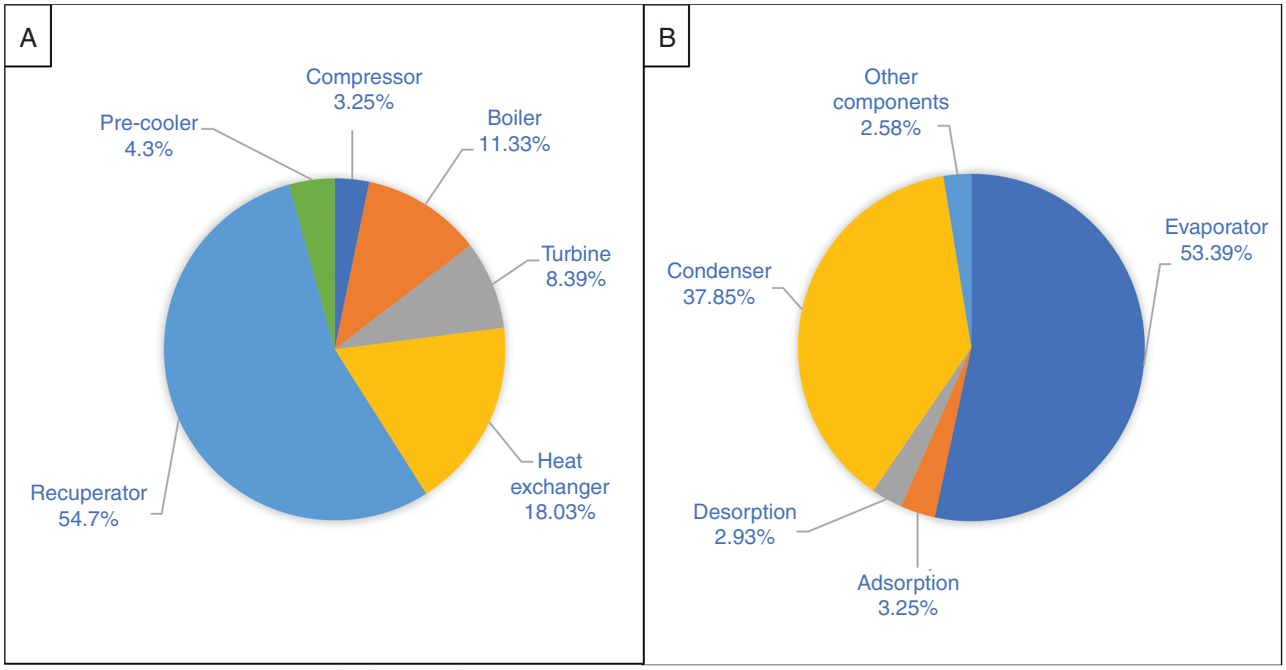


Fig. 12: The shares of the exergy destruction rate per component of the (a) S-CO_2 power cycle and (b) AD cycle.

In the condenser, the water vapour at ~85°C condenses by heat rejection to ambient and the exergy destruction is higher due to this phase change.

5 Concluding remarks

The thermodynamic analysis was performed on a novel solar-based supercritical CO_2 power and low-temperature AD system for power, heating and potable water. The effectiveness of an integrated system was evaluated using a mathematical model. Different varying parameters of the solar irradiation intensity, inlet temperature of the collector, turbine inlet pressure, water flow rate, half-cycle time, and cooling and hot water temperatures were examined to ascertain their influence on the thermal and exergy efficiency, GOR and SDWP. The main conclusions of this study are summarized as follows:

- The thermal efficiency of the PTC decreases with an increase in the inlet temperature of the HTF for the five selected mass flow rates of the collector. The thermal efficiency of the PTC is

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

highest for a mass flow rate of 0.3 kg/s and varies from 69.5% to 45.12% with increasing HTF input temperature.
- When the inlet temperature of the PTC and the intensity of the solar irradiation intensity (G_b) are fixed at 327°C and 800 W/m^2, the thermal and exergy efficiency of the PTC is determined as 67.2% and 41.2%, respectively.
- The thermal efficiency of the electrical and heating are 24.68% and 9.85% under fixed operating conditions.
- The variation in the thermal efficiency of the S-CO_2 power cycle is higher and ranges from 26.12% to 44.31% at 1000 W/m^2 with an increase in the turbine inlet pressure from 11 000 to 15 000 kPa.
- The highest SDWP achieved is 12.5 m^3/ton/day and 0.62 GOR at fixed operating conditions of the cooling water temperature (30°C), hot water temperature (85°C), water flow rate (12 L/min) and half-cycle time (400 s).
- The exergy analysis performed on the system reveals that the recuperator in the S-CO_2 cycle is responsible for 54.7% of the exergy dissipation, while in the AD cycle, 91% of the total exergy dissipation occurs in the evaporator and condenser.
- The development of such systems appears to be capable of meeting the energy requirements (power, heating and freshwater) of commercial units and industries in a sustainable and eco-friendly manner.
- The economic evaluation of the proposed solar-based integrated system will be carried out in further studies to reveal the economic feasibility of the system. Furthermore, considering the intermittent nature of solar energy, it is advisable to explore various energy storage methods that are applicable to the present system for further research.

Nomenclature

A_c: collector area (m^2);
C_p: specific heat at constant pressure (kJ/kg K);
$\dot{E}_x$: rate of exergy (kW);
G_b: solar irradiation intensity (W/m^2);
H_{ads}: isosteric heat of adsorption (kJ/kg);
h: specific enthalpy (kJ/kg);
$\dot{m}$: flowing rate of mass (kg/s);
P: pressure (kPa);
$\dot{Q}$: rate of heat transfer (kW);
T: temperature (°C);
s: specific entropy (kJ/kg K);
UA: overall conductance (W/K);
$\dot{W}$: work output (kW). Subscripts:
ads: adsorption;
am: adsorbent material;
C: compressor;
cond: condenser;
chill: chilled water;
cw: cooling water;
D: destruction;
des: desorption;
hw: hot water;
evap: evaporator;
pc: pre-cooler;
T, turbine.

Acknowledgements

The Authors acknowledge the support provided by King Abdullah City for Atomic and Renewable Energy (K.A.CARE) under K.A.CARE-King Abdulaziz University Collaboration Program.

Conflict of interest statement

None declared.

Data Availability

No new data were generated or analysed in support of this research.

References

[1] US Energy Information Administration (EIA). *International Energy Outlook 2019*. https://doi.org/https://www.eia.gov/outlooks/archive/ieo19/ (19 June 2023, date last accessed).
[2] Hinze J, Rapp L, Nellis G, *et al.* Modeling and experimental testing of periodic flow regenerators for sCO$_2$ cycles. *Appl Therm Eng*, 2019, 147:789–803. https://doi.org/10.1016/j.applthermaleng.2018.10.110.
[3] Quéré CL, Moriarty R, Andrew RM, *et al.* Global carbon budget 2015. *Earth Syst Sci Data*, 2015, 7:349–396. https://doi.org/10.5194/essd-7-349-2015.
[4] Bellos E, Tzivanidis C, Antonopoulos KA, *et al.* The use of gas working fluids in parabolic trough collectors: an energetic and exergetic analysis. *Appl Therm Eng*, 2016, 109:1–14. https://doi.org/10.1016/j.applthermaleng.2016.08.043.
[5] Khan Y, Mishra RS. Thermo-economic analysis of the combined solar based pre-compression supercritical CO_2 cycle and organic Rankine cycle using ultra low GWP fluids. *Thermal Science and Engineering Progress*, 2021, 23:100925. https://doi.org/10.1016/j.tsep.2021.100925.
[6] Bashir AAA, Özbey M. Modelling and analysis of an 80-MW parabolic trough concentrated solar power plant in Sudan. *Clean Energy*, 2022, 6:512–527. https://doi.org/10.1093/ce/zkac032.
[7] Katardjiev I. A fractal concentrator of direct and indirect sunlight. *Clean Energy*, 2020, 4:305–315. https://doi.org/10.1093/ce/zkaa017.
[8] Grena R. Optical simulation of a parabolic solar trough collector. *Int J Sustain Energy*, 2010, 29:19–36.
[9] Price H, Lópfert E, Kearney D, *et al.* Advances in parabolic trough solar power technology. *J Sol Energy Eng*, 2002, 124:109–125.
[10] Binotti M, Zhu G, Gray A, *et al.* Geometric analysis of three-dimensional effects of parabolic trough collectors. *Sol Energy*, 2013, 88:88–96.
[11] Khan Y, Mishra RS. Parametric (exergy-energy) analysis of parabolic trough solar collector-driven combined partial heating supercritical CO_2 cycle and organic Rankine cycle. *Energy Sources Part A*, 2020, 42:1–27. https://doi.org/10.1080/15567036.2020.1788676.
[12] Aybar HS. Desalination system using waste heat of power plant. *Desalination*, 2004, 166:167–170.
[13] Lienhard JH, Antar MA, Bilton A, *et al.* Solar desalination. In: Prasad V, Jaluria Y, Zhang ZM (eds). *Annual Review of Heat Transfer*. New York: Begell House, Inc., 2012, 277–347.
[14] Khan Y, Mishra RS. Performance evaluation of solar based combined pre-compression supercritical CO_2 cycle and organic Rankine cycle. *Int J Green Energy*, 2021, 18:172–186. https://doi.org/10.1080/15435075.2020.1847115.

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

[15] Al-Sulaiman F, Atif M. Performance comparison of different supercritical carbon dioxide Brayton cycles integrated with a solar power tower. *Energy*, 2015, 82:61–71.
[16] Moisseytsev A, Sienicki JJ. *Performance Improvement Options for the Supercritical Carbon Dioxide Brayton Cycle (No. ANL-GenIV-103)*. Argonne, IL, USA: Argonne National Lab. (ANL), 2007.
[17] Mehos M, Turchi C, Vidal J, *et al. Concentrating Solar Power Gen 3 Demonstration Roadmap. NREL Report No. TP-5500-67464*. Golden, CO, USA: National Renewable Energy Laboratory, 2017.
[18] Ettouney H, Wilf M. Commercial desalination technologies. In: Micale G, Rizzuti L, Cipollina A (eds). *Seawater Desalination: Green Energy and Technology*. Berlin, Germany: Springer, 2009. https://doi.org/10.1007/978-3-642-01150-4_4.
[19] Farsi A, Rosen MA. Multi-objective optimization of a geothermal steam turbine combined with reverse osmosis and multi-effect desalination for sustainable freshwater production. *ASME J. Energy Resour Technol*, 2022, 144:052102. https://doi.org/10.1115/1.4053298.
[20] Shaaban S. Performance optimization of an integrated solar combined cycle power plant equipped with a brine circulation MSF desalination unit. *Energy Convers Manage*, 2019, 198:111794.
[21] Elsayed E, Raya AD, Mahmoud S, *et al.* CPO-27 (Ni), aluminium fumarate and MIL-101 (Cr) MOF materials for adsorption water desalination. *Desalination*, 2017, 406:25–36.
[22] Mezher T, Fath H, Abbas Z, *et al.* Techno-economic assessment and environmental impacts of desalination technologies. *Desalination*, 2011, 266:263–273.
[23] Askalany AA. Innovative mechanical vapor compression adsorption desalination (MVC-AD) system. *Appl Therm Eng*, 2016, 106:286–292.
[24] Alsaman AS, Askalany AA, Harby K, *et al.* Performance evaluation of a solar-driven adsorption desalination-cooling system. *Energy*, 2017, 128:196–207. https://doi.org/10.1016/j.energy.2017.04.010.
[25] Thu K, Kim Y-D, Amy G, *et al.* A hybrid multi-effect distillation and adsorption cycle. *Appl Energy*, 2013, 104:810–821.
[26] Ng KC, Thu K, Kim Y, *et al.* Adsorption desalination: an emerging low-cost thermal desalination method. *Desalination*, 2013, 308:161–179.
[27] Mitra S, Srinivasan K, Kumar P, *et al.* Solar driven adsorption desalination system. *Energy Procedia*, 2014, 49:2261–2269.
[28] Ng KC, Thu K, Saha BB, *et al.* Study on a waste heat-driven adsorption cooling cum desalination cycle. *Int J Refrigeration*, 2012, 35:685–693.
[29] Elsayed E, Al-Dadah R, Mahmoud S, *et al.* Experimental testing of aluminium fumarate MOF for adsorption desalination. *Desalination*, 2020, 475:114170. https://doi.org/10.1016/j.desal.2019.114170.
[30] Kim Y-D, Thu K, Masry M, *et al.* Water quality assessment of solar-assisted adsorption desalination cycle. *Desalination*, 2014, 344:144–151.
[31] Stuber M, Sullivan C, Kirk S, *et al.* Pilot demonstration of concentrated solar-powered desalination of subsurface agricultural drainage water and other brackish groundwater sources. *Desalination*, 2015, 355:186–196.
[32] Sharaf M, Nafey A, García-Rodríguez L. Thermo-economic analysis of solar thermal power cycles assisted MED-VC (multi effect distillation-vapor compression) desalination processes. *Energy*, 2011, 36:2753–2764.
[33] Ibarra M, Rovira A, Alarcon-Padilla D, *et al.* Performance of a 5 kWe solar-only organic Rankine unit coupled to a reverse osmosis plant. *Energy Procedia*, 2014, 49:2251–2260. https://doi.org/10.1016/j.egypro.2014.03.238.
[34] Casimiroa S, Ahmedb M, Cardosoa J, *et al.* Reverse osmosis powered by concentrating solar power (CSP): a case study for Trapani, Sicily. *Desalin Water Treat*, 2017, 61:183–195.
[35] Palenzuela P, Zaragoza G, Alarcon D, *et al.* Simulation and evaluation of the coupling of desalination units to parabolic-trough solar power plants in the Mediterranean region. *Desalination*, 2011, 281:379–387.
[36] Assareh E, Alirahmi SM, Ahmadi P. A sustainable model for the integration of solar and geothermal energy boosted with thermoelectric generators (TEGs) for electricity, cooling and desalination purpose. *Geothermics*, 2021, 92:102042. https://doi.org/10.1016/j.geothermics.2021.102042.
[37] Olwig R, Hirsch T, Sattler C, *et al.* Techno-economic analysis of combined concentrating solar power and desalination plant configurations in Israel and Jordan. *Desalin Water Treat*, 2012, 41:9–25.
[38] Demir M, Dincer I. Development of an integrated hybrid solar thermal power system with thermoelectric generator for desalination and power production. *Desalination*, 2017, 404:59–71.
[39] Wang G, Dong B, Chen Z. Design and behaviour estimate of a novel concentrated solar-driven power and desalination system using S–CO_2 Brayton cycle and MSF technology. *Renew Energy*, 2021, 176:555–564. https://doi.org/10.1016/j.renene.2021.05.091.
[40] Amirfakhraei A, Zarei T, Khorshidi J. Performance improvement of adsorption desalination system by applying mass and heat recovery processes. *Therm Sci Eng Prog*, 2020, 18:100516. https://doi.org/10.1016/j.tsep.2020.100516.
[41] Siddiqui MA, Khaliq A, Kumar R. Proposal and analysis of a novel cooling-power cogeneration system driven by the exhaust gas heat of HCCI engine fuelled by wet-ethanol. *Energy*, 2021, 232:120954. https://doi.org/10.1016/j.energy.2021.120954.
[42] CIEMAT. *EuroTrough: Development of a Low Cost European Parabolic Trough Collector: EuroTrough*. Final Report, Research funded in part by The European Commission in the framework of the Non-Nuclear Energy Programme JOULE III. Contract JOR3-CT98-0231, 2001. https://doi.org/https://cordis.europa.eu/docs/projects/files/JOR/JOR3980231/51762451-6_en.pdf (19 June 2023, date late accessed).
[43] Kearney DW. Parabolic trough collector overview: notes on a bit of history, development after Luz, and a recent surge in trough collector technology offerings. In: *Parabolic Trough Workshop 2007*, Golden, CO, USA, 8–9 March 2007.
[44] Geyer M, Lüpfert E, Osuna R, *et al.* EUROTROUGH: parabolic trough collector developed for cost efficient solar power generation. In: *11th SolarPACES International Symposium on Concentrated Solar Power and Chemical Energy Technologies*, Zurich, Switzerland, 4–6 September 2002.
[45] Bellos E, Tzivanidis C, Tsimpoukis D. Thermal, hydraulic and exergetic evaluation of a parabolic trough collector operating with thermal oil and molten salt based nanofluids. *Energy Convers Manag*, 2018, 156:388–402.
[46] Loni R, Askari Asli-ardeh E, Ghobadian B, *et al.* Thermodynamic analysis of a solar dish receiver using different nanofluids. *Energy*, 2017, 133:749–760. https://doi.org/10.1016/j.energy.2017.05.016.
[47] Petela R. Exergy analysis of solar radiation, (Chapter 2). In: Enteria N, Akbarzadeh A (eds). *Solar Thermal Sciences and Engineering Applications*. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group, 2013.

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

[48] Ghazy M, Askalany A, Kamel A, *et al.* Performance enhancement of adsorption cooling cycle by pyrolysis of Maxsorb III activated carbon with ammonium carbonate. *Int J Refrigeration*, 2021, 126:210–221. https://doi.org/10.1016/j.ijrefrig.2020.12.036.

[49] Ali ES, Asfahan HM, Sultan M, *et al.* A novel ejectors integration with two-stages adsorption desalination: away to scavenge the ambient energy. *Sustainable Energy Technol Assess*, 2021, 48:101658. https://doi.org/10.1016/j.seta.2021.101658.

[50] Ghazy M, Askalany AA, Ibrahim EMM, *et al.* Solar powered adsorption desalination system employing CPO-27(Ni). *J Energy Storage*, 2022, 53:105174. https://doi.org/10.1016/j.est.2022.105174.

[51] Muñoz-Anton J, Biencinto M, Zarza E, *et al.* Theoretical basis and experimental facility for parabolic trough collectors at high temperature using gas as heat transfer fluid. *Appl Energy*, 2014, 135:373–381.

[52] Manente G, Lazzaretto A. Innovative biomass to power conversion systems based on cascaded supercritical CO_2 Brayton cycles. *Biomass Bioenergy*, 2014, 69:155–168.

[53] Youssef PG, AL-Dadah RK, Mahmoud S, *et al.* Effect of evaporator and condenser temperatures on the performance of adsorption desalination cooling cycle. *Energy Procedia*, 2015, 75:1464–1469. https://doi.org/10.1016/j.egypro.2015.07.263.

[54] Thu K, Saha BB, Chakraborty A, *et al.* Study on an advanced adsorption desalination cycle with evaporator-condenser heat recovery circuit. *Int J Heat Mass Transf*, 2011, 54:43–51. https://doi.org/10.1016/j.ijheatmasstransfer.2010.09.065.

[55] Ali ES, Askalany AA, Harbi K, *et al.* Experimental adsorption water desalination system utilizing activated clay for low grade heat source applications. *J Energy Storage*, 2021, 43:103219. https://doi.org/10.1016/j.est.2021.103219.

[56] Dakkama HJ, Youssef PG, Al-Dadah RK, *et al.* Adsorption ice making and water desalination system using metal organic frameworks/water pair. *Energy Convers Manage*, 2017, 142:53–61. https://doi.org/10.1016/j.enconman.2017.03.036.

[57] Elsheniti MB, Rezk A, Shaaban M, *et al.* Performance of a solar adsorption cooling and desalination system using aluminum fumarate and silica gel. *Appl Therm Eng*, 2021, 194:117116. doi: https://doi.org/10.1016/j.applthermaleng.2021.117116.

Downloaded from https://academic.oup.com/ce/article/7/5/1110/7329122 by Delhi Technological University user on 07 December 2023

arXiv:2311.02587v1 [cond-mat.mtrl-sci] 5 Nov 2023

Dielectric behavior and impedance spectroscopy of Niobium substituted Lanthanum based orthovanadates at high temperatures

Ashok Kumar,[1,2] Vikas N. Thakur,[3,*] Ajay Kumar,[3,†] Vinod Singh,[1] Anita Dhaka,[4] and Rajendra S. Dhaka[3,‡]

[1] *Department of Applied Physics, Delhi Technological University, Delhi-110042, India*
[2] *Department of Physics, Atma Ram Sanatan Dharma College, University of Delhi, New Delhi-110021, India*
[3] *Department of Physics, Indian Institute of Technology Delhi, Hauz Khas, New Delhi-110016, India*
[4] *Department of Physics, Hindu College, University of Delhi, New Delhi-110007, India*

(Dated: November 7, 2023)

We present the detailed study of the temperature dependent dielectric properties, impedance spectroscopy and electrical conductivity of $LaV_{1-x}Nb_xO_4$ ($x = 0$–1) samples prepared by the solid-state reaction method. The dielectric constant (ϵ_r') increases (decreases) with increase in the temperature (frequency); while, the magnitude of ϵ_r' remains almost invariant (order of 10^4 at 100 Hz and 600°C) with x. The single phase $x = 0$ (monoclinic-monazite, $P2_1/n$) and $x = 1$ (monoclinic fergusonite, $I2/a$) samples show the lower values of loss factor [tanδ = (4–8)] as compared to the $x = 0$–0.8 samples [tanδ =(12–18)] having mixed tetragonal and monoclinic phases, which indicates the strong correlation between the crystal structure and the dielectric properties. The real part of impedance (Z') decreases with both temperature and frequency, and the observed weak relaxation remains almost unaltered with x. The imaginary part of impedance (Z'') shows strong relaxation peaks shifting towards higher temperatures with frequency, which is attributed to the effect of grains, grain boundaries, and electrodes in the samples. The activation energy of the relaxation process is estimated to be 0.8–1.0 eV for the $x = 0$–0.6 samples, and ≈1.4 eV for the $x = 0.8$; whereas the $x = 1$ sample shows two values (≈0.5 eV and ≈1.0 eV) in the higher and lower temperature range, respectively. Further, the change in total conductivity with the angular frequency, which found to be in the range of 10^{-3} to 10^{-5} S/m, is fitted using the Jonsher power law. The analysis suggests the overlapping large polaron tunneling (OLPT) model for all the samples, whereas the $x = 1$ sample exhibit a transition near 480°C and at higher temperatures it shows non-overlapping small polaron tunneling (NSPT) and quantum mechanical tunneling (QMT). This transition is corroborated by the tangent loss curves and may be associated with the change in structure.

I. INTRODUCTION

In last few decades, the orthovanadates with general formula RVO_4, where R refers to rare earth elements, have widely attracted the attention of scientific community due to their vast technological applications in catalysts, as luminescent materials, polarizers, and laser host materials [1]. In this family, the $LaVO_4$ is turned out to be particularly important in the recent time because of its use as an electrode material for Lithium-ion batteries (LIBs) as it shows large specific capacity as well as long cyclic life [2]. The $LaVO_4$ crystallises in tetragonal zircon type polymorphs with space group $I4_1/amd$ and monoclinic monazite type polymorphs with space group $P2_1/n$, following the trend of the rare-earth family [3]. However, it also deviates from the trend by thermally stabilizing in the monazite type structure and being metastable at ambient temperatures [3]. This is due to the higher oxygen coordination number in the monazite as compared to zircon type structure and largest ionic radius of La^{3+} in the lanthanide family [4–6]. A pattern of VO_4 tetrahedra with four distinct V–O bonds and RO_8 dodecahedra sharing edges alternately and connected in chains along the c-axis can be found in the zircon structure [7]. In the monazite structure, RO_9 polyhedra are linked to deformed VO_4 tetrahedra with four distinct V–O links, sharing their edges [8]. According to the refs. [9] and [10], the Zircon type $LaVO_4$ cannot be synthesized at ambient temperatures using the traditional solid state reaction technique; however, it can be produced and stabilized using the hydrothermal method.

The influence of substituting La by another rare earth elements has been significantly explored for the $LaVO_4$ family; however, the doping/substation at the V-site is still in its infancy, as evidenced by the fact that only a handful of studies have been conducted so far [3, 11, 12]. The rare-earth niobate $LaNbO_4$ on the other hand is well-known to show the temperature and chemical pressure (doping) induced structural transformation [13, 14]. The $LaNbO_4$ transforms from the monoclinic fergusonite (space group $I2/a$) structure at the low temperature to the tetragonal scheelite (space group $I4_1/a$) phase above ∼495°C [13]. Similarly, it also shows a structural transformation on replacing 25% of Nb^{5+} with V^{5+} [14]. Because of its intriguing characteristics, including proton conductivity [15], excellent dielectric, and high energy release using X–ray stimulation [16]and hence wide technological applications in sensors [17], contrast agents, waveguides, ferroelectrics [18], phosphors [19], laser crys-

* Present Address: Department of Physics, School of Physical Sciences, DIT University, Dehradun 248009, India
† Present Address: Ames National Laboratory, U.S. Department of Energy, Iowa State University, Ames, Iowa 50011, USA
‡ Corresponding author: rsdhaka@physics.iitd.ac.in

tals [20], scintillators [21], luminophores, LEDs, etc., the $LaNbO_4$ is getting a special attention of the scientific community from several decades. Therefore, we choose to replace V by Nb atoms in $LaV_{1-x}Nb_xO_4$, as Nb shares many properties with vanadium and is found right beneath it in the periodic chart, to examine the influence of B-site atoms on the high temperature structural transition and hence the physical properties of these compounds. Though, the vanadium costs have historically lagged behind those of niobium, but it has lately increased by about 300%, which has created strong economic incentives to transition to niobium. The tetrahedrally coordinated niobium has bigger ionic radius (0.48 Å), and is isoelectronic to the vanadium ion (V^{+5}) [22] and hence most suitable to exert the chemical pressure in the crystal lattice. Further, the nontoxic nature of Nb make it suitable candidate for the industrial applications. Moreover, the La-Nb based perovskite oxides have been also extensively studied due to their important fundamental as well as functional properties [23–29].

Recently, we report the vibrational, structural as well as electronic properties of the $LaV_{1-x}Nb_xO_4$ (x =0–1) samples using the x-ray diffraction (XRD), Raman and x-ray photoemission spectroscopy [3]. The $x = 0$ and 1 samples were found to crystallize in the single phase monoclinic-monazite ($P2_1/n$) and monoclinic fergusonite ($I2/a$) type structures at the room temperature, respectively; whereas the monoclinic-monazite ($P2_1/n$) and tetragonal scheelite ($I4_1/a$) phases coexist in the x = 0.2–0.8 samples and phase fraction of later increases with the Nb substitution [3]. The fact that the $LaNbO_4$ possesses the tetragonal structure when measured at room temperature by substituting about 20% of Nb^{5+} concentration at the V^{5+} site is an important finding, which makes a variety of room-temperature applications possible, as this transformation plays a significant role in regulating the protonic conductivity of $LaNbO_4$ [30]. Moreover, for certain compositions of $LaV_{1-x}Nb_xO_4$, the transition temperature can occur close to room temperature [14]. At x=0.75, for instance, temperature-dependent XRD measurements indicate that $LaV_{1-x}Nb_xO_4$ should possess a tetragonal structure when measured at room temperature, with its transition at around 250 K. Nevertheless, the XRD pattern below this transition also shows minor residual intensity in the form of broadened lines, which may be the effect of precursors [14]. Similarly, the XRD patterns of x=0.8 samples also exhibit broad peaks, likely due to the same effect [14]. Therefore, it is vital to investigate the high temperature phase transition and their influence on the dielectric and ac impedance spectroscopy with Nb substitution in $LaV_{1-x}Nb_xO_4$ samples. The impedance spectroscopy is widely used to study the electrical properties of materials to understand their various possible applications such as solid oxide fuel cells (SOFCs). One of the advantages of impedance spectroscopy is its ability to provide information about the properties of a material or system over a wide frequency range, typically from a few Hertz to several mega Hertz. This can be particularly useful for studying materials that exhibit complex electrical behavior, such as those with multiple structural phases or with a distribution of relaxation times [31–34].

Therefore, in this article, our objectives are to comprehensively investigate the temperature dependent conductivity, impedance spectroscopy, and dielectric properties of $LaV_{1-x}Nb_xO_4$ (x =0–1) samples. These measurements are made of the dielectric constant and tangent loss at temperatures ranging from 25°C to 600°C and in a wide range of frequency (f) from 100 Hz to 1 MHz. We demonstrate that the dielectric constant increases with temperature and decreases in magnitude with frequency. The magnitude of the dielectric constant is not significantly changed by the addition of Nb^{5+} in place of V^{5+}, and anomalous behaviour is observed from $x = 0.2$ to 0.8 due to their mixed states, whereas the $x = 0$ sample shows no peak and the $x = 1$ sample shows a weak relaxation in higher temperature range. For the $x = 0$–0.8 samples, the tangent loss displays the peaks in the higher temperature range above 360°C, whereas the x = 1 displays a weak relaxation peak, shifting towards higher temperatures. The real component of the impedance (Z') decreases with temperature and frequency, and a weak relaxation is observed in all the samples, which remains almost invariant with x. The imaginary component of impedance (Z'') exhibits prominent relaxation peaks that shift towards higher temperatures with frequency, which further illuminating the influence of the grain, grain boundary, and electrode in the samples. In the Z''–f plots, the behaviour from $x = 0$ to 0.6 is identical, but the $x = 0.8$ and 1 samples exhibit a relaxation in the lower frequency area that exhibits both the grain effect and the grain boundary effect. In order to look at the lower frequency components, such as the grain boundary and electrode influence, the modulous spectroscopy is also used. Additionally, the total conductivity is measured with varied frequency in the 360°C–560°C temperature range and fitted with the Jonsher power law. The $x = 0.8$ sample shows the lowest conductivity, which is related to the largest grain size and highest Nb concentration in the coexisting phases of these samples. Interestingly, we find the overlapping large polaron tunneling (OLPT) behavior for all the samples, whereas the $x = 1$ sample undergoes non-overlapping small polaron tunneling (NSPT) and quantum mechanical tunneling (QMT) behavior at higher temperatures.

II. EXPERIMENTAL

We use the conventional solid state reaction method to prepare the $LaV_{1-x}Nb_xO_4$ ($x = 0$–1, Δx=0.2) bulk samples. The details of the synthesis process, and their structural and electronic properties can be found in our earlier report [3]. The prepared samples were cold pressed in the form of circular pellets of 10 mm diameter and 1-2 mm thickness at 14 MPa before the final annealing at 1250°C.

The pellets were then polished and painted with sliver paste on both the sides for the dielectric measurements. The temperature dependent dielectric measurements and ac impedance spectroscopy were then performed on all the samples using HIOKI 3536 LCR meter from 25°C to 600°C in the large frequency range from 100 Hz to 1 MHz at 1 V excitation voltage, under the vacuum conditions of $\sim 3\times10^{-2}$ mbar.

The complex dielectric constant (ϵ_r), tangent loss ($\tan\delta$), complex electric modulus (M), and ac conductivity are calculated using the temperature dependent complex impedance (Z) and phase angle (θ) measured using the LCR meter. It involves applying a small, alternating voltage to a sample and measuring the resulting current response over a range of frequencies. By analyzing the relationship between the current and voltage, the impedance of the sample can be determined. We use the following formulas to calculate various electrical parameters:

$$Z = Z' + iZ'' \tag{1}$$

$$M = i\omega C_0 Z = M' + iM'', \tag{2}$$

where, $Z'=|Z|cos\theta$ and $Z''=|Z|sin\theta$ are real and imaginary parts of complex impedance, M' and M'' are real and imaginary parts of complex electric modulus, $\omega=2\pi f$ is the angular frequency, C_0 is the capacitance in the vacuum, defined as $C_0=A\epsilon_0/d$, where A is the area of electrodes and d is the separation between them.

Using equations (1) and (2), the M' and M'' can be expressed as:

$$M' = \omega C_0 Z''; M'' = \omega C_0 Z' \tag{3}$$

Here, the dielectric constant can be written in terms of the inverse of electric modulus, i.e.,

$$\epsilon_r = \epsilon_r' + i\epsilon_r'' = 1/M \tag{4}$$

where the ϵ_r' and ϵ_r'' are real and imaginary parts of ϵ_r, respectively.

Using equations (2), (3) and (4), we get

$$\epsilon_r' = Z''/\omega C_0 Z^2; \epsilon_r'' = Z'/\omega C_0 Z^2 \tag{5}$$

We utilize the above equations to compute the ϵ_r', Z', Z'', and M'' parameters, as discussed in detail below.

III. RESULTS AND DISCUSSION

The temperature dependent real part of dielectric constant (ϵ_r') is presented in Figs. 1(a–f) for $LaV_{1-x}Nb_xO_4$ (x =0–1) samples at the selected frequencies. The dielectric constant of a material determines the amount of electric charge that can be stored in a capacitor made from the material, as well as the strength of the electric field that can be sustained before the material breaks down. For the $x = 0$ sample, no peak is observed in the measured temperature range and the ϵ_r' decreases with increase in the frequency [see Fig. 1(a)]. The ϵ_r' remains almost temperature independent up to around 300°C and then increases abruptly with further increase in the temperature. The decrease in the ϵ_r' with frequency is because of the reduction of space charge polarization [35–38]. The presence of space charge polarization at lower frequency creates a potential barrier which leads to the building-up of charges at the grain boundary, resulting the higher values of ϵ_r' [35]. It is conventional to use the thermal energy to demonstrate the relation between the dielectric constant and the change in the temperature [39]. The mobility of charge carriers and, therefore, their rate of hopping are improved at higher temperatures by the additional thermal energy, which is insufficient at the lower temperatures, causing the observed enhancement in the ϵ_r' with temperature. The dielectric polarization increases with increase in the temperature, which enhances the ϵ_r' at higher temperatures up to a certain limit, called 'phase transition temperature' and at this temperature the polarization decreases to almost zero, there-

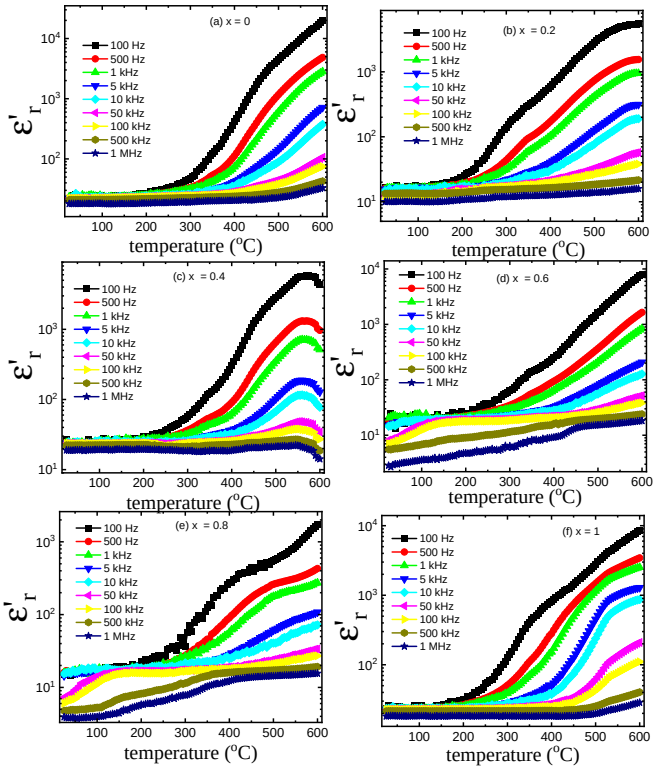


FIG. 1. (a–f) The temperature dependent real part of dielectric constant (ϵ_r') for $LaV_{1-x}Nb_xO_4$ (x = 0–1) samples from 25°C–600°C in the frequency range of 100 Hz–1 MHz.

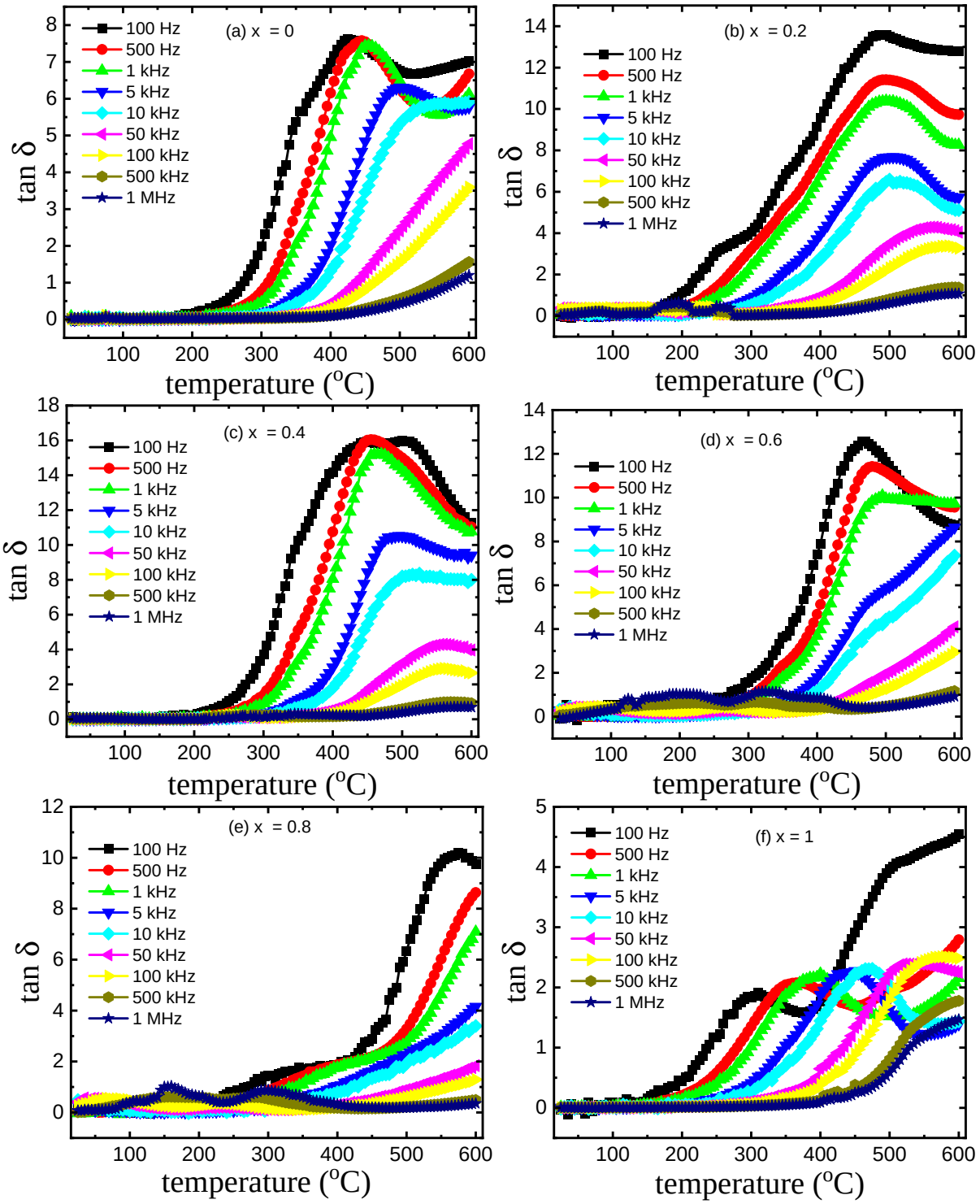


FIG. 2. (a–f) The dielectric loss (tan δ) for $LaV_{1-x}Nb_xO_4$ (x = 0–1) samples plotted as a function of temperature at the selected frequencies.

fore the ϵ_r' starts decreasing with temperature [40, 41]. However, no phase transition has been observed for the x =0 sample in the measured temperature range (up to 600°C). Similarly, for the $x = 0.2$ sample, no peak is observed; however, the curves start saturating near 600°C, as shown in Fig. 1(b). The substitution of Nb^{5+} may increases the conductivity of the sample, which led to decrease in the phase transition temperature and hence

saturation in the dielectric constant. Moreover, a peak is observed near 550°C with further increase in the Nb substitution for the $x = 0.4$ sample [see Fig. 1(c)], the position of which remains almost invariant with increase in the frequency. Interestingly, for the x =0.6 and 0.8 samples, an additional dielectric relaxation peak is observed at significantly lower temperatures (200–300°C), which shifts to the higher values with increase in the excitation frequencies [see Figs. 1(d, e)]. The increase in the tetragonal scheelite ($I4_1/a$) phases with the Nb substitution is expected to give rise to this additional relaxation peak in these samples [3]. Moreover, the $x = 1$ sample shows a similar behavior to the $x = 0$ [except a weak relaxation around 500°C; see Fig. 1(f)], which can be attributed to the presence of single phase in both of these samples [3].

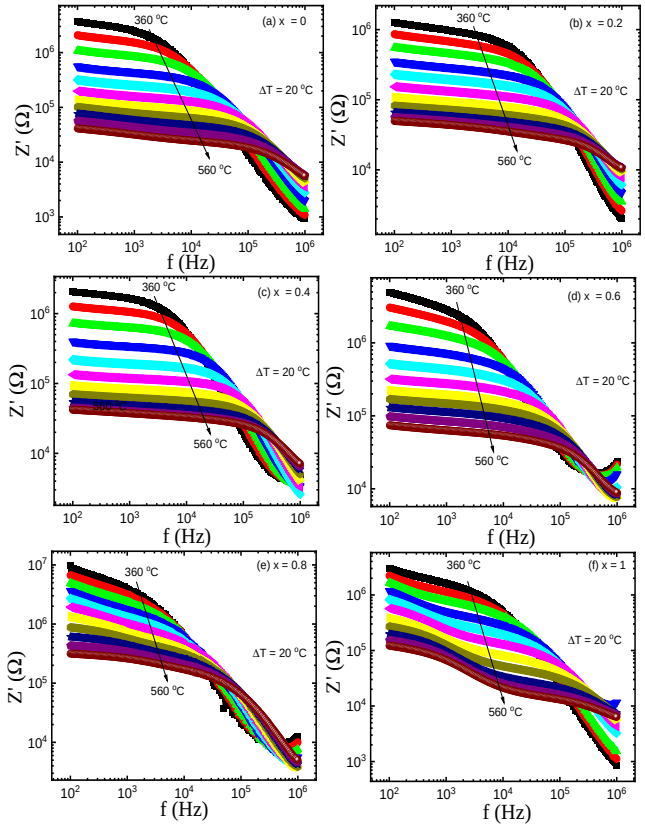


FIG. 3. (a–f) The real part of impedance (Z') of $LaV_{1-x}Nb_xO_4$ (x = 0–1) samples plotted as a function of frequency from 100 Hz to 1 MHz at the selected temperature from 360°C–560°C.

To gain a deeper understanding of the material's behavior, we perform detailed analysis of the temperature-dependent tangent dielectric loss, represented as tan $\delta = \epsilon_r''/\epsilon_r'$, for all the samples, as shown in Figs. 2(a–f). Here, we investigate how different samples respond to changes in temperature and excitation frequency by analyzing their tangent dielectric loss, which is a measure of how the material dissipates energy when subjected to an alternating electric field. Note that all the samples in this study exhibit peaks in their tan δ curves, which are a significant indicator of how the material responds to changes with the variation in temperature and excitation frequency. The location of the peak in tanδ shifts to higher temperatures as the excitation frequency increases. This shift implies that with the application of higher frequencies, the material's response to temperature changes at higher temperatures. It is an important behavior to note because it indicates a thermally triggered relaxation process [42]. The magnitude of the peak in tanδ decreases as the excitation frequency increases for most of the samples. This behavior suggests that at higher frequencies, the material's dielectric loss is reduced. However, there is an exception with the sample where $x = 1$, where the height of the peak in tanδ actually increases with the frequency until it reaches around 100 kHz. After this point, it starts to decrease as the frequency increases further [see Fig. 2(f)].

It should be noted that the increase in tanδ at higher temperatures is primarily due to the direct current (dc) conductivity. In other words, the behavior of the material at high temperatures is influenced significantly by its ability to conduct the electric current. Interestingly, the tanδ values increases with Nb substitution up to $x = 0.4$ sample, and then start decreasing at higher Nb concentration. It is observed that the samples with single phases (with $x = 0$ and $x = 1$) have lower dielectric loss values compared to those with mixed phases (with $x = 0.2$–0.8). This indicates that the presence of mixed phases affects the dielectric properties of the material. For the samples with mixed phases (specifically, $x = 0.2$–0.8), the interaction between the different phases results in anomalous behavior in the tanδ curves. This anomalous behavior is observed in higher frequencies and lower temperature regimes [3]. It suggests that the combination of different phases introduces complex interactions that affect how the material responds to changes in temperature and excitation frequency. The observed shifts in peak positions, frequency-dependent magnitudes, and differences between single-phase and mixed-phase samples provide insights into the material's electrical properties and its complex behavior under varying conditions. These findings are important for understanding the material's suitability for specific applications and may have implications for the design of electronic devices. From the above analysis and discussion we reveal that the samples with mixed phases ($x = 0.2$–0.8) exhibit anomalous behavior in their ϵ_r' and tanδ characteristics. The presence of mixed phases introduces complex interactions that govern how the electric dipoles respond to the change in temperature and excitation frequency.

Now, we further delved insight into the investigation by analyzing the impedance (Z) and modulus (M) spectra of all the samples. The aim is to unravel how the grains (individual crystal structures) and grain boundaries (interfaces between these crystal structures) contribute to the observed relaxation processes in the ϵ_r'(T) (real part of dielectric permittivity as a function of temperature)

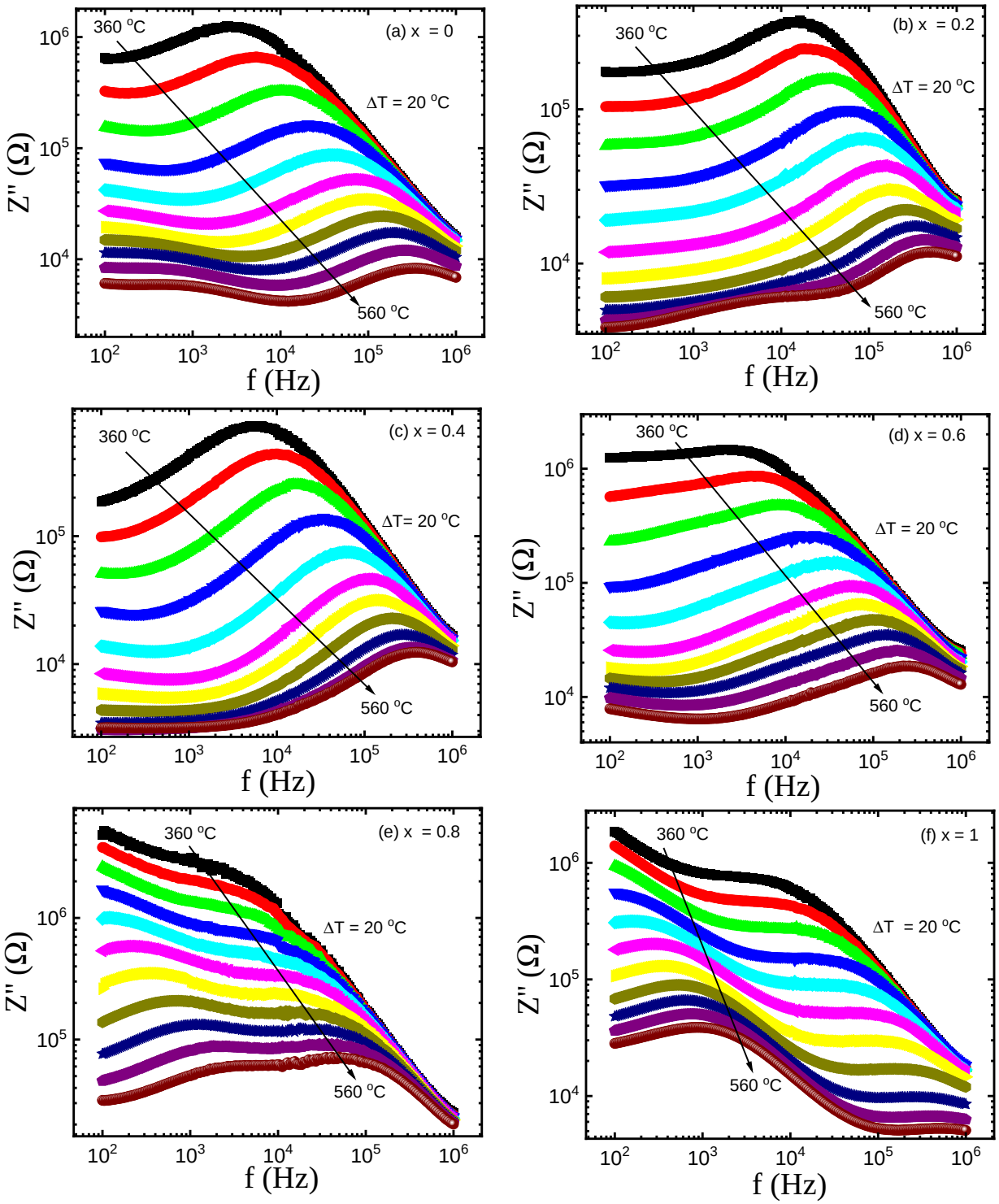


FIG. 4. (a–f) The imaginary part of impedance (Z'') of $LaV_{1-x}Nb_xO_4$ (x = 0–1) samples plotted as a function of frequency from from 100 Hz to 1 MHz at the selected temperatures from 360°C–560°C.

and tanδ(T) (tangent of dielectric loss as a function of temperature) spectra. From the temperature-dependent data of ϵ_r' and $tan\delta$ in the temperature range of 25 to 600°C for all the samples in Figures 1 and 2, respectively, we observed that the significant and rapid increase in the dielectric constant is within the range of 360 °C to 560 °C. Also, this temperature range exhibited the most pronounced peaks in the tangent loss for all the samples.

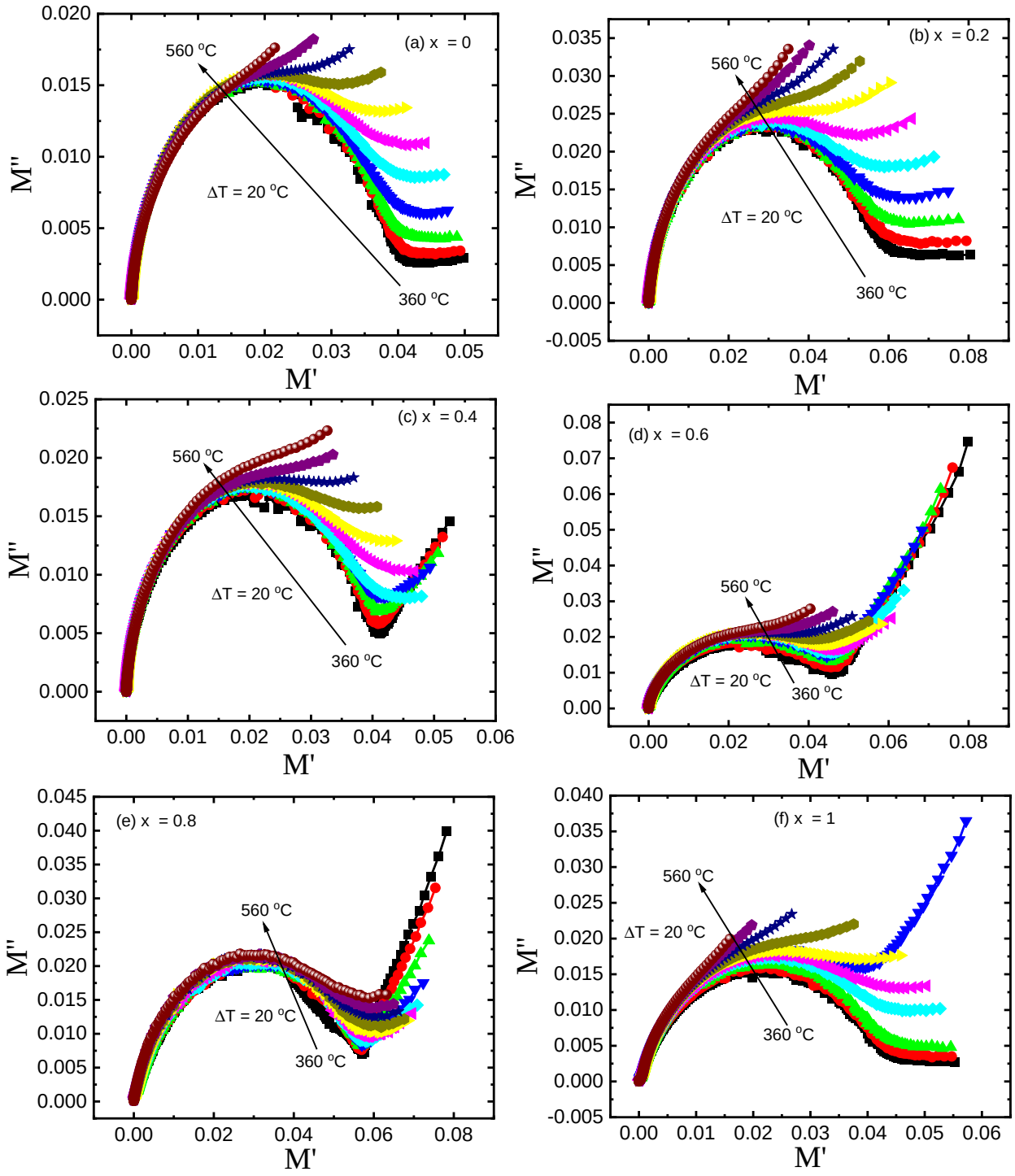


FIG. 5. (a-f) The imaginary (M'') versus real (M') parts of electric modulus as a function of frequency from 100 Hz to 1 MHz at various temperatures from 360 °C to 560 °C for the $LaV_{1-x}Nb_xO_4$ (x = 0–1) samples.

Therefore, we focus our analysis solely in this specific temperature range (360 °C to 560 °C) to investigate parameters such as impedance, electric modulus, and conductivity. This approach allow us to gain insights into the behavior of the grain and the effects on grain boundaries resulting from Nb doping in $LaV_{1-x}Nb_xO_4$. These findings provide valuable information about the electrical properties of the material and how they relate to its

microstructure, which is crucial for understanding its potential applications. Figs. 3(a–f) display the real part of impedance (Z') as a function of frequency for all the samples at selected temperatures. The Z' is a measure of the opposition that a material presents to the flow of alternating current. The key observation is that the order of Z' remains nearly constant regardless of the Nb substitution in the samples. However, the value of Z' decreases with increasing temperature for all the samples, which is indicative of the semiconducting or insulating nature of these materials. In other words, as the temperature rises, the ability of the material to conduct electric current decreases. The impedance behavior can be divided into two regions: (1) frequency-independent region and (2) frequency-dependent region. The first region is associated with *dc* conductivity, which means that the charge carriers (electrons or ions) can move over extended distances successfully through a hopping mechanism. In this process, the neighboring charge carriers relax to the location of the moving carrier. This mechanism leads to the *dc* conductivity at low frequencies. The second region is linked to the alternating current (*ac*) conductivity. Here, the charge carriers have limited or localized movement and fail to hop effectively, instead, they return to their original positions. This mechanism results in dispersive *ac* conductivity at higher frequencies. We find that the *ac* conductivity dominates above a critical frequency and this critical frequency is determined by the sample's temperature and remains nearly constant across different values of the Nb concentration (x). In other words, the transition from *dc* to *ac* conductivity occurs at a specific frequency that depends on temperature, but found to be independent of x [see Figs. 3(a–f)].

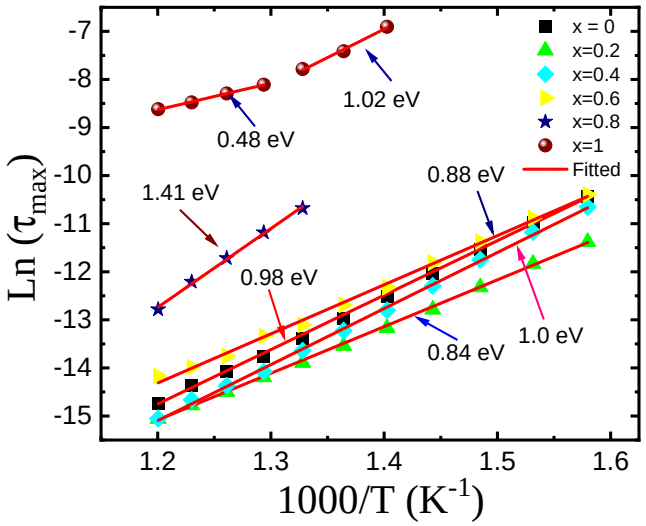


FIG. 6. The variation in the relaxation time with inverse of temperature for $LaV_{1-x}Nb_xO_4$ ($x = 0$–1) samples. The solid red lines represent the best fit using the Arrhenius equation.

Further, the imaginary part of the impedance (Z'') for all the samples is investigated to gain more understanding into various contributing factors, including the effects of electrodes, individual grains, and grain boundaries. Figs. 4(a–f) present the imaginary part of the impedance (Z'') for all the samples, which is a measure of the phase difference between the applied voltage and the resulting current. In the context of this study, it provides information about the relaxation processes within the material. In the Z'' spectra, relaxation maxima are observed, which are characteristic peaks at specific frequencies. These peaks offer details about the contribution of various factors: grain (bulk), grain boundary and ceramic-electrode interface. The samples with x values ranging from 0 to 0.6 exhibit well-defined relaxation peaks in the higher frequency range [see Figs. 4(a–d)]. These peaks are attributed to the grain or bulk effect within the samples. In other words, they represent how the individual crystalline grains in the material respond to the applied electrical field. In the samples with $x = 0$–0.6, a weaker relaxation is also observed in the low-frequency range at higher temperatures. This lower-frequency relaxation is associated with the grain boundary mechanism, which involves how charge carriers move and interact at the interfaces between different crystalline grains. Additionally, the electrode interfaces between the material and the measuring electrodes can contribute to relaxation peaks in the low-frequency region [43–45]. As the temperature increases, the relaxation maxima in Z'' tend to migrate

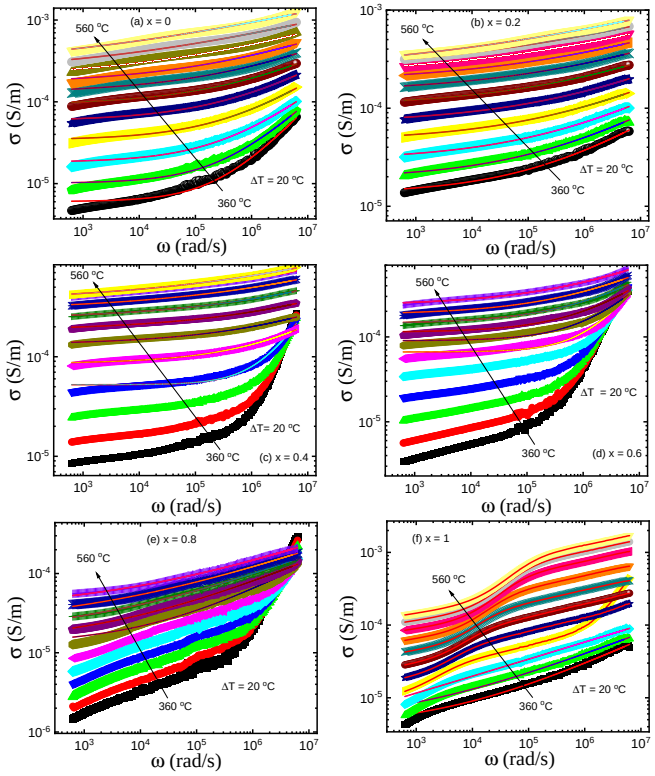


FIG. 7. (a–f) The angular frequency dependent *ac* conductivity for the $LaV_{1-x}Nb_xO_4$ ($x = 0$–1) samples at various temperatures between 360°C and 560°C.

toward higher frequencies, and their intensity weakens. The temperature-dependent behavior suggests that the relaxation processes within the material are influenced by temperature, with relaxation occurring at higher frequencies as the temperature increases. For the samples with higher Nb substitution ($x = 0.8$ and 1), stronger relaxation peaks are observed in the lower-frequency regime compared to the samples with x values of 0 to 0.6 as well as at higher frequencies. This suggests that the substitution of Nb^{5+} at the V^{5+} site triggers grain boundary relaxation mechanisms in these samples, especially when x is equal to and greater than 0.8.

It is important to plot the Nyquist modulus to comprehend the process underlying the relaxation behavior in these samples [46]. The modulus spectra [imaginary (M'') versus real (M')] that shows the behavior of different semicircular arcs, which are illustrated to distinguish each relaxation mechanism, as shown in Figs. 5(a–f) for the samples. One semicircular arc can be seen in the real (M') versus imaginary (M'') modulus spectra for all the samples, which indicates that one of the relaxation processes has been severely inhibited by the surface polarization effect. When a small arc is formed at the end of the M' axis, it indicates a minor grain contribution, and when the temperature increases the arc eventually vanishes. This clearly demonstrate how the grain contribution finally diminishes with increase in the temperature. The M' versus M'' modulus spectra show that the electrode polarization at high temperature, extracted by the impedance spectroscopy, is drastically inhibited in this condition [47, 48]. At the higher frequencies, the grain boundary and electrode ceramic effects can also be seen, which increases with Nb^{5+} concentration and are not visible in impedance plots. In the samples $x = 0.4$, 0.6 and 0.8, electrode ceramic interface effects are more prominent which may be due to occurrence of mixed phases because of the significant contribution of both Nb^{5+} and V^{5+} elements.

In order to understand this mechanism, we use the relationship $2\pi\tau f_{max}$=1, where f_{max} represent the frequency at the peak in Z'' to determine the dielectric relaxation time (τ). In Fig. 6, we plot the τ as a function of inverse of temperature for all the samples. We find that the relaxation time decreases when the temperature increases and it follow the Arrhenius relation for all the samples:

$$\tau = \tau_0 \exp^{-E_a/k_B T} \tag{6}$$

where E_a is the activation energy of the charge carriers, k_B is the Boltzmann's constant, and τ_0 represents the pre-exponential factor in the relaxation time. The activation energy is typically associated with breaking and forming the chemical bonds, i.e., the minimum amount of energy required for a given relaxation process to activate. The estimated values of activation energy for the $x = 0$–0.8 samples are found to be 0.98 eV, 0.84 eV, 1.0 eV, 0.88 eV, and 1.41 eV, respectively. The $x = 0.8$ sample shows maximum activation energy among all, which may be associated with the highest Nb concentration for the coexisting monoclinic-monazite ($P2_1/n$) and tetragonal scheelite ($I4_1/a$) phases, and hence largest average particle size of this samples, observed from the scanning electron microscopy (SEM) [3]. In case of the $x = 1$ sample, a change in the activation energy is observed near 460°C from 1.02 eV to 0.48 eV, which can also be re-

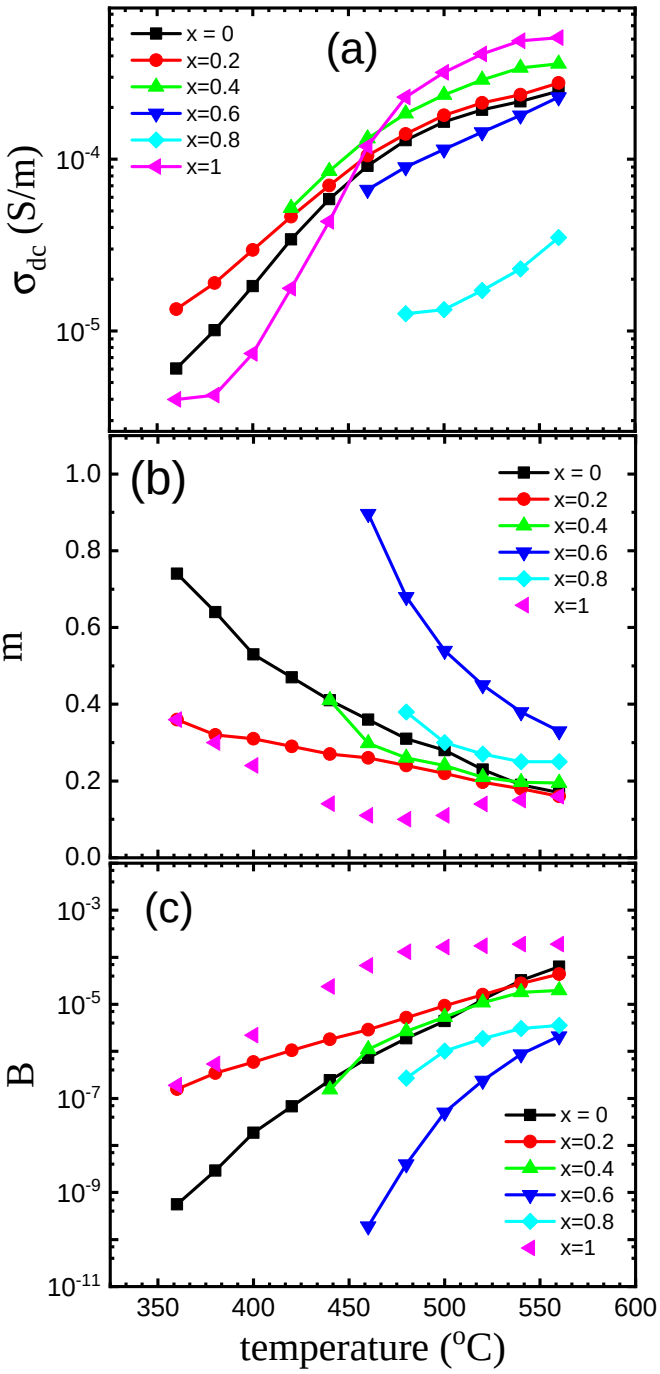


FIG. 8. The best fit parameters of the *ac* conductivity using Jonscher power law, (a) the *dc* conductivity (b) the exponent m, and (c) the constant B as a function of temperature for the $LaV_{1-x}Nb_xO_4$ ($x = 0$–1) samples.

flected in Fig. 2(f), where an additional relaxation peak is observed in this sample. The transformation of the $x = 1$ sample from the monoclinic fergusonite (space group $I2/a$) structure at the low temperature to the tetragonal scheelite (space group $I4_1/a$) phase at the higher temperature, and hence availability of the large thermal energy of the charge carriers results in this observed reduction in the activation energy above a threshold temperature of around 460°C [13].

In Figs. 7(a–f), we present the total conductivity (σ) of the samples as a function of angular frequency (ω) in the temperature range of 360°C–560°C. The solid lines represent the best fit of the experimental data using Jonscher Power Law (JPL) equation:

$$\sigma = \sigma_{dc} + B\omega^m \tag{7}$$

where σ_{dc} represents the dc conductivity, $B\omega^m$ represents the ac conductivity (σ_{ac}), where both the B and m are the measure of strength of polarization and degrees of interaction of charge carriers, respectively. Both of these quantities are suppose to be dependent on temperature. We find that the σ values are strongly frequency dependent and gradually increases at the higher frequencies. Further, the conductivity increases monotonically with increase in temperature for all the samples. Note that at lower temperatures the conductivity increases towards higher frequency; however, it shift towards the saturation at higher temperatures, which indicates a threshold in the conductance mechanism in the materials. At the lower temperature and high frequency region, the conductivity becomes frequency dependent and this frequency dependence increases from $x = 0$ to 0.8, whereas for the $x = 1$ sample a weak relaxation in the conductivity is observed in the low frequency region, which is more clearly visible above 400°C, which may be attributed to the presence of single monoclinic ($I2/a$) phase in this sample [3]. The total conductivity for the $x = 0$, 0.2 and 1 follows Jonscher Power Law equation in the temperature range of 360°C–560°C, whereas for the $x = 0.4$, 0.6 and 0.8 sample it follows the Jonscher Power Law equation above 420°C, 460°C and 480°C temperatures, respectively.

Finally, in Fig. 8(a) we show the variation of the σ_{dc} with the temperature where we observe that the σ_{dc} increases non-linearly in the temperature range of 360°C–560°C, which shows the semiconducting/insulating nature of the materials. There have been various models reported to understand the conductivity of different complex oxide samples on the basis of exponent m as well as the polaron tunneling. It is important to note that a local lattice deformation can result due to the addition of charge transporters to a site in a solid. Further, a polaron can be created by the arrangement of electrons as well as the local distortion. In the current study, we find that the m values decreases with temperature for all the samples, except for $x = 1$, which shows the decreasing valley near 480°C and the subsequent increase in value indicates the presence of the overlapping large polaron tunneling (OLPT) conduction mechanism [49, 50]. In this mechanism, the polaron clouds extend considerably beyond the typical interatomic distances [51]. As the temperature increases, the m values shown in Fig. 8(b) exhibits a linear ascent until it reaches approximately 540°C, elucidating the behavior of non-overlapping small polaron tunneling (NSPT) [52]. During this phase, the polarons are confined to the extent that their distortion clouds do not intersect. Consequently, the electron's tunnel between states near the Fermi level, and the tunneling energy remains unaffected by inter-site distances [51]. The m value eventually levels off at around 560°C, a phenomenon explicable by quantum mechanical tunneling (QMT) [53]. Notably, the temperature independence of m in this model distinguishes it [54]. This characteristic underscores that conduction occurs primarily through the tunneling of electrons between defect states near the Fermi level, without the generation of polarons. Similarly, the values of B shown in Fig. 8(c) increases nonlinearly with temperature for all the samples, except for the $x = 1$, which shows the saturation near 560°C, which means the polarization strength is also getting saturated at higher temperatures, which is found to be consistent with the dielectric behavior, as shown in Fig. 2.

IV. SUMMARY AND CONCLUSIONS

We measured the dielectric constant and tangent loss at temperatures ranging from 25°C to 600°C and a wide range of frequencies (100 Hz to 1 MHz) for $LaV_{1-x}Nb_xO_4$ ($x = 0$–1) samples. We found that the dielectric constant increased with temperatures, but decreased in magnitude with frequency having the value $\approx 10^4$ at 100 Hz and 600°C. The iso-electronic substitution of Nb^{5+} at V^{5+} site did not significantly affect the magnitude of the dielectric constant. The anomalous behavior in $\epsilon_r'(T)$ curves have been observed for the $x = 0.2$–0.8 samples due to their mixed structural phases, whereas no relaxation peak was observed for the $x = 0$ sample, and a weak relaxation was seen for the $x = 1$ sample at the higher temperatures. The mixed phase in the $x = 0$ to 0.8 samples show the significantly higher values of the tangent loss as compared to the single phase $x = 0$ and 1 samples. The real part of the impedance (Z') spectroscopy decreased with temperature and frequency, and a weak relaxation was observed in all the samples with negligible changes in magnitude with x. The imaginary part of the impedance (Z'') exhibited strong relaxation peaks that shifted toward higher temperatures with frequency, which described the grain, grain boundary, and electrode effects in the samples. The Z''–f plots exhibited similar behavior for the $x = 0$ to 0.6 samples, but for the $x = 0.8$ and 1, the relaxation was observed in the lower frequency region, indicating the grain boundary effect along with the grain effect. The activation energy of the relaxation processes for the $x = 0$–0.6 samples found to be in the range of 0.8 to 1.0 eV, whereas for the $x = 0.8$, the

activation energy was calculated to be around 1.4 eV due to highest Nb concentration in the multiphase samples. Further, for the $x = 1$ sample, two values of the activation energy were observed, i.e., around 1.0 eV and 0.5 eV, which indicated the transition near 460°C. We measured the total conductivity (10^{-3} to 10^{-5} S/m) with varying frequency in the temperature range of 360°–560°C and fitted it with the Jonsher power law. The fitted *dc* conductivity behavior showed the semiconducting characteristics of all the samples, with $x = 0.8$ having the lowest conductivity. The analysis showed that all the samples exhibited OLPT mechanism, whereas the $x = 1$ exhibited a transition near 480°C, which was also reflected in the tangent loss plots. As an exception, we find NSPT and QMT mechanisms for the $x = 1$ sample at higher temperatures.

ACKNOWLEDGMENTS

V.N.T. and Ajay Kumar thank DST (project: DST/TMD/MECSP/2K17/07) and UGC, respectively, for the fellowship. RSD thanks the IIT Delhi for supporting the development of temperature dependent dielectric set-up in the group. RSD acknowledges SERB–DST for the financial support through a core research grant (project reference no. CRG/2020/003436).

[1] H. R. Xia, L. X. Li, H. J. Zhang, X. L. Meng, L. Zhu, Z. H. Yang, X. S. Liu, and J. Y. Wang, Raman spectra and laser properties of Yb-doped yttrium orthovanadate crystals, J. Appl. Phys. **87** (2000) 269. https://doi.org/10.1063/1.371855

[2] M. Yi, S.-K. Park, C.-Y. Seong, Y. Piao, and T. Yu, The general synthesis and characterization of rare earth orthovanadate nanocrystals and their electrochemical applications, J. Alloys Compd. **693** (2017) 825. https://doi.org/10.1016/j.jallcom.2016.09.249

[3] A. Kumar, A. Sharma, M. Sharma, V. Singh, A. Dhaka, and R. S. Dhaka, Structural, vibrational and electronic properties of Nb substituted orthovanadates $LaV_{1-x}Nb_xO_4$, J. Alloys Compd. **966** (2023) 171506. https://doi.org/10.1016/j.jallcom.2023.171506

[4] L. Sun, X. Zhao, Y. Li, P. Li, H. Sun, X. Cheng, and W. Fan, First-principles studies of electronic, optical, and vibrational properties of $LaVO_4$ polymorph, J. Appl. Phys. **108** (2010) 093519. https://doi.org/10.1063/1.3499308

[5] M. Inguscio, R. Wallenstein, and North Atlantic Treaty Organization, editors, Solid State Lasers: New developments and applications (Plenum Press, New York, 1993).

[6] F. Axel and centre de Physique, editors, From Quasicrystal to more complex systems Les Houches School, February 23 to March 6, 1998 order Chance and Risk Aperiodic Phenomena from Solid State to Finance (Springer, 2000).

[7] C. E. Rice and W. R. Robinson, Lanthanum Orthovanadate, Acta Crystallogr B Struct Sci **32** (1976) 2232. https://doi.org/10.1107/S0567740876007450

[8] B. C. Chakoumakos, M. M. Abraham, and L. A. Boatner, Crystal structure refinements of zircon-type MVO_4 (M = Sc, Y, Ce, Pr, Nd, Tb, Ho, Er, Tm, Yb, Lu), J. Solid State Chem. **109** (1994) 197. https://doi.org/10.1006/jssc.1994.1091

[9] M. Zábranský and V. Růžička, Erratum: "Estimation of the heat capacities of organic liquids as a function of temperature using group additivity. An amendment." [J. Phys. Chem. Ref. Data **33**, 1071 (2004)], Journal of Physical and Chemical Reference Data **34** (2005) 39. https://doi.org/10.1063/1.1889439

[10] X. Liu, S. W. Or, S. L. Ho, C. C. Cheung, C. M. Leung, Z. Han, D. Geng, and Z. Zhang, Corrigendum to "Full X–Ku band microwave absorption by Fe(Mn)/Mn_7C_3/C core/shell/shell structured nanocapsules" [J. Alloys Compd. **509**, 9071 (2011)], Journal of Alloys and Compounds **531** (2012) 96. https://doi.org/10.1016/j.jallcom.2011.06.031

[11] Himanshu Dua, R. Shukla, and R. S. Dhaka, Structural phase transition and its consequences for the optical behavior of $LaV_{1-x}Nb_xO_4$, Phys. Rev. B **103** (2021) 174107. https://doi.org/10.1103/PhysRevB.103.174107

[12] S. Verma, B. N. Wani, and N. M. Gupta, Synthesis, characterisation, TPR/TPO and activity studies on $LaMn_xV_{1-x}O_{4-\delta}$–catalysts, Appl. Catal. A: Gen. **205** (2001) 295. https://doi.org/10.1016/S0926-860X(00)00579-2

[13] H. Takei and S. Tsunekawa, Growth and properties of $LaNbO_4$ and $NdNbO_4$ single crystals, Journal of Crystal Growth **38** (1977) 55. https://doi.org/10.1016/0022-0248(77)90372-4

[14] A. T. Aldred, Unusual cell volume behavior in the $LaNb_{1-x}V_xO_4$ system, Materials Letters **1** (1983) 197. https://doi.org/10.1016/0167-577X(83)90017-4

[15] R. Haugsrud and T. Norby, Proton conduction in rare-earth ortho-niobates and ortho-tantalates. Nature Mater **5** (2006) 193. https://doi.org/10.1038/nmat1591

[16] G. Blasse, and L. H. Brixner, Ultraviolet emission from ABO_4-type niobates, tantalates and tungstates, Chem. Phys. Lett. **173** (1990) 409. https://doi.org/10.1016/0009-2614(90)87224-F

[17] C. Balamurugan, D.-W. Lee, and A. Subramania, Preparation and LPG-gas sensing characteristics of p-type semiconducting $LaNbO_4$ ceramic material, Applied Surface Science **283** (2013) 58. https://doi.org/10.1016/j.apsusc.2013.06.013

[18] B. Y. Gu, H. Z. Cummins, S. L. Qiu, and M. Copic, Acoustic soft modes and the ferroelastic phase transition in $BiVO_4$ and $LaNbO_4$, Ferroelectrics **52** (1983) 45. https://doi.org/10.1080/00150198308208237

[19] K. Li, Y. Zhang, X. Li, M. Shang, H. Lian, and J. Lin, Host-sensitized luminescence in $LaNbO_4$:Ln^{3+} (Ln^{3+} = Eu^{3+}/Tb^{3+}/Dy^{3+}) with different emission colors, Phys. Chem. Chem. Phys. **17** (2015) 4283. https://doi.org/10.1039/C4CP03894K

[20] S. Ding, Q. Zhang, W. Liu, J. Luo, F. Peng, X. Wang, G. Sun, and D. Sun, Crystal growth and characterization of a mixed laser crystal: Nd-doped $Gd_{0.89}La_{0.1}NbO_4$, RSC Adv. **7** (2017) 35666.

https://doi.org/10.1039/C7RA05380K
[21] Y. D. Porter-Chapman, E. Bourret-Courchesne, S. E. Taylor, M. J. Weber, and S. E. Derenzo, in: Systematic search for new Lanthanum scintillators, IEEE Nuclear Science Symposium Conference Record, **3** (2006) 1578. https://doi.org/10.1109/NSSMIC.2006.354200
[22] D. Errandonea and F. J. Manjón, Pressure effects on the structural and electronic properties of ABX_4 scintillating crystals, Progress in Materials Science **53** (2008) 711. https://doi.org/10.1016/j.pmatsci.2008.02.001
[23] A. Kumar, and R. S. Dhaka, Unraveling the magnetic interactions and spin state in insulating $Sr_{2-x}La_xCoNbO_6$, Phys. Rev. B **101** (2020) 094434. https://doi.org/10.1103/PhysRevB.101.094434
[24] A. Kumar, B. Schwarz, H. Ehrenberg, and R. S. Dhaka, Evidence of discrete energy states and cluster-glass behavior of $Sr_{2-x}La_xCoNbO_6$, Phys. Rev. B **102** (2020) 184414. https://doi.org/10.1103/PhysRevB.102.184414
[25] A. Kumar, R. Shukla, R. Kumar, R. J. Choudhary, S. N. Jha, and R. S. Dhaka, Electronic and local structure investigation of $Sr_{2-x}La_xCoNbO_6$ using near-edge and extended x-ray absorption fine structures, Phys. Rev. B **105** (2022) 245155. https://doi.org/10.1103/PhysRevB.105.245155
[26] R. Shukla, R. S. Dhaka, Anomalous magnetic and spin glass behavior in Nb substituted $LaCo_{1-x}Nb_xO_3$, Phys. Rev. B **97** (2018) 024430. https://doi.org/10.1103/PhysRevB.97.024430
[27] R. Shukla, A. Jain, M. Miryala, M. Murakami, K. Ueno, S. M. Yusuf, and R. S. Dhaka, Spin dynamics and unconventional magnetism in insulating $La_{(1-2x)}Sr_{2x}Co_{(1-x)}Nb_xO_3$, Journal of Physical Chemistry C **123** (2019) 22457. https://doi.org/10.1021/acs.jpcc.9b05423
[28] R. Shukla, A. Kumar, R. Kumar, S. N. Jha, and R. S. Dhaka, X-ray absorption spectroscopy study of $La_{1-y}Sr_yCo_{1-x}Nb_xO_3$, Journal of Physical Chemistry C **125** (2021) 10130. https://doi.org/10.1021/acs.jpcc.1c02211
[29] V. N. Thakur, A. Kumar, A. Kaushik, G. D. Gupt, R. S. Dhaka, High temperature dielectric and impedance spectroscopy study of $LaCo_{0.7}Nb_{0.3}O_3$, Materials Research Bulletin, **158** (2023) 112070. https://doi.org/10.1016/j.materresbull.2022.112070
[30] M. Huse, A. W. B. Skilbred, M. Karlsson, S. G. Eriksson, T. Norby, R. Haugsrud, and C. S. Knee, Neutron diffraction study of the monoclinic to tetragonal structural transition in $LaNbO_4$ and its relation to proton mobility, Journal of Solid State Chemistry **187** (2012) 27. https://doi.org/10.1016/j.jssc.2011.12.025
[31] Y. Nagao and I. Terasaki, Dielectric constant and ac conductivity of the layered cobalt oxide $Bi_2Sr_2CoO_{6+}$: a possible metal-dielectric composite made by self-organization of Co^{2+} and Co^{3+} ions, Phys. Rev. B **76** (2007) 144203. https://doi.org/10.1103/PhysRevB.76.144203
[32] Z. Y. Cheng, R.S. Katiyar, X. Yao, A.S. Bhalla, Temperature dependence of the dielectric constant of relaxor ferroelectrics, Phys. Rev. B. 57 (1998) 8166–8177. https://doi.org/10.1103/PhysRevB.57.8166
[33] C. Rayssi, S. El Kossi, J. Dhahri, K. Khirouni, Frequency and temperature-dependence of dielectric permittivity and electric modulus studies of the solid solution $Ca_{0.85}Er_{0.1}Ti_{1-x}Co_{4x/3}O_3$ ($0 \leq x \leq 0.1$), RSC Adv. 8 (2018) 17139–17150. https://doi.org/10.1039/C8RA00794B
[34] C. C. Homes, T. Vogt, S. M. Shapiro, S. Wakimoto, and A. P. Ramirez, Optical response of high-dielectric-constant perovskite-related oxide, Science **293** (2001) 673. https://doi.org/10.1126/science.1061655
[35] M. Abdullah Dar, K. Majid, K. M. Batoo, and R.K. Kotnala, Dielectric and impedance study of polycrystalline $Li_{0.35}$–$0.5XCd_{0.3}NiXFe_{2.35}$–$0.5XO_4$ ferrites synthesized via a citrate-gel auto combustion method, Journal of Alloys and Compounds 632 (2015) 307–320. https://doi.org/10.1016/j.jallcom.2015.01.190
[36] R. Gao, Q. Zhang, Z. Xu, Z. Wang, G. Chen, X. Deng, C. Fu, and W. Cai, A comparative study on the structural, dielectric and multiferroic properties of $Co_{0.6}Cu_{0.3}Zn_{0.1}Fe_2O_4/Ba_{0.9}Sr_{0.1}Zr_{0.1}Ti_{0.9}O_3$ composite ceramics, Composites Part B 166 (2019) 204–212. https://doi.org/10.1016/j.compositesb.2018.12.010
[37] R. Gao, Z. Wang, G. Chen, X. Deng, W. Cai, and C. Fu, Influence of core size on the multiferroic properties of $CoFe_2O_4$@$BaTiO_3$ core shell structured composites, Ceramics International 44 (2018) S84–S87. 10.1016/j.ceramint.2018.08.234
[38] R. Gao, X. Qin, Q. Zhang, Z. Xu, Z. Wang, C. Fu, G. Chen, X. Deng, and W. Cai, Enhancement of magnetoelectric properties of $(1-x)Mn_{0.5}Zn_{0.5}Fe_2O_4-xBa_{0.85}Sr_{0.15}Ti_{0.9}Hf_{0.1}O_3$ composite ceramics, Journal of Alloys and Compounds 795 (2019) 501–512. 10.1016/j.jallcom.2019.05.013
[39] Y. Song, Q. Sun, Y. Lu, X. Liu, F. Wang, Low-temperature sintering and enhanced thermoelectric properties of $LaCoO_3$ ceramics with B_2O_3–CuO addition, Journal of Alloys and Compounds **536** (2012) 150. https://doi.org/10.1016/j.jallcom.2012.05.001
[40] R. Meena, R. S. Dhaka, Dielectric properties and impedance spectroscopy of NASICON type $Na_3Zr_2Si_2PO_12$, Ceramics International 48(23), (2022) 35150–35159. https://doi.org/10.1016/j.ceramint.2022.08.111
[41] C. Rayssi, S. E. Kossi, K. Khirouni, Frequency and temperature-dependence of dielectric permittivity and electric modulus studies of the solid solution $Ca_{0.85}Er_{0.1}Ti_{1-x}Co_{4x/3}O_3$ ($0 \leq x \leq 0.1$). RSC Advances 8(31), (2018) 17139–17150. https://doi.org/10.1039/C8RA00794B
[42] I. S. Zheludev, Physics of crystalline dielectrics, electrical properties, Plenum Press, New York, London (1971).
[43] N. Biškup, A. de Andrés, J. L. Martinez, and C. Perca, Origin of the colossal dielectric response of $Pr_{0.6}Ca_{0.4}MnO_3$, Phys. Rev. B **72**, 024115 (2005). https://doi.org/10.1103/PhysRevB.72.024115
[44] P. Lunkenheimer, R. Fichtl, S. G. Ebbinghaus, and A. Loidl, Nonintrinsic origin of the colossal dielectric constants in $CaCu_3Ti_4O_{12}$, Phys. Rev. B **70**, 172102 (2004). https://doi.org/10.1103/PhysRevB.70.172102
[45] P. Lunkenheimer, V. Bobnar, A. V. Pronin, A. I. Ritus, A. A. Volkov, and A. Loidl, Origin of apparent colossal dielectric constants, Phys. Rev. B **66**, 052105 (2002). https://doi.org/10.1103/PhysRevB.66.052105
[46] D. Triyono, S.N. Fitria, and U. Hanifah, Dielectric analysis and electrical conduction mechanism of $La_{1-x}Bi_xFeO_3$ ceramics, RSC Adv. **10** (2020) 18323. https://doi.org/10.1039/D0RA02402C

[47] N. Sivakumar, A. Narayanasamy, C. N. Chinnasamy, and B. Jeyadevan, Influence of thermal annealing on the dielectric properties and electrical relaxation behaviour in nanostructured $CoFe_2O_4$ ferrite, J. Phys.: Condens. Matter **19** (2007) 386201. 10.1088/0953-8984/19/38/386201

[48] A. Srivastava, A. Garg, and F. D. Morrison, Impedance spectroscopy studies on polycrystalline $BiFeO_3$ thin films on Pt/Si substrates, J. Appl. Phys. **105** (2009) 054103. https://doi.org/10.1063/1.3078822

[49] A. Kahouli, A. Sylvestre, F. Jomni, B. Yangui, and J. Legrand, Experimental and theoretical study of ac electrical conduction mechanisms of semicrystalline parylene C thin films, J. Phys. Chem. A **116** (2012) 1051. https://doi.org/10.1021/jp207114u

[50] A. R. Long, Frequency-dependent loss in amorphous semiconductors, Advances in Physics **31** (1982) 553. https://doi.org/10.1080/00018738200101418

[51] M. Ben Bechir, K. Karoui, M. Tabellout, K. Guidara, and A. Ben Rhaiem, Alternative current conduction mechanisms of organic-inorganic compound $[N(CH_3)_3H]_2CuCl_4$, J. Appl. Phys. 115 (2014) 203712. https://doi.org/10.1063/1.4880735

[52] A. Ghosh, Transport properties of vanadium germanate glassy semiconductors, Phys. Rev. B **42** (1990) 5665. https://doi.org/10.1103/PhysRevB.42.5665

[53] A. Ghosh, Frequency-dependent conductivity in bismuth-vanadate glassy semiconductors, Phys. Rev. B **41** (1990) 1479. https://doi.org/10.1103/PhysRevB.41.1479

[54] S. R. Elliott, Temperature dependence of a.c. conductivity of chalcogenide glasses, Philos. Mag. B. 37 (1978) 553–560. https://doi.org/10.1080/01418637808226448

DO Modeling for the Yamuna River, Delhi

Nibedita Verma[1*] Dr. Geeta Singh [2] Prof. Naved Ahsan[3]
[1]*Research Scholar, Environmental Engineering Department, Delhi Technological University, Delhi, India*
[2]*Assistant professor, Environmental Engineering Department, Delhi Technological University, Delhi, India*
[3]*Professor, Civil Engineering Department, Jamia Millia Islamia University, New Delhi, India*

[1*]nibedita_2k19phd@dtu.ac.in [2] geeta.singh@dce.ac.in
[3]*nahsan@jmi.ac.in*

Abstract- **Yamuna River Delhi stretch is a severely polluted reach with dissolved oxygen concentration below the detected limit almost the entire length with a higher concentration of oxygen-demanding substances. This stretch needs a management tool to predict the conditions and management of the river with the contributing load. QUAL2Kw is used to simulate the parameter DO. The framework has been calibrated and validated for the secondary data of February data of 1999-2006 and 2007-2009 respectively. DO prediction was done by keeping BOD and COD for all outfalling drains 10 mg/l and 50 mg/l as the effluent standard set for this reach. The scenarios have been developed changing the upstream flow. With the fixed pollutant load, it was observed that if the flow is increased by 50 cumecs of the minimum suggested flow, the reach could not attain the required minimum DO concentration.**

Keywords- **Dissolved Oxygen, Water Quality Model, QUAL2Kw, Urban river reach, Biochemical Oxygen Demand**

1. INTRODUCTION

Yamuna River originated from the Yamunotri, north of Haridwar in the Himalayan Mountains. After flowing total length of 1386 kilometers covering five states Uttaranchal, Uttar Pradesh, Himachal Pradesh, Haryana, Rajasthan, Madhya Pradesh & NCT – Delhi with a watershed area of 366223 km^2, its confluence with the river Ganga at Prayag Raj, Uttar Pradesh. Before reaching the Delhi stretch, the river water diverts and controls by a weir to produce electricity in Uttaranchal. In Tajewala / Hathnikund in Haryana state. During the dry season, the river becomes dry in some portions of the Tajewala and Delhi segment and regains by groundwater discharge and Som Nadi, a seasonal stream.

The river enters Delhi near Palla village and courses through Delhi for about 22 km. In Delhi, the river is tapped at Wazirabad through a barrage for supplying drinking water in the city, and no water is allowed to flow downstream during the dry season beyond the Wazirabad barrage to fulfill the water supply requirement in the capital Delhi [1]. After the Wazirabad barrage, in the Delhi stretch, the river is fed by 16 major drains along with the water transported by Haryana Irrigation Department from Western Yamuna Canal to Agra Canal via Najafgarh Drain, with partially treated or untreated domestic and industrial wastewater from different places and the river becomes a sewer line [1]. In the Delhi stretch, wastewater from the National Capital Region contributes about 80% of the pollution of the total river length [1] and becomes a sewerage line. This massive pollution is due to the enormous abstraction of water for irrigation, water supply for drinking and industrial purposes, and also the outfalling of untreated or partially treated industrial and domestic wastewater. it is an urgent requirement to develop a proper waste load allocation process to maintain the ecological health of the river and the sustainable development of society. Data from WRIS, India, at Delhi Railway Bridge station - pH value is7.65, DO is around 0.37 mg/l (In few years, it was not detectable), Biochemical Oxygen demand 38.78mg/l, fecal Coliform 2201851.85 in the year 2019, and Ammonia 10.75 mg/l (the last 10-year average value is very higher than the acceptable limit of 0.5 mg/l).

The diminishes of DO mainly occur in the low flow season. A high concentration of Biochemical Oxygen and Chemical oxygen demand with nutrients reduces the DO and consequently, the reach becomes anoxic. Hence it requires developing management options that can predict the DO concentrations and manage the required standard throughout the reach.

QUAL2kw which originated from QUAL2k, and modified versions of QUAL2E or QUAL2E- UNCAS [2] which was deliberated as a water quality assurance tool to interact with algal growth, nutrients, and dissolved oxygen, boundary conditions, nonconservative constituents with input-output consideration for steady and dynamic states and uncertainty analysis like sensitivity analysis [3] and deterministic tool for

waste load allocation, as well as total maximum daily load analyses[4]. Applications of QUAL2Kw are found in various research works such as [5][6][7][8] and successfully determined the waste load allocation and Total maximum daily load allocation. The present study develops a modeling approach using the QUAL2Kw framework to maintain the DO concentration of the urban reach of Yamuna, Delhi.

2. METHODOLOGY

The present study developed DO profiles for the Yamuna River, Delhi, which is from Palla to Wazirabad. In Palla, the river enters Delhi and after 22 km it reaches Wazirabad. At Wazirabad, the water is collected and supplied to Capital City, and after Wazirabad, 16 drains dispose of the partially and untreated water into the river, and at Okhla 22 km downstream from Wazirabad river leaves Delhi. This study includes four monitoring stations Palla (M1), Old Railway Bridge (M2), Nizamuddin (M3), and Okhla (M4), and 16 drains were taken as point sources. The upstream flow was taken constant for both calibration and validation Two abstraction points are taken as Wazirabad and Agra Canal. The data for water quality was collected from secondary sources for the year 1999-2009 and the month of February. The average data of the first seven years has been used for calibration and the last three years have been used for validation. The O'Conner -Dobbins equation has been used to determine the reaeration coefficient [6]. Root mean square values have been calculated. The total 44 km stretch of river was segmented into 20 reaches. The model was calibrated and validated manually to get minimum RMSE values. The model has been run to determine the minimum load for maintaining the required minimum DO concentration (4 mg/l) throughout the reach. In the DO modeling for this stretch, the flow of water has been kept minimum at 10 cumecs at Wazirabad to Okhla, and BOD and COD have been kept at 10 mg/l and 50 mg/l for all contributing drains as directed by NGT, Delhi. The flow has been increased 10 cumecs intervals to attain the required DO concentration.

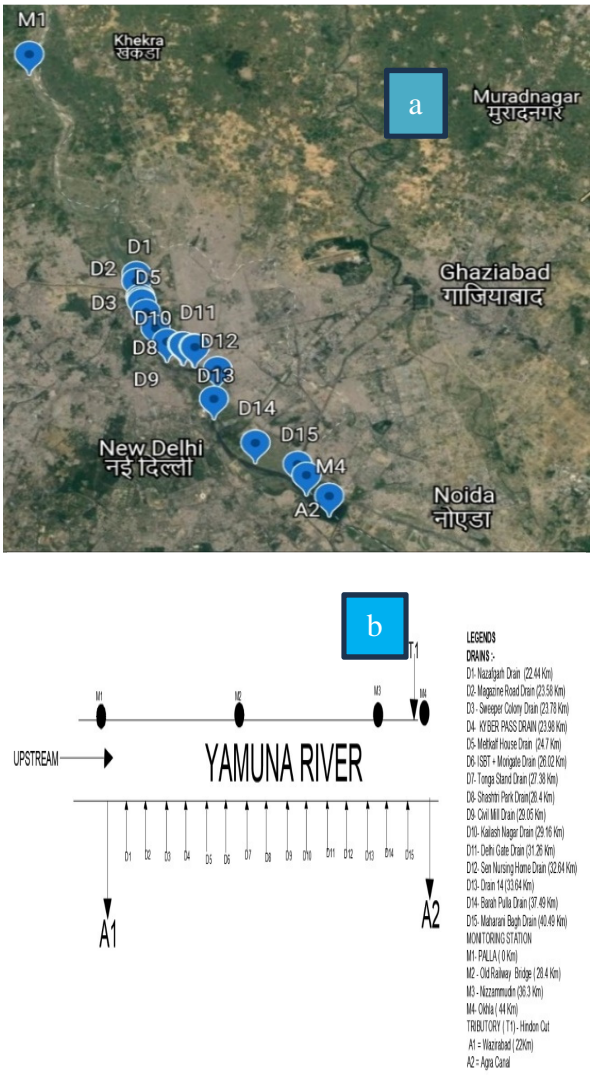


Fig.1a-b Outfalling drains locations and name [10]

TABLE I
Head Water Input

Parameter	Calibration	Validation
Temperature	15.18°C	14.5°C
DO	8.47 mg/l	8.65 mg/l
BOD	1.61mg/l	1.75mg/l
COD	5.51 mg/l	5.13 mg/l
pH	7.82 s.u.	7.85 s.u.

Source:[9];[1]

TABLE II
Water Quality for Different Stations

Station Name	Distance Km	DO mg/l	BOD mg/l	COD mg/l	pH
For Calibration					
M2	28.4	0.085	42.9	65.43	7.06
M3	37.5	0.01	35.91	49.67	6.82
M4	44	0	22.7	40.43	6.88
For Validation					
M2	28.4	0.0825	34.025	63.33	7.2
M3	37.5	0.02	22.5	39	6.93
M4	44	0.03	19.66	29	6.975

Source:[9][1]

TABLE III
RMSE Values for Calibration and Validation

RMSE	Calibration	Prediction
DO	0.021	0.021
BOD	0.4167	0.3568
COD	0.3755	0.411

TABLE IV
Headwater Input for DO Modeling

Parameter	Values
Dissolved Oxygen	8.40 mg/l
BOD	2.40 mg/l
COD	12.00 mg/l
pH	7.9 s.u.

Source: DPCC domain

TABLE V
Point Source Input

Drain No	Flow m^3/s *	BOD, mg/l	COD, mg/l
1	24.91	10	50
2	0.22	10	50
3	0.43	10	50
4	0.150	10	50
5	0	0	0
6	0.16	10	50
7	0.45	10	50
8	0.09	10	50
9	0.58	10	50
10	0.79	10	50
11	0.58	10	50
12	1.07	10	50
13	1.81	10	50
14	0.81	10	50
15	10.69	10	50

* DPCC domain

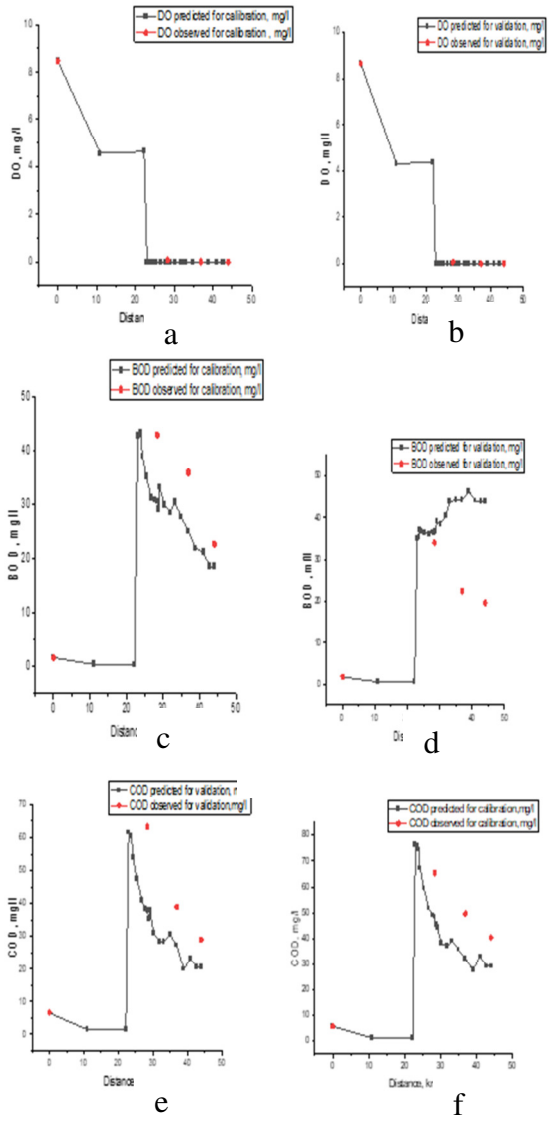


Fig.2 Calibration and validation for DO, BOD, and COD (a-f)

3. RESULTS

Fig. 2 depicts the observed and predicted profiles for calibration and validation. Fig 2(a) and Fig 2(b) are shown the DO prediction for calibration and Validation. It was observed that after the meeting of the D1 (Najafgarh drain), the dissolved oxygen became zero for the entire reach. From the simulation results, it has been observed that the predicted values for DO were quite the same as the observed values. BOD values were different from observed values in Fig 2(a) and 2(b). This might be due to the sampling collection time. But RMSEV for the calibration and validation were 0.41 and 0.35 which were very close. Table III shows the RMSE values for calibration and validation which are quite acceptable. Hence, this model can be used to manage urban

river stretches like the Yamuna River Delhi [6][10]. Fig. 3 shows the

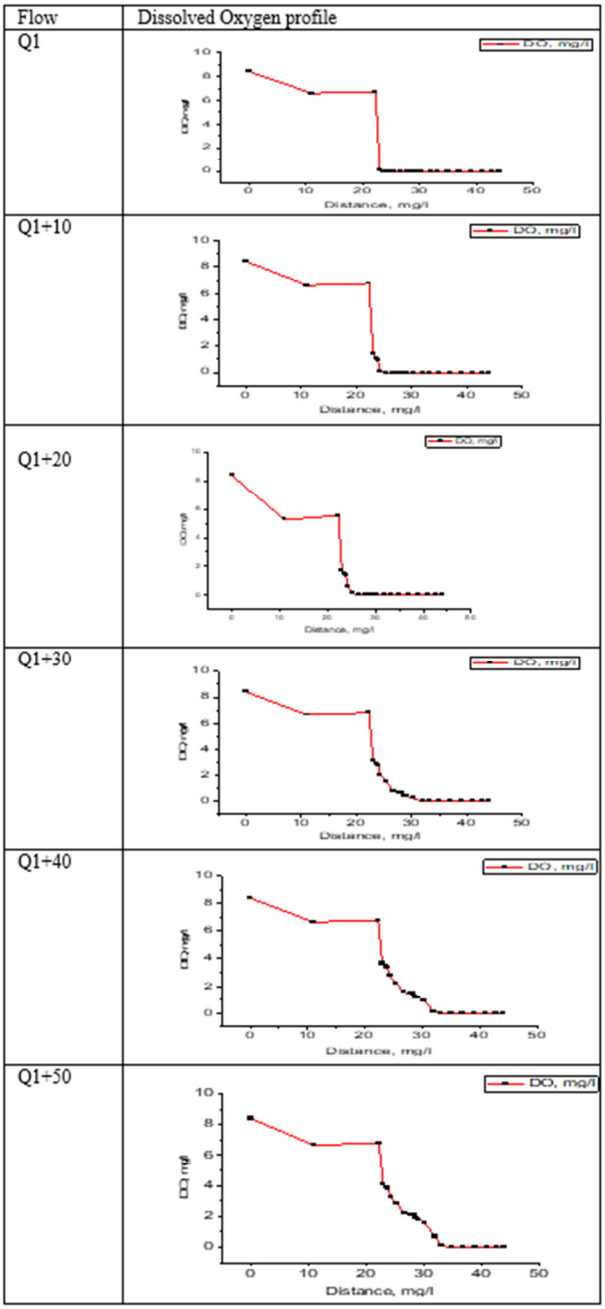


Fig. 3 DO profiles with flow variations

DO profile for different flow conditions. Fig. 3 shows that with an increment of flow, DO conditions got better conditions. Although, it was observed that 50 cumecs increment flow could not attain the required 4 mg/l of DO concentration.

4. CONCLUSIONS

The water quality Model QUAL2Kw is a one-dimensional and less complex model requiring less data. In many developed countries like the USA, South Korea used this model for waste load allocation and Total maximum daily load allocation. For India, which is a developing country can use this model for river management and maintain the ecological health of the river. Although this study shows that DO concentration increased with the flow increase, more management options should be adopted to attain the required DO concentration of this reach. Furthermore, due to data constraints, this study did not include nutrient data, Hence, it is suggested that more parameters need to be monitored and the frequency of the monitoring should be increased.

REFERENCES

[1] CPCB. 2006. "Water Quality Status of Yamuna River, Assessment and Development of River Basin." : 1–115. www.cpcb.nic.in.

[2] Aliffia, A., and N. Karnaningroem. 2019. "Simulation of Pollution Load Capacity Using QUAL2Kw Model in Kali Surabaya River (Cangkir-Sepanjang Segment)." *IOP Conference Series: Earth and Environmental Science* 259(1).

[3] Greogory Pellettier, Steven Chapra, Hua tao. 2006. "QUAL2Kw – A Framework for Modeling Water Quality in Streams and Rivers Using a Genetic Algorithm for CalibrationNo Title." *environmental modelling and software* 21(3): 419–25.

[4] Hobson, A. J. et al. 2015. "Development of a Minimalistic Data Collection Strategy for QUAL2Kw." *Journal of Water Resources Planning and Management* 141(8).

[5] Teegavarapu, Ramesh, Lindell Ormsbee, Amin Elshorbagy, and Ramesh S V Teegavarapu. 2014. "Reclamation of Oil Sand Field & Reclamation of Saline Heavy Clay Soils View Project Artificial Intelligence Modeling of Natural Phenomena View Project Total Maximum Daily Load (TMDL) Approach to Surface Water Quality Management: Concepts, Issues, and Applications."

[6] Sharma, Deepshikha, Arun Kansal, and • Greg Pelletier. 1999. "Water Quality Modeling for Urban Reach of Yamuna River, India (1999–2009), Using QUAL2Kw."

[7] Lakshmi, E, and And G Madhu. 2020. "EVALUATION AND MODELLING OF DISSOLVED OXYGEN AND RE-AERATION RATE IN RIVER PERIYAR, SOUTH INDIA." *Indian J.Sci.Res* 10(2): 1–14.

[8] Gikas, Georgios D. 2014. "Water Quality of Drainage Canals and Assessment of Nutrient Loads Using QUAL2Kw." *Environmental Processes* 1(4): 369–85.

[9]Sharma, Deepshikha. 2013. "Evaluation of River Quality Restoration Plan and Intervention Analysis Using Water Quality Modeling with Focus on the River Yamuna, Delhi (India). PhD Thesis, TERI University, Delhi, India."

[10]Verma, Nibedita, Geeta Singh, and Naved Ahsan. 2022. "Development of Water Quality Management Strategies for an Urban River Reach: A Case Study of the River Yamuna, Delhi, India." *Arabian Journal of Geosciences* 15(24).

Microsystem Technologies
https://doi.org/10.1007/s00542-023-05560-4

TECHNICAL PAPER

Dual metal dual layer GAA NW–FET (DMDL–GAA–NW–FET) biosensor for label free SARS-CoV-2 detection

Shivani Yadav[1] · Sonam Rewari[1]

Received: 22 December 2022 / Accepted: 24 October 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
The current Covid-19 epidemic has necessitated the development of a biosensor that is expected to be extraordinarily sensitive, rapid, precise, and economical for the detection of the SARS-CoV-2 virus. This paper proposes a novel structure named Dual Metal Dual Layer Gate-All-Around Nanowire Field Effect Transistor (DMDL-GAA-NW-FET) Biosensor for detection of SARS-CoV-2 in terms of the S (Spike) protein and DNA of the virus using the concept of dielectric modulation and charge modulation. The proposed device employs gate work function engineering to segregate the gate into two layers with two distinct work functions, improving gate control. As a result, the biosensor's capability to identify the SARS-CoV-2 virus is significantly improved. Analysis of the proposed device's electrostatic behavior in terms of channel potential, energy band variation and electron concentration has been performed and then SARS-CoV-2 viral detection has been validated using the threshold voltage drift (ΔV_{TH}), I_{ON} current drift (ΔI_{ON}), g_m, I_{ON}/I_{OFF} ratio and g_d upon virus immobilization in the nanocavity. The effectiveness of the DMDL-GAA-NW-FET biosensor in detecting SARS-CoV-2 is further highlighted by a comparison of the biosensor's threshold voltage and I_{ON} current sensitivities with those of a conventional Gate-All-Around Nanowire FET (GAA-NW-FET) biosensor.

1 Introduction

The severe acute respiratory syndrome corona virus (SARS-CoV-2) that led in coronavirus disease 2019 (covid-19) outbreak was proclaimed a pandemic by the World Health Organization on March 11, 2020 (Cucinotta and Vanelli 2020). This ongoing COVID-19 pandemic, with its rapidly changing variants, is looming as the biggest threat to world health in a century (Gorbalenya et al. 2020).

The most common approach for diagnosis of this disease at the moment is real-time reverse transcription–polymerase chain reaction (RT-PCR), which uses an inner nasal swab material to identify the genes for E-protein, N- protein, open frame reading b1 (OFRb1), and OFRb2. Placing the viral RNA in the test kit is the first step in the detection process. After that, primer is applied to the viral Ribonucleic acid (RNA) to create a copy of it. Further, this virus's RNA is transformed into complementary deoxyribonucleic acid (C-DNA) by the reverse transcriptase process, and different copies are created utilizing the DNA polymerization procedure (Quest Diagnostics 2020; Weissleder et al. 2020). Although the RT-PCR detection method is sensitive and has excellent specificity for COVID-19 detection, this approach requires expensive reagents, sophisticated equipment, trained staff and has a lengthy processing time of almost 4 h (Liu and Rusling 2021).

Understanding the structural intricacies of the SARS-CoV-2 virus (Hardenbrook and Zhang 2022) is pivotal not only in deciphering its behavioral patterns but also as a cornerstone for the development of vaccines (Pelletier et al. 2022), antiviral drugs (Yang and Rao 2021; Wu et al. 2022), and diagnostic tools tools (Nag et al. 2020).

While traditional vaccination strategies have been invaluable in curbing the virus's spread, novel approaches rooted in mechanical principles and natural frequencies show promise as innovative alternatives. Dastjerdi et al. have probed the mechanical characteristics of SARS-CoV-2 using a viscoelastic spherical model, suggesting that targeted application of specific natural frequencies, delivered through mechanical loads and vibrations, could hold the potential for virus destruction (Dastjerdi et al. 2022).

✉ Shivani Yadav
shivaniyadav.ece@gmail.com

Sonam Rewari
rewarisonam@gmail.com

[1] Electronics & Communication Engineering Department, Delhi Technological University, Delhi, India

Published online: 17 November 2023

Warsame et al. further supports this, highlighting that the natural frequencies of 88.52–125.13 MHz computed through modal analysis can denature SARS-CoV-2 when subjected to external frequencies, potentially utilizing ultrasound or electromagnetic waves (Warsame et al. 2022). These findings align with the work of Wierzbicki et al., which identifies distinct resonant frequencies specific to the SARS-CoV-2 virus and introduces ultrasound-induced vibrations as a groundbreaking alternative to conventional vaccination strategies, particularly in addressing emerging viral strains (Wierzbicki et al. 2021).

FET based biosensors are among the most recent developments in rapid diagnostic techniques and have a number of benefits including compact size, real-time detection, high sensitivity, and ability for integrated multiplexing (Shao et al. 2021; Chen et al. 2017; Ishikawa et al. 2009; Fathi-Hafshejani et al. 2021; Seo et al. 2020). GAA Nanowire FETs in which the Gate material encircles the whole channel, are best suited for biosensing applications due to their maximum gate controllability, compatibility with current semiconductor industry trends, intrinsic simplicity of fabrication, and greater sensitivity (Das et al. 2022b)(Yadav and Rewari 2022). Biomolecules are immobilized in the carved nanocavities created by etching off the dielectric material below the gate in a Dielectric Modulated-FET (DM-FET) biosensor (Gu et al. 2009; Im et al. 2007). The immobilized biomolecules in the cavity region cause the gate capacitance to change in DM-FET biosensors, which in turn causes changes in electrical characteristics like the threshold voltage and current (Gu et al. 2009),(Im et al. 2007),(Kinsella and Ivanisevic 2007). It is worth noticing that the device characteristics such as threshold voltage and current are significantly influenced by doping (Das et al. 2022a) and hence, sensitivity also gets affected.

Labels are not used in label-free biosensing technology for measurements. For sensing purposes, they employ the analytes' innate physical characteristics, such as molecular weight, size, electrical impedance, charge, refractive index, or dielectric permittivity (Dixit et al. 2021).

This paper presents the label free detection method of the Covid-19 with the help of S-protein and C-DNA immobilization of SARS-CoV-2 in the nanocavity region of DMDL-GAA-NW-FET. In this work, the junctionless accumulation mode (JAM) nanowire FET device has mainly been implemented using the dual-metal (DM) and cylindrical Gate ALL Around (GAA) schemes because it has been already illustrated (Lou et al. 2012; Singh and Raman 2018) that the dual metal gate provides better drive current, transconductance, and reduced SCEs and is thus suitable for augmented analog performance, while the cylindrical GAA helps to increase the gate control over the channel. As the channel length is reduced, several effects start to occur in FETs, such as impact ionization, drain-induced barrier lowering (DIBL), carrier velocity saturation, gate oxide leakage, hot carrier effects, increased subthreshold swing, and increased leakage current (Kaul et al. 2022). These effects are known as short channel effects. Dual Metal Gate (DMG) is a technique that involves using two different metal gates with different work functions so that $\phi_{m1} > \phi_{m2.}$ When the electrons travels from source to drain region its velocity is reduced for DMG structure as $\phi_{m1} > \phi_{m2}$ in comparison to single metal gate (SMG) structure (Goel et al. 2020). The dual metal gates also allow for better electrostatic control over the channel, as the combination of different work functions helps in optimizing the threshold voltage (V_{TH}) of the device. This can help reduce the DIBL effect, which occurs when the drain voltage induces a lowering of the barrier at the source end of the channel, causing increased leakage current and reduced performance. Also, by carefully selecting the work functions of the two gates, the subthreshold swing is optimized, leading to improved device performance (Pal and Sarkar 2014; Rony et al. 2017).

Four sections of this paper are presented in the following order. Section 1 presents an introduction and a review of the pertinent literature for different detection techniques of SARS-COV-2 and emerging nanowire FETs biosensors. In Sect. 2, the design, and simulation strategy of DMDL-GAA-NW-FET have all been covered. Section 3 demonstrates the results analysis, and Section 4 summarises and draws the conclusions.

2 Device design and simulation approach

Spike (S), Membrane (M), Nucleocapsid (N), and Envelope (E) proteins are the four different kinds of proteins present in SARS-CoV-2. Figure 1a (Chen et al. 2022) shows the same. Spike protein (S-protein) is on the viral envelope, and its S1 subunit is the virus' outermost part (Prof Nanshan Chen et al. 2020; Zhu et al. 2020; Lu et al. 2020). Human ACE2 serves as a host for spike (S) proteins, which initiate an infection. Following this, SARS-CoV-2 enters the host cell and generate greater number of viruses that are later spread the infection to new cells Fig. 1b (Xie et al. 2020).

A three-dimensional illustration of the proposed DMDL-GAA-NW-FET biosensor is shown in Fig. 2a. It is a junctionless accumulation mode device in which the doping in the channel region is marginally lower ($N_{D-} = 1 \times 10^{18}$ /cm^3) than in the source and drain areas ($N_{D+} = 1 \times 10^{19}$ /cm^3).

Figure 2b depicts the proposed biosensor structure in vertical cross-sectional aspect. The DMDL-GAA-NW-FET biosensor's two-dimensional view and its capacitance

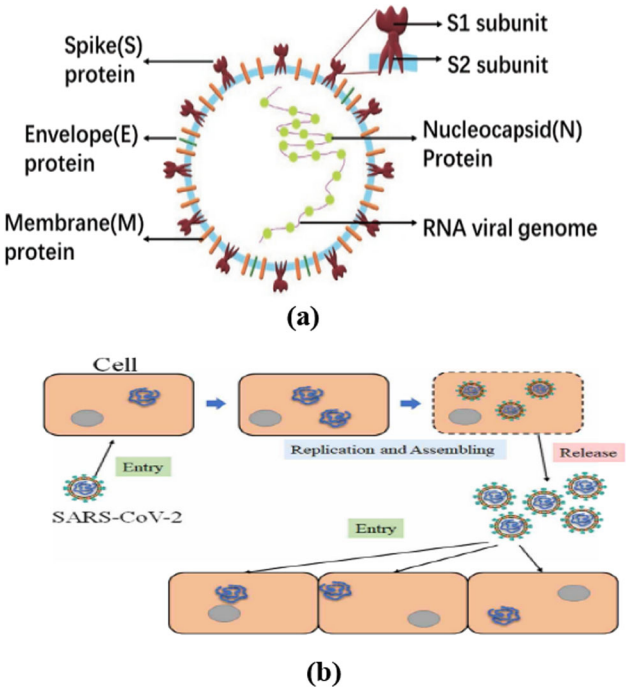


Fig. 1 **a** Structure of SARS-CoV-2 (Chen et al. 2022). **b** Process of SARS-CoV-2 infecting host cells (Xie et al. 2020)

model is depicted in Fig. 2c. The internal device construction can be explained with the help of the Fig. 2c. Deposition of high-K dielectric materials directly on silicon substrates would reduce performance, predominantly because of the fringing field that permeates the channel from the source and drain terminals, weakening the gate control (Kasturi et al. 2005). Therefore, in this work, gate oxide stack engineering (SiO_2 + High K material HfO_2) has been used to reduce the fringing Filed and to improve device performance (Das et al. 2020). The device is having Dual metal Dual Layer structure in which the two layers of Gate are made of different metals having different work functions. Gate Layer 1 is formed by depositing gate metal of work function 5 eV in a trench that is created by etching the HfO_2 dielectric layer, which is 7 nm thick. After that, GATE Layer 2 is created by selectively depositing a metal with a 5.5 eV work-function on top of the HfO_2 layer. Further, plasma dry etching process can be used to realize the nanocavity region in the biosensor device for immobilization of biomolecules. In the nanocavity, as shown by the 2D structure in Fig. 2c, biomolecules have been introduced evenly, by varying the dielectric permittivity. In the capacitance analogous circuit, the lumped capacitances of the DMDL-GAA-NW-FET biosensor are represented by C_{bio} for the variable capacitance of nanogap cavity having biomolecule species, C_{SiO2} for the capacitance of SiO_2 layer, and C_{HfO2} for the capacitance of HfO_2 layer. The capacitance of the nanocavity will vary depending on the physical parameters of the virus, modifying the biosensor characteristics and thus sensing the virus according to the capacitance model of the device. Table 1 illustrates various device parameters for both proposed DMDL–GAA–NW–FET and conventional GAA–NW–FET.

The simulation results were evaluated using the Atlas-3D device simulation tool. The simulations include models for carrier recombination and carrier mobility generation. CONMOB model has been employed for doping versus mobility validation and BGN model has been used for high channel doping. The CONMOB model is used to calculate the effective mobility of carriers in the channel region of the GAA FET. This model takes into account the effect of impurities, interface roughness, and scattering mechanisms on carrier mobility. The mobility values calculated using the CONMOB model can be used to determine the device performance parameters such as drain current, transconductance, and subthreshold swing (Sarkar 2016). On the other hand, the BGN model is used to model the narrowing of the bandgap in the channel region of the GAA FET due to the presence of an electric field. This model takes into account the quantum confinement effect and the variation of the effective mass of carriers in the channel region. The BGN model is important for accurately predicting the device performance in the sub-threshold regime (Sarkar 2016). Both of these models are important for simulating the electrical behavior of GAA FET biosensors, as they affect the device performance in both the on-state and off-state conditions. The recombination of minority carriers in semiconductors has also been taken into account through the use of Shockley–Read–Hall (SRH) and Boltzmann transport equations. The Shockley–Read–Hall recombination is modeled as follows:

$$R_{SRH} = \frac{pn - {n_{ie}}^2}{TAUPO\left[n + n_{ie}\exp\left(\frac{ETRAP}{KT_L}\right)\right] + TAUNO\left[p + n_{ie}\exp\left(\frac{-ETRAP}{KT_L}\right)\right]} \quad (i)$$

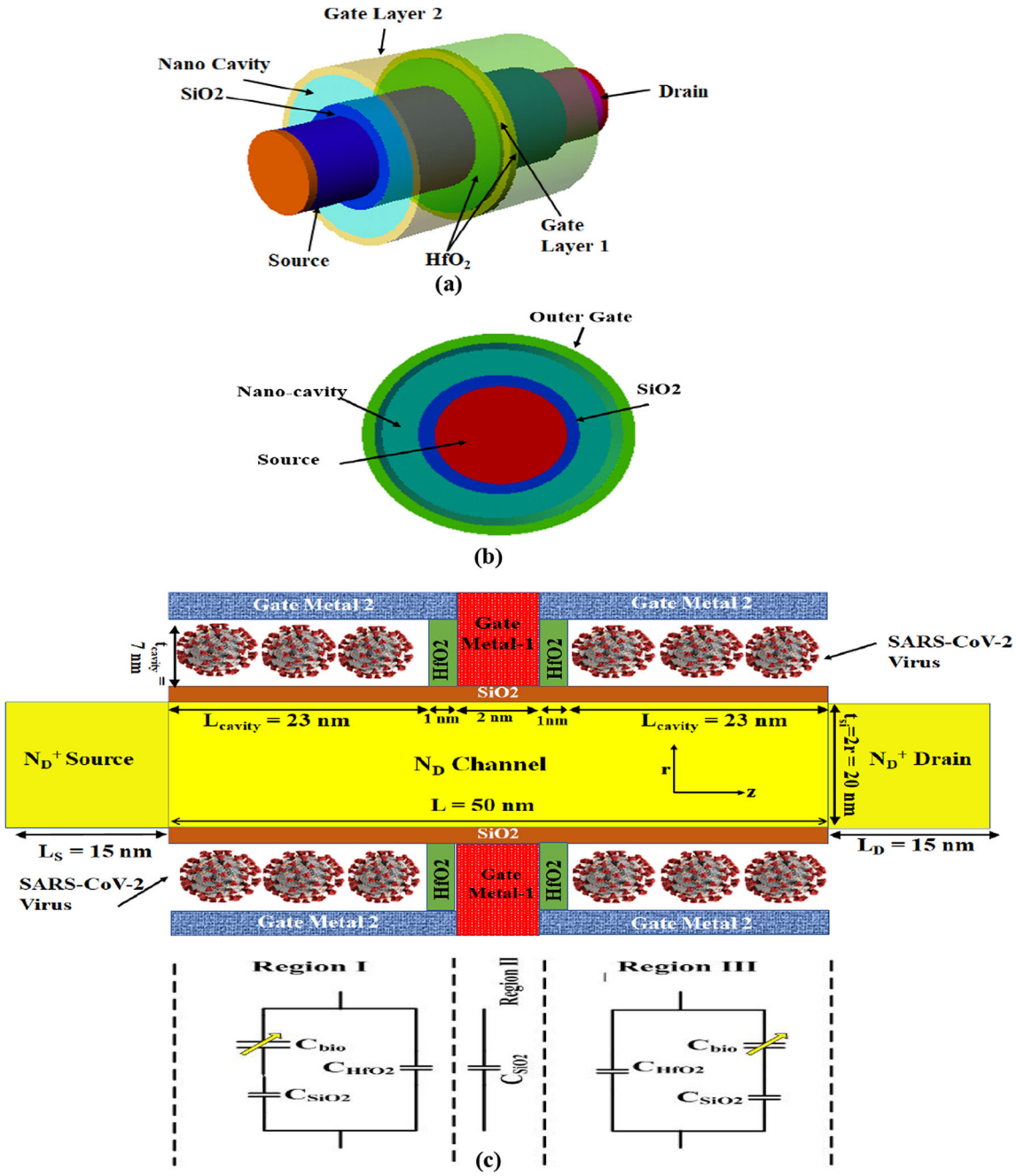


Fig. 2 **a** 3-Dimensional View of DMDL–GAA–NW–FET **b** Vertical Cross-sectional view of DMDL–GAA–NW–FET **c** 2- Dimensional Structure of DMDL–GAA–NW–FET

where ETRAP is the difference between the trap energy level and the intrinsic Fermi level, T_L is the lattice temperature in kelvin and TAUN0 and TAUP0 are the electron and hole lifetimes (Sarkar 2016).

The Boltzmann transport statistics equation is represented as:

$$f(\varepsilon) = \frac{1}{1 + \exp\left(\frac{\varepsilon - E_F}{kT_L}\right)} \quad (ii)$$

where f(ε) is the probability that an available electron state with energy ε is occupied by an electron, E_F spatially

Table 1 Design parameters

Physical device parameter	DMDL–GAA–NW–FET	Conventional GAA–NW–FET
Channel Length (nm)	50	50
Thickness of siliconfilm (nm)	20	20
Length of source/drain (nm)	15	15
Oxide Thickness, SiO2 & HfO2 (nm)	3.0 & 7.0	3.0 & 7.0
Doping of Channel Region, N_D (/cm^3)	1×10^{18}	1×10^{18}
Doping of Source and Drain Region, N_D^+ (/cm^3)	1×10^{19}	1×10^{19}
Thickness of Cavity (nm)	7	7
Gate Layer-1 Thickness (nm)	9	–
Gate Layer-2 Thickness (nm)	2	2
Length of Cavity (nm)	23	23
HfO2 and SiO2 dielectric constants	25.0 & 3.9	25.0 & 3.9

independent reference energy known as the Fermi level, T_L temperature and k is Boltzmann's constant (Sarkar 2016).

3 Device calibration and fabrication

In order to validate the reliability of the simulation models, the simulation data have been calibrated with experimental data from a pre-fabricated JAM nanowire FET (Najmzadeh et al. 2012). The simulation and experimental results are fairly well matched for the transfer characteristics (I_{DS} versus V_{GS} curve), as shown in this Fig. 3a.

The fabrication of the biosensor begins with some fundamental fabrication steps (Li et al. 2021) such as substrate preparation, silicon epitaxial growth, sacrificial layer deposition, Gate Dielectric Deposition using techniques such as thermal oxidation or chemical vapor deposition (CVD), Gate Electrode Deposition using techniques such as PVD or CVD (Das et al. 2023a), and Nanowire Formation (Han et al. 2020; Ye et al. 2021), which will serve as the sensing element of the biosensor. The proposed detailed fabrication steps of Dual Metal Dual Layer Gate-All-Around Nanowire Field Effect Transistor (DMDL-GAA-NW-FET) Biosensor are mentioned in a flowchart in Fig. 3b.

4 Results and discussion

The SARS-CoV-2 virus's spike (S) protein protrusions have a dielectric constant that ranges from 2 to 4 (Gilson and Honig 1986; Talley et al. 2008). Furthermore, the DNA of the virus contains positive and negative charge densities, which make it possible to identify viruses based on charge density modulation. DNA has a constant dielectric constant between 1 and 64. Therefore, S-protein and C-DNA are represented in this work by $K = 2.1$ and $K = 4$.

4.1 Sensitivity comparison of proposed biosensor

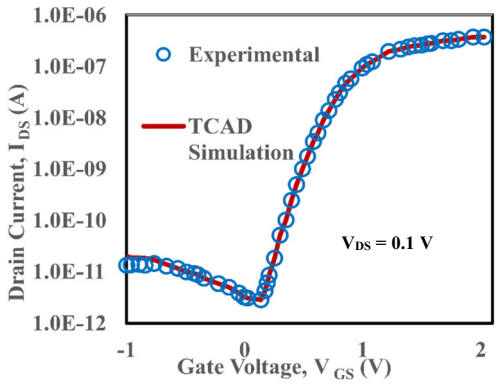
The conventional GAA–NW–FET biosensor has been designed and simulated for the same dimensions as the proposed biosensor for the detection of the SARS-CoV-2 virus in order to compare the proposed DMDL–GAA–NW–FET biosensor's performance to that of the reported conventional GAA biosensor (Priyadarshani et al. 2022). For sensitivity performance comparison the V_{TH} sensitivity and I_{ON} current sensitivity has been taken as a measure for S-Protein and DNA based charged biomolecule detection of SARS-CoV-2 virus.

For Label Free SARS-CoV-2 Detection using DMDL-GAA-NW-FET, the gate dielectric constant of the FET is altered depending upon the type of S-protein immobilized. Furthermore, the device's threshold voltage, on-state current, and off-state current are all calculated as per the S-Protein and its sensitivity. A biosensor's sensitivity can be determined using one of three methods: drift in an electrical parameter, fractional change in an electrical parameter, or the ratio of an electrical parameter with a biomolecule present to an electrical parameter without a biomolecule.

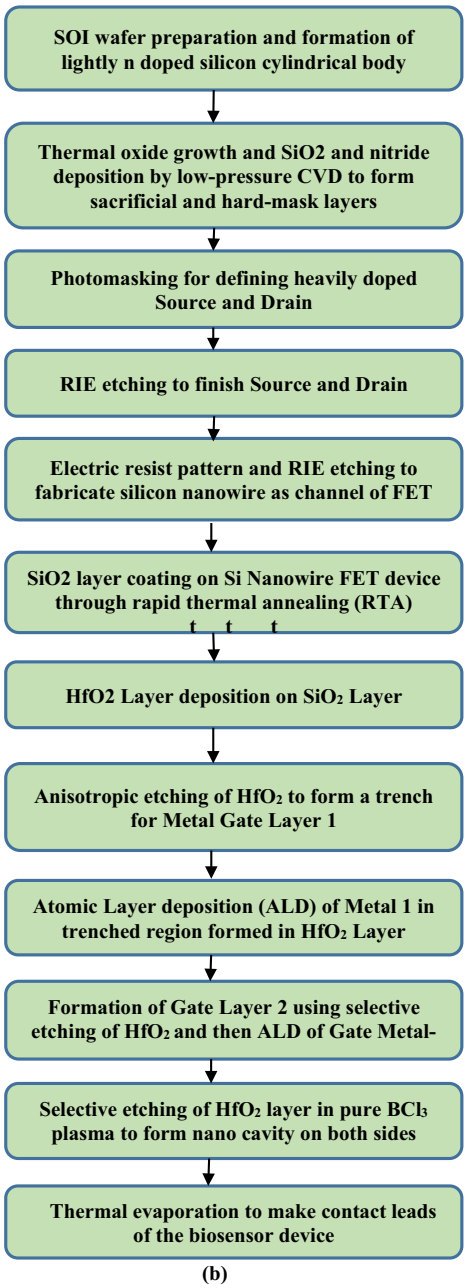
Threshold Voltage sensitivity ($S_{Vt_{Nbio}}$) is defined as the fractional change in threshold voltage (V_{TH}) with and without S-Protein immobilized into the nanocavity. It can be formulated as (Kumar et al. 2021):

$$S_{Vt_{Nbio}} = \frac{V_{TH[with\,S-protein,K=2.1/4]} - V_{TH[without\,S-protein]}}{V_{TH[without\,S-protein]}} \quad (1)$$

The impact of S-protein on the V_{TH} sensitivity of the DMDL–GAA–NW–FET and GAA–NW–FET is shown in



(a)



(b)

◀**Fig. 3** **a** Calibration of simulation setup with experimental data (Najmzadeh et al. 2012). **b** Proposed Fabrication Flowchart of DMDL–GAA–NW–FET Biosensor

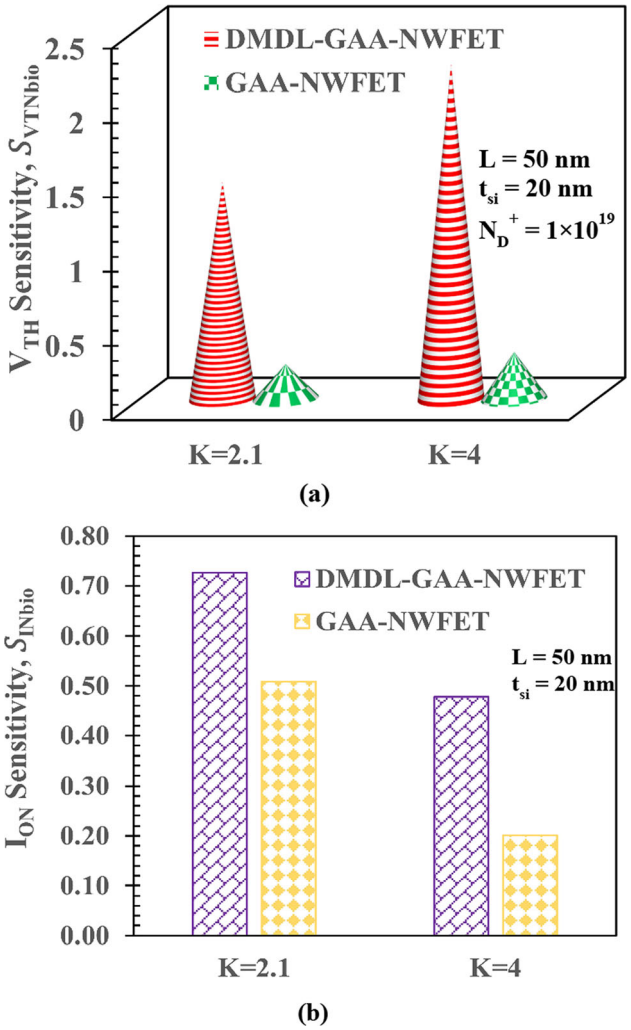


Fig. 4 Sensitivity Comparison of DMDL–GAA–NW–FET and GAA-NW-FET biosensor for Dielectric constant variation of S-Protein (**a**) V_{TH} sensitivity (**b**) I_{ON} current sensitivity

Fig. 4a, revealing that the proposed device is more sensitive to S-Protein than the GAA-NW-FET in terms of threshold voltage sensitivity. The $S_{Vt_{Nbio}}$ is 7.08 times higher than $S_{Vt_{Nbio}}$ of the conventional GAA-NW-FET biosensor for S-protein with K = 4.

The ratio of drain current (I_{ON}) method has been used to calculate the sensitivity of both the devices in terms of I_{ON} current sensitivity (S_{INbio}) for S-Protein. The I_{ON} current is defined as the value of drain current at $V_{GS} = V_{DS} = 1V$. The following mathematical equations have been used to determine the I_{ON} current sensitivity (Kumar et al. 2021):

$$S_{I_{Nbio}} = \frac{I_{ON[with\ S-protein,K=2.1/4]}}{I_{ON[Withwithout\ S-protein]}} \quad (2)$$

I_{ON} sensitivity is plotted for DMDL–GAA–NW–FET and GAA–NW–FET for S-Protein having K = 2.1 and 4 in Fig. 4b. The I_{ON} sensitivity for DMDL–GAA–NW–FET is 2.38 times higher than that of the conventional GAA-NW-FET biosensor for S-protein with dielectric constant K = 4.

The density of the charges carried by the biomolecules affects a biosensor's sensitivity. For negatively charged biomolecules, the channel center potential of the DMDL-GAA-NW-FET decreases because the flat band voltage (V_{FB}) rises due to the negative charge density. In turn, this causes the threshold voltage for negatively charged biomolecules to rise, and vice versa. This change in threshold voltage is utilized to calculate the sensitivity of biosensor. Equation (3) has been used to determine the threshold voltage sensitivity of charged DNA biomolecules for negative and positive charge associated with the biomolecule.

$$S_{Vt_{Cbio}} = \frac{V_{TH[Neutral]} - V_{TH[Charge]}}{V_{TH[Neutral]}} \quad (3)$$

The drain current is decreased for negatively charged biomolecules, whereas it is increased for positive charge density. The reason for this is that the positively charged biomolecules produce a shift in channel potential upward, whereas negatively charged biomolecules cause a shift in channel potential downward (Goel et al. 2021). This shift in the drain current magnitude directly affects the sensitivity of the DMDL-GAA-NW-FET biosensor. Equation (4) has been utilised as the I_{ON} current sensitivity in terms of shift in drain current caused by the insertion of positive and negative charge DNA biomolecules into the nanocavity (Rashid et al. 2022):

$$S_{I_{Cbio}} = I_{ON[Charge]} - I_{ON[Neutral]} \quad (4)$$

As demonstrated in Fig. 5a, b, the DMDL–GAA–NW–FET biosensor outperforms the GAA–NW–FET biosensor for the DNA charged biomolecules detection.

For DNA charged biomolecules, the average V_{TH} sensitivity of DMDL–GAA–NW–FET is 0.19 which is 3.2 times higher than the V_{TH} sensitivity of the conventional GAA–NW–FET biosensor. However, the I_{ON} sensitivity is only 1.68 times greater than that of the conventional GAA-NW-FET biosensor.

The aforementioned discussion leads to the conclusion that the proposed device is more sensitive to S-protein for changing dielectric constants than DNA charged biomolecules. Also, the V_{TH} Sensitivity is dominant over I_{ON} sensitivity.

Table 2 compares the DMDL-GAA-NW-FET biosensor's sensitivities to those of the conventional GAA-NW-FET biosensor for both the S-protein and DNA biomolecules. A common sensitivity parameter is required in order to compare the sensitivity of proposed DMDL-GAA-NW-FET biosensor with that of similar existing biosensors in the recent literature. An effective sensitivity metric known as relative drift in threshold voltage (ΔV_{TH}) aids in determining which biosensor device structure has the highest sensitivity. Table 3 compares the threshold voltage sensitivity (in terms of ΔV_{TH}) of the DMDL-GAA-NW-FET biosensor with that of similar existing biosensors for various biomolecules.

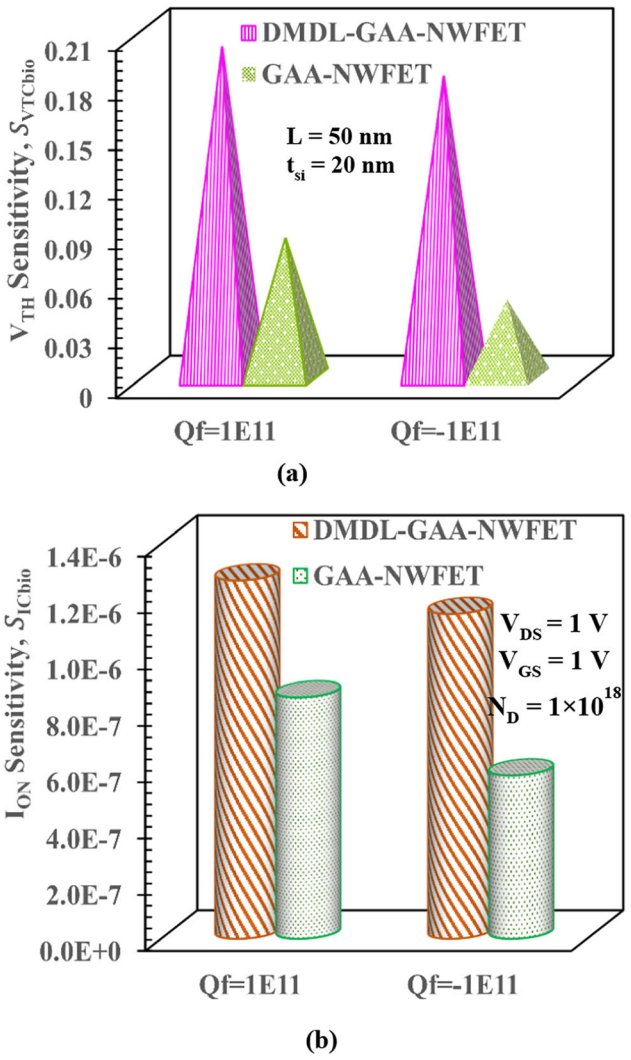


Fig. 5 Sensitivity Comparison of DMDL–GAA–NW–FET and GAA–NW–FET biosensor for charged DNA biomolecules at K = 2.1 (**a**) V_{TH} sensitivity (**b**) I_{ON} current sensitivity

4.2 Device electrostatic analysis

Electrostatics Analysis is crucial to thoroughly characterize and comprehend the electrical behavior of the FET biosensor with varying dielectric constant and charge of the

Table 2 Sensitivity comparison

Sensitivity parameter	DMDL–GAA–NW–FET	Conventional GAA–NW–FET	% improvement
I_{ON} sensitivity for S-protein	0.73 for K = 2.1, 0.48 for K = 4	0.51 for K = 2.1, 0.20 for K = 4	43% for K = 2.1, 140% for K = 4
V_{TH} sensitivity for S-protein	1.46 for K = 2.1, 2.25 for K = 4	0.23 for K = 2.1, 0.32 for K = 4	534.7% for K = 2.1, 603% for K = 4
I_{ON} sensitivity for DNA biomolecules (in μA)	1.27 for Q_f = 1E11, 1.15 for Q_f = − 1E11	0.86 for Q_f = 1E11, 0.58, for Q_f = − 1E11	47.7% for Q_f = 1E11, 98.3% for Q_f = −1E11
V_{TH} sensitivity for DNA biomolecules	0.20 for Q_f = 1E11, 0.18 for Q_f = −1E11	0.08 for Q_f = 1E11, 0.05 for Q_f = − 1E11	150% for Q_f = 1E11, 260%, for Q_f = −1E11

Table 3 Threshold voltage sensitivity comparison of DMDL–GAA–NW–FET to contemporary research on similar biosensors

Structural Parameters	Ref. (Kumari et al. 2022)	Ref. (Sharma et al. 2022)	Ref. (Maiti et al. 2022)	Ref. (Getnet and Chaujar 2022)	Ref. (Das et al. 2023b)	Ref. (Singh and Dutta 2020)	Ref. (Kumar et al. 2020)	Our work
Cavity length (nm)	25	15	20	10	29	20	8	23
Cavity thickness	9	4	9	1	6	10	2	7
Channel length	100	50	40	20	60	40	20	50
(K_{BIO})	5	8	12	2.1	2.63	5	5	4
S_{VTH} in terms of ΔV_{TH} (mV)	130	105	161	17	182	137	87	226

biomolecule to ensure proper functioning of the biosensor. Understanding the electrostatic interactions between the biomolecules and the FET surface provides insights into the sensitivity of the biosensor device and its ability to accurately detect the biomolecules. Also, electrostatic analysis can be used to optimize the design and operation of the device such as the choice of materials, the size and shape of the sensing area (nano cavity), and the operating conditions. Figure 6a depicts the variation of the DMDL–GAA–NW–FET biosensor's center potential along the channel as the dielectric constant for S-Protein is varied. When a bias voltage is provided to the gate electrode, the electrostatic potential at the center of the nanowire channel region is referred to as the channel center potential. The behavior of the charge carriers in the channel is influenced by the channel center potential, which makes it a crucial parameter in biosensing. When the nanogap cavity is devoid of biomolecules, the potential in the channel at source and drain sides are 0.48 V, 0.58 V, and the channel potential minimum (ψc) is 0.03 V. When the S-protein analytes having dielectric constants K = 2.1 and K = 4, are immobilized in the nanocavity, the channel center potential reaches a minimum of − 0.06 V and − 0.16 V respectively, indicating that the channel potential minimum (ψc) is at its maximum value for the case when nanogap cavity has no biomolecules (K = 1).

A rise in the dielectric constant in the cavity region, which increases the vertical electric field and leads to an increase in gate capacitance, is the cause of this drop in ψc (Tayal et al. 2022).

The threshold voltage, subthreshold slope, and other electrical properties of the transistor are all impacted by the decrease in the channel center potential minimum that occurs with an increase in the K value of the biomolecules. The channel center potential minimum reduces when the dielectric constant of biomolecules rises, which also raises the gate capacitance. This decreasesd potential minimum increases the voltage difference between the gate and the channel, which leads to an increase in the threshold voltage. Therefore, the nanogap cavity without biomolecules will reflect the smallest threshold voltage (V_{TH}) and as we increase the K value, the V_{TH} rises as well.

Figure 6b depicts the center potential for positive and negative charge on DNA biomolecules along the channel's position. The center potential is found to be around − 0.067 V for Q_f = 0 Coulomb/cm^2 and − 0.301 V for Q_f = − 5 × 10^{11} C/cm^2 indicating that the channel potential reduces for negative charge analytes. This decrease in center potential is the result of an increase in flat band voltage. This also suggests higher depletion in the channel area, which leads to greater gate control. Also, negatively charged analytes produce an electric field that repels the

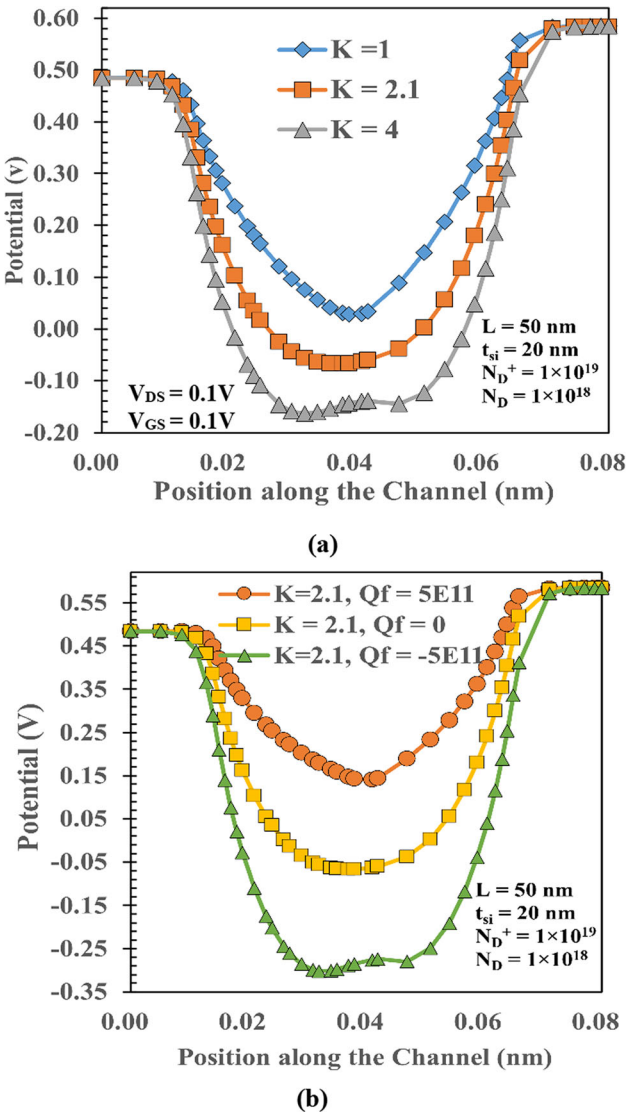


Fig. 6 **a** Channel potential for S-Protein with dielectric constant variations of DMDL–GAA–NW–FET along the channel length). **b** Channel potential for K = 2.1 and varying charge density of DMDL–GAA–NW–FET along the channel length

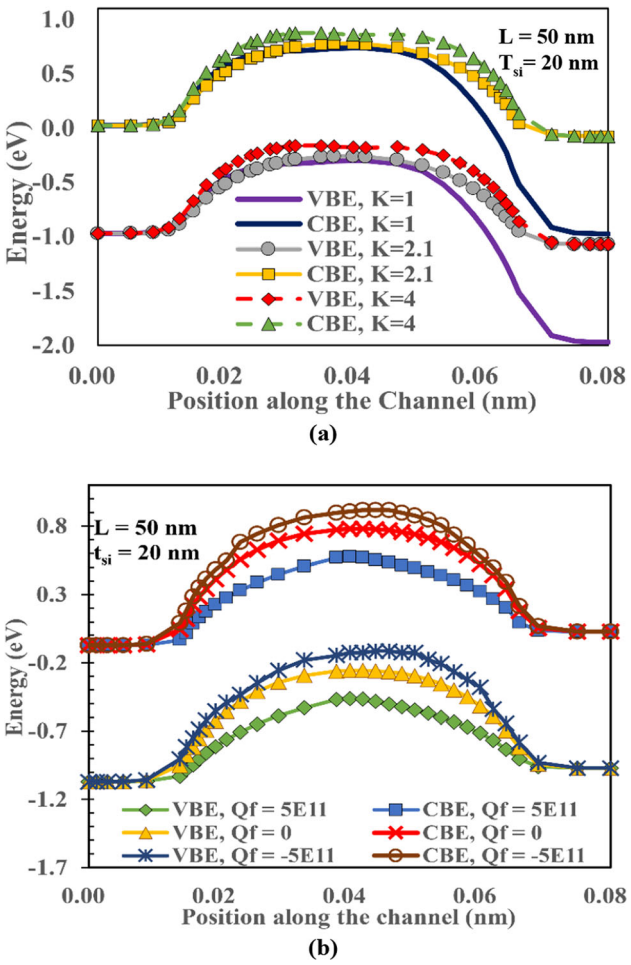


Fig. 7 Energy band diagrams of DMDL–GAA–NW–FET along the channel length (**a**) For S-Protein with dielectric constant variation (**b**) For DNA biomolecule with charge variation

channel's majority carriers (electrons), pushing them out of the sensing region. As a result, the channel's majority carrier concentration declines, which lowers the channel potential. The center potential for positive charged DNA biomolecule ($Q_f = 5 \times 10^{11}$ Coulomb/cm^2) is 0.143 V, showing that the central potential increases for positive charge biomolecules. This is because the charges in the cavity and channel have a stronger attraction. The greater central potential of positive charged DNA biomolecules signifies that the charges in the channel are not depleted, implying that the gate control is weakening. This lowers the threshold voltage even more.

When the dielectric constant value of the biosensor changes, it can affect the electric field distribution in the device. A higher dielectric constant will result in a stronger electric field, while a lower dielectric constant will result in a weaker electric field. This change in the electric field can in turn affect the energy levels of the conduction and valence bands. Overall, changing the dielectric constant value of the biomolecule in a DMDL-GAA-NW-FET biosensor impacts the energy levels of both the conduction and valence bands, affecting the charge carrier behaviour and ultimately the electrical performance of the biosensor. The specific changes in conduction and valence band energies would depend on the biomolecule properties, biosensor device design, and operating conditions of the biosensor. The energy band profiles have been used to examine the impact of channel center potential variation on the carrier transport mechanism. Figure 7a depicts the energy band diagram for both conduction band energy (CBE) and valence band energy (VBE) of DMDL–GAA–NW–FET biosensor with Spike protein analytes having various dielectric constants in the nanogap cavity. It is

evident from Fig. 7a that the energy band barrier at the channel-drain interface is greatly impacted by changes in the dielectric constant of biomolecules in the nanogap, which causes a significant variation in device conduction. S-Protein analytes into the cavity change the channel's gate capacitance and lateral electric field, which marginally boosts the band energies (CBE and VBE).

In contrast to the energy band profile with no charge for a constant dielectric constant K = 2.1, Fig. 7b illustrates the shift in the energy band profile for charged DNA biomolecules. The CBE and VBE levels are observed to shift downward for positively charged ($Q_f = 5 \times 10^{11}$ Coulomb/cm^2) biomolecules and slightly upward for negatively charged (Q_f = -5 $\times$ 10^{11} C/cm^2) biomolecules relative to the level of CBE and VBE for neutral biomolecules ($Q_f = 0$) when these biomolecules are inserted into the nanogap cavity. The potential barrier for the thermionic emission of electrons from the source to channel decreases for positively charged DNA biomolecules and increases for negatively charged DNA biomolecules as a result of this shift in band energy (Tayal et al. 2022). Changing the charge concentration value of the DNA biomolecules alter the position of the Fermi level and, subsequently, the energy required to excite an electron from the valence band to the conduction band. This change in energy can be detected by FET biosensor and used to determine the presence of the DNA biomolecules.

In Fig. 8a, the contours of the electron concentration are shown before (K = 1) and after (K = 2.1 and K = 4) conjugation of the biomolecules. While Fig. 8b depicts the electron concentration contours for positive charge concentration ($Q_f = 5 \times 10^{11}$ Coulomb/cm^2), no charge (Q_f = 0), and negative charge concentration ($Q_f = -5 \times 10^{11}$ Coulomb/cm^2) on DNA biomolecule species having dielectric constant K = 2.1. These electron concentration contours have been plotted at $V_{DS} = V_{GS}$ = 0.1 V. On the right-hand side of the contour plots in Fig. 8a, b, the level of electron concentration is depicted using a color scale.

In a FET biosensor, the concentration of electrons in the channel region is influenced by the dielectric constant and charge of the surrounding medium. The contour plot displays the electron concentration as a function of the dielectric constant, charge and the position along the

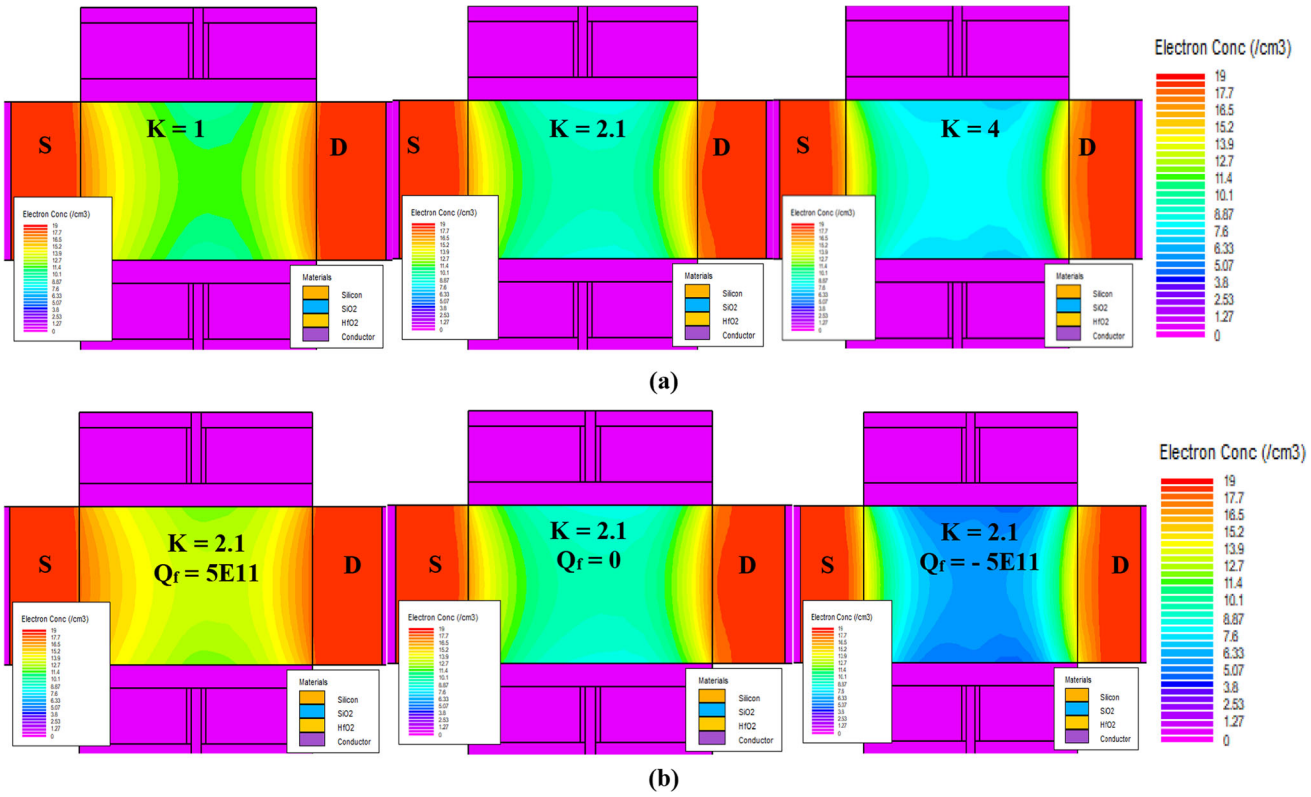


Fig. 8 Countour Plots of electron concentration (**a**) For different biomolecule dielectric constants (**b**) For variation in charge concentration of DNA biomolecule having K = 2.1

channel region. Dielectric constant is a measure of a material's ability to store electrical energy in an electric field. It affects the capacitance of the FET biosensor, which in turn influences the electron concentration at the surface of the transistor. As depicted in Fig. 8a, the electron concentration in the centre of channel becomes lower (as indicated by the colour scale) with the immobilisation of S-Protein (K = 2.1 and K = 4) in the nanocavity, indicating a decrease in the electron concentration. A decrease in electron concentration shifts the threshold voltage, requiring a higher gate voltage to turn on the transistor. The decrease in electron concentration also leads to a fall in current value in the subthreshold region as we increase the value of the dielectric constant.

Charged biomolecules, such as DNA, introduce additional charges at the surface of the FET biosensor, altering the electric field and potential distribution in the channel region. This result in changes in the electron concentration which is reflected in the contour plot in Fig. 8b. According to Fig. 8b, the electron concentration rises when positively charged DNA biomolecules are inserted into the nanocavity region, while it falls when negatively charged biomolecules are inserted, indicating an increase in the current value for positively charged biomolecules and a decrease in the current value for negatively charged biomolecules in subthreshold regime. The threshold voltage is inversely proportional to the electron concentration. As the electron concentration decreases, the threshold voltage tends to increase, and vice versa.

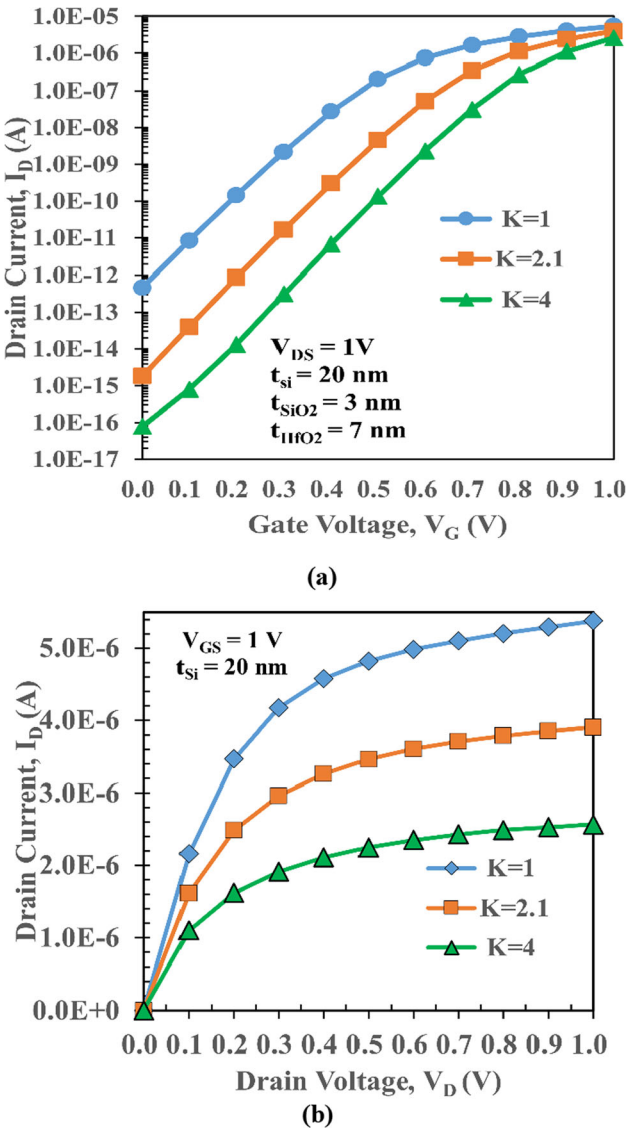


Fig. 9 **a** I_D-V_G characteristics at V_{DS} = 1 and for S-Protein with dielectric constant variations. **b** I_D-V_D characteristics at V_{GS} = 1 V for S-Protein with dielectric constant variations

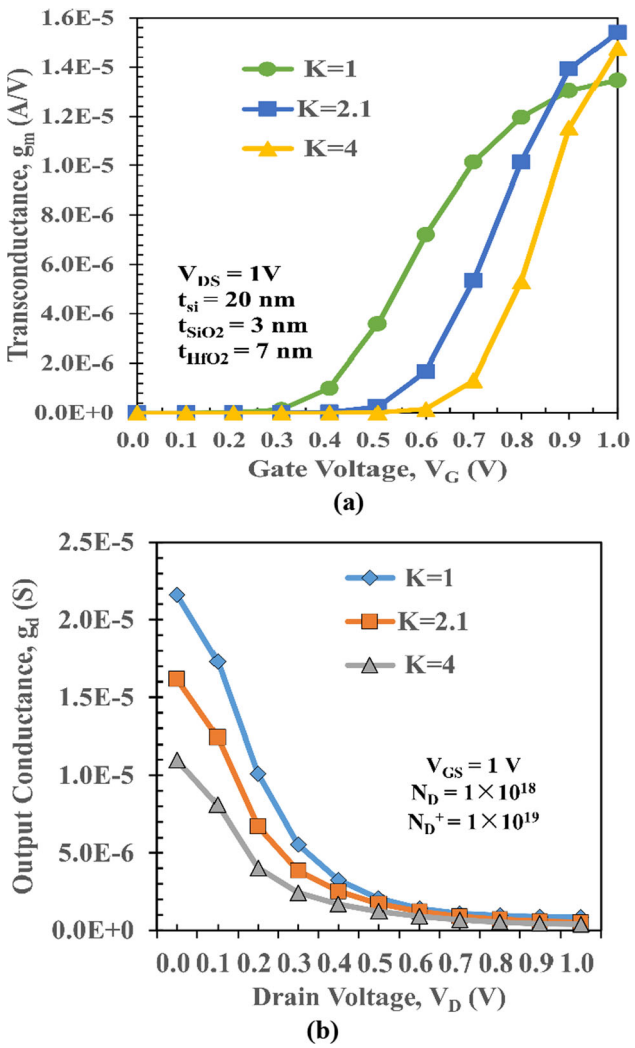


Fig. 10 **a** Transconductance g_m and **b** Output conductance g_d for S-Protein with dielectric constant variations

4.3 Sensitivity analysis of DMDL-GAA-NW-FET biosensor for SARS-CoV-2 spike protein

Figure 9a examines I_D -V_G characteristics (drain current versus gate voltage) of the DMDL-GAA-NW-FET biosensor for dielectric constants, K = 2.1 and K = 4 of S-Protein, assuming that they are neutrally charged. The subthreshold current dramatically decreases from 4.62×10^{-13} to 2.78×10^{-16} A/μm after immobilization of the biomolecules such as air whose K = 1 to K = 4 for the instance of S-Protein. This decrease in drain sub-threshold current is caused by the efficient control of Dual Metal Dual Layer Gate over the channel and the decreased concentration of charge carriers within the channel area when the S-Protein biomolecules having high dielectric constants are immobilized into nanocavity region. This reduction in subthreshold current is accompanied by a drift in the ON current (ΔI_{ON}) as well which is clearly depicted in Fig. 9b showing the drain current versus drain Voltage curve for V_{GS} = 1 V. As we immobilize the biomolecules with increased dielectric constant, they effectively screen the gate electric field, lowering the effective gate voltage that reaches the channel. This reduces the electrostatic potential at the channel surface, therefore decreases the drain current.

Transconductance (g_m) is demonstrated with a change in gate voltage, V_G in Fig. 10a for the proposed design. Transconductance is defined as the ratio of the change in drain current to the change in gate voltage across a pre-set, arbitrarily short interval on the I_{DS} vs V_{GS} curve. Mathematically, g_m is given by Eq. (5)

$$g_m = \left.\frac{\partial I_{DS}}{\partial V_{GS}}\right|_{V_{DS}=constant} \tag{5}$$

A high transconductance value indicates that a slight change in V_G results in a substantial change in I_D. The unique Dual Metal Dual Layer Gate of the DMDL-GAA-NW-FET increases electrostatic coupling and enhances carrier mobility, all of which improve the performance of the proposed device and increases its transconductance. As S-Protein biomolecules are incorporated into the nano cavity, there is a substantial drift in g_m, as illustrated in this Fig. 10a. The highest value of the drift in transconductance is 7.05×10^{-06} A/V which is at V_{GS} = 0.6 V. This variation in g_m is caused by the variation in drain on current on immobilization of S-protein at different dielectric constants.

The impact of S-Protein immobilization on the output conductance (g_d) of the DMDL-GAA-NW-FET biosensor is shown in Fig. 10b. At constant V_{GS} = 1.0 V, the output conductance (g_d) is evaluated by the first order derivative of the drain current with respect to the drain voltage (V_D).

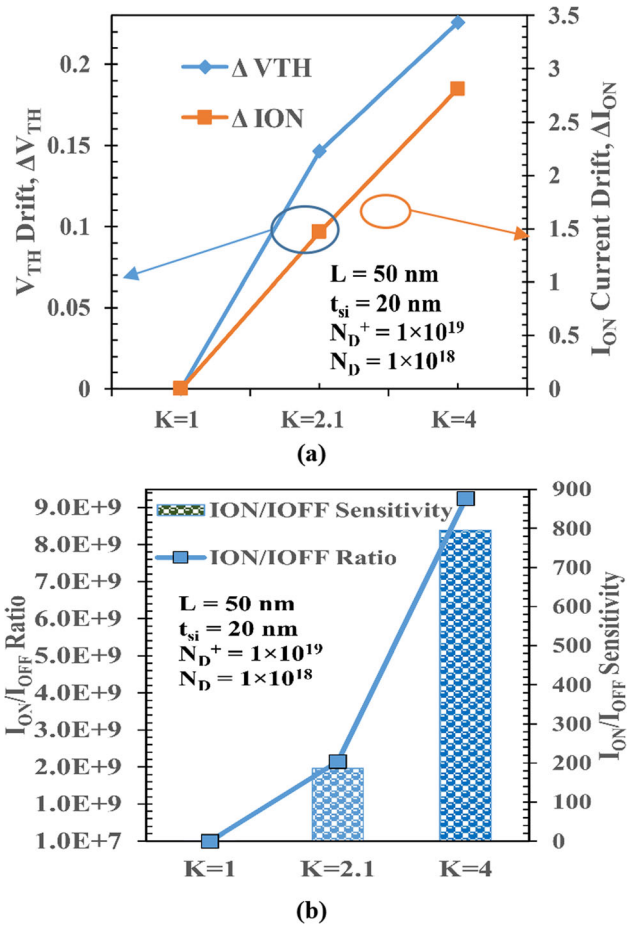


Fig. 11 Sensitivity of S-protein with dielectric constant variation for (**a**) Drift in threshold voltage (ΔV_{TH}), and Drift in Drain ON current (ΔI_{ON}) with respect to air (K = 1) (**b**) I_{ON}/I_{OFF} ratio and I_{ON}/I_{OFF} sensitivity with change in dielectric constant.

According to Fig. 10b, output conductance decreases as S-protein of SARS-CoV-2 virus is injected into the nanocavity compared to the case in which no biomolecule is present. The reason for this drop in output conductance is a decrease in drain current caused by a rise in the dielectric constant of the biomolecules in the nanocavity region.

Figure 11a shows the drift in the threshold voltage (ΔV_{TH}) and drift in I_{ON} drain current (ΔI_{ON}), which are utilized as sensitivity measures of the DMDL–GAA–NW–FET biosensor. The threshold voltage of the proposed biosensor increases with an increase in the dielectric constant of biomolecule in the nano-cavity which can be explained by the fact that the channel centre potential is maximum when there are no biomolecules inside the nanocavity region as shown in Fig. 6a, and that as soon as biomolecules are inserted, the centre potential decreases with an increase in the K value of the biomolecules and subsequently the threshold voltage. Figure 11b shows the variation in I_{ON}/I_{OFF} ratio for different dielectric constants (K values) of S-Protein considered in this work. It is

Table 4 Electrical parameters of DMDL-GAA-NW-FET biosensor for SARS-CoV-2 spike protein

Device electrical parameter	K value	Parameter value	Sensitivity of the parameter
Subthreshold current, I_{OFF} (μA)	K = 1	4.62E-13	2.5×10^2 for K = 2.1
	K = 2.1	1.81E-15	5.93×10^3 for K = 4
	K = 4	7.79E-17	
Drain I_{ON} current (μA)	K = 1	5.38E-06	72.6% for K = 2.1
	K = 2.1	3.91E-06	47.8% for K = 4
	K = 4	2.57E-06	
Threshold voltage drift (ΔV_T) (mV)	K = 2.1	146.4	146.1% for K = 2.1
	K = 4	225.7	225.3% for K = 4
I_{ON}/I_{OFF} current ratio	K = 1	1.16E + 07	1.9×10^2 for K = 2.1
	K = 2.1	2.16E + 09	7.96×10^2 for K = 4

evident from this figure that when the value of the dielectric constant rises, the I_{ON}/I_{OFF} ratio also rises because of the decreased subthreshold current (I_{OFF}) caused by improved DMDL GATE control over the channel, which raises the potential barrier height. Secondary axis of Fig. 11b shows the I_{ON}/ I_{OFF} sensitivity for varying K values of S-Protein. The ratio of I_{ON}/I_{OFF} current at various S-protein dielectric constants and the I_{ON}/I_{OFF} current at K = 1 have been used to compute the I_{ON}/I_{OFF} sensitivity illustrated in Fig. 11b. The value of I_{ON}/I_{OFF} sensitivity for K = 4 of S-Protein is 795.63.

Table 4 summarises the various device electrical parameters on the immobilisation of S-protein with changing dielectric constants within the nanocavity region.

4.4 Sensitivity investigation of DNA charge density modulation/trap charge concentration

The reverse transcriptase process transforms the virus's RNA into DNA, enabling the detection of the virus through its DNA. A biosensor's sensitivity can be changed based on the density of the charges carried by the DNA biomolecules. This section examines the reactions of the proposed biosensor to charged DNA immobilized in the nanocavity regions. The positive and negative charges taken into account for the investigation are -1×10^{11}, -5×10^{11}, 1×10^{11}, and 5×10^{11} cm^{-2} and the DNA biomolecule's fixed dielectric constant is 2.1. Consideration of fixed oxide interface trapped charges ($Q_f = \pm 1 \times 10^{11}$ and $\pm 5 \times 10^{11}$) in the oxide-Air interface can be used to illustrate the effects of the presence of a charged biomolecule in the nanogap cavity. The flat band voltage (V_{FB}) rises due to the negative trap charge (and vice versa) which results in a decrease in the channel centre potential of the DMDL–GAA–NW–FET biosensor. The concept is similar to a well-known approach used to model the interface trapped charges present in the oxide–semiconductor interface in MOSFETs (Kinsella and Ivanisevic 2007; Chakraborty and Sarkar 2017).

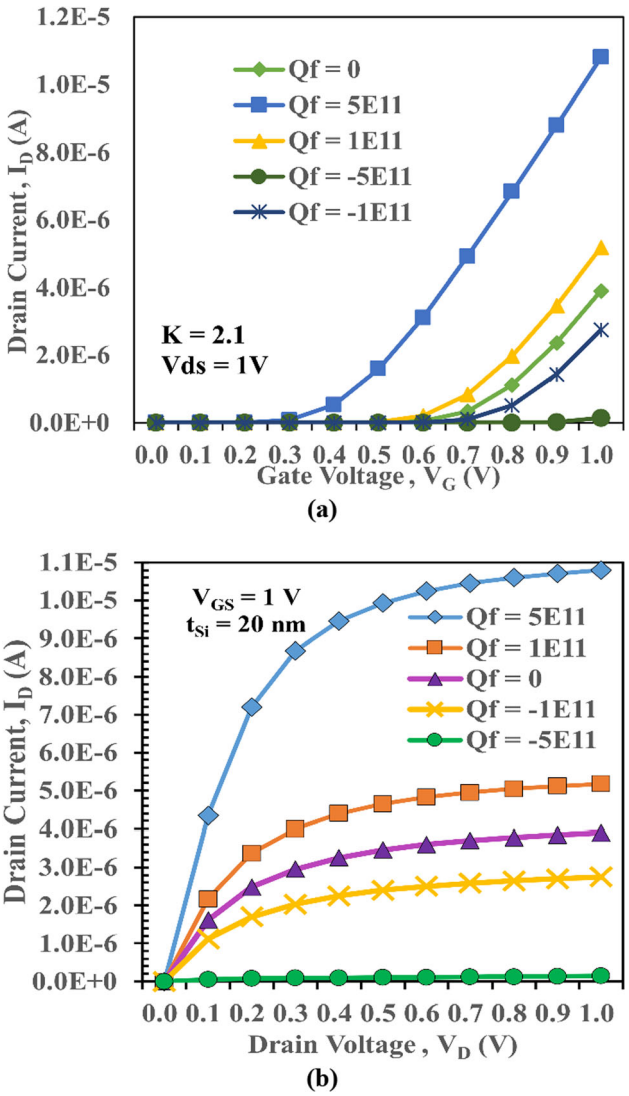


Fig. 12 **a** I_D-V_G characteristics at V_{DS} = 1 V and **b** I_D–V_D characteristics at V_{GS} = 1 V for DNA (K = 2.1) charged biomolecules

Figure 12 (a) depicts the change in I_D with respect to V_G for different negative and positive charge densities of C-DNA. These I_D–V_G characteristics are displayed on the assumption that DNA biomolecules with K = 2.1 completely occupy the nanocavities. The drain current is shown to be falling as the negative charged density increases, whereas it is found to be rising when the positive charge density increases. The reason for this is that positively charged biomolecules produce a shift in channel potential upward, whereas negatively charged biomolecules cause a shift in channel potential downward. And this shift in the drain current magnitude directly affects the sensitivity of the DMDL-GAA-NW-FET biosensor.

The drain current against drain voltage (I_D–V_D curve) for the proposed biosensor is shown in Fig. 12b with the variation in charge of DNA biomolecules. The Fig. 12b clearly shows that the drain current continued to rise as the charges on biomolecules shifted from negative to positive.

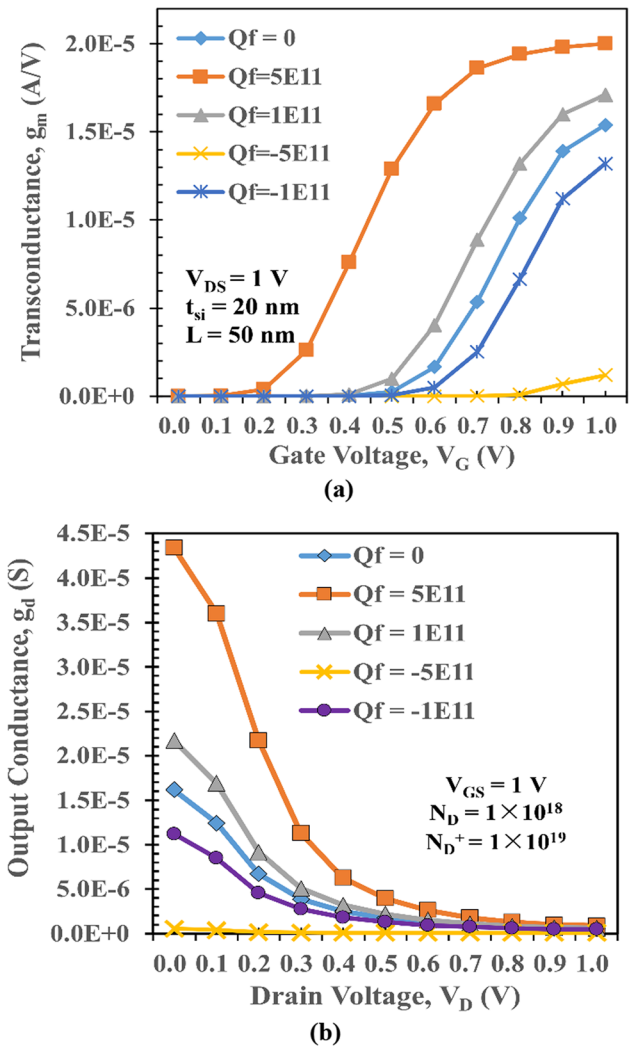


Fig. 13 **a** Transconductance, g_m and **b** Output conductance, g_d for DNA (K = 2.1) charged biomolecules

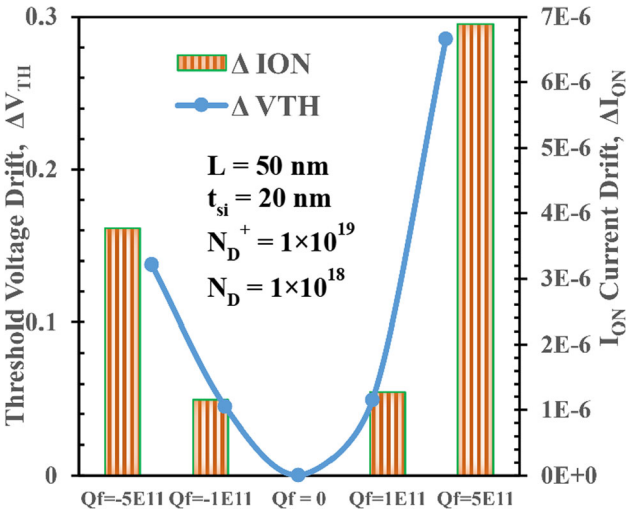


Fig. 14 Drift in threshold voltage (ΔV_{TH}), and Drift in Drain ON current (ΔI_{ON}) for charged DNA biomolecules at constant K = 2.1

For biomolecules with positive charges, the positive charge draws electrons to the surface and reduces the space charge region. This leads to more carriers entering the channel, which lowers resistance and boosts drain current. Also, negative charges on biomolecules deflect electrons away from the surface, broadening the space charge region of the channel and lowering drain current.

The transconductance (g_m) curves of the proposed biosensor with biomolecule's positive and negative charge density variation are demonstrated in Fig. 13a. With rising positive charge density, peak transconductance's V_{GS} value falls. While, when negative charge density increases, the V_{GS} value of the peak transconductance increases. This is the cause of a decrease in the threshold voltage for positively charged biomolecules and an increase in the V_{TH} for negatively charged biomolecules.

The variation in output conductance (g_d) with variation in charge on the DNA biomolecule of the SARS-COV-2 virus is shown in Fig. 13b. The increase in positive charge increases the output conductance, while increasing the negative charge decreases the g_d value.)

The impact of positive and negative charges on the DNA biomolecule with a fixed dielectric constant of 2.1 is summarised in Fig. 14. With the immobilization of positive and negative DNA biomolecules into the nanocavity, the variation in drain current and transconductance (which in turn alters the threshold voltage) as shown in Figs. 12 and 13 can be translated into drift in I_{ON} drain current (ΔI_{ON}) and drift in threshold voltage (ΔV_{TH}). ΔV_{TH} and ΔI_{ON} are calculated as the difference of V_{TH} and I_{ON} at $Q_f = 0$ and V_{TH} and I_{ON} at $Q_f = 1 \times 10^{11}$ cm^{-2}, -5×10^{11} cm^{-2}, 1×10^{11} cm^{-2}, and 5×10^{11} cm^{-2}. It is clear from Fig. 14 that proposed biosensor device exhibits a significant drift in the drain current and threshold voltage with the

Table 5 Electrical parameters of DMDL–GAA–NW–FET biosensor for DNA biomolecules of SARS-CoV-2

Device electrical parameter	Charge value	Parameter value	Sensitivity of the parameter
Subthreshold current, I_{OFF} (μA)	Qf = 0	1.81E-15	6.5×10^3 for $Q_f = 5 \times 10^{11}$, 0.42 for $Q_f = -5 \times 10^{11}$
	Qf = 5×10^{11}	1.18E-11	
	Qf = -5×10^{11}	7.77E-16	
Drain I_{ON} current (μA)	Qf = 0	3.91E-06	2.7×10^2% for $Q_f = 5 \times 10^{11}$
	Qf = 5×10^{11}	1.08E-05	3.56% for $Q_f = -5 \times 10^{11}$
	Qf = -5×10^{11}	1.39E-07	
Threshold voltage drift (ΔV_T) (mV)	Qf = 5×10^{11}	285.5	50.54% for $Q_f = 5 \times 10^{11}$, 56.81% for $Q_f = -5 \times 10^{11}$
	Qf = -5×10^{11}	137.5	

change in charged biomolecules which directly affects the sensitivity of the biosensor. Furthermore, as shown in Fig. 14, the DMDL–GAA–NW–FET biosensor exhibits larger V_{TH} and I_{ON} drift for positively charged DNA biomolecules than for negatively charged DNA biomolecules.

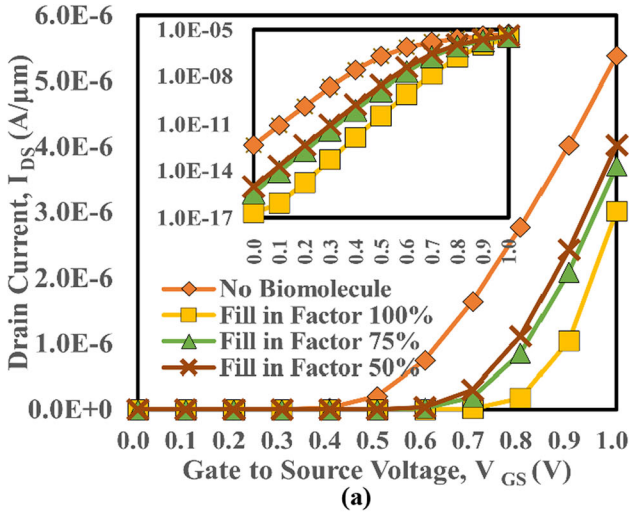
The several electrical parameters that are sensitive to the immobilisation of the SARS-CoV-2 virus' DNA biomolecules in the nanocavity region are listed in Table 5.

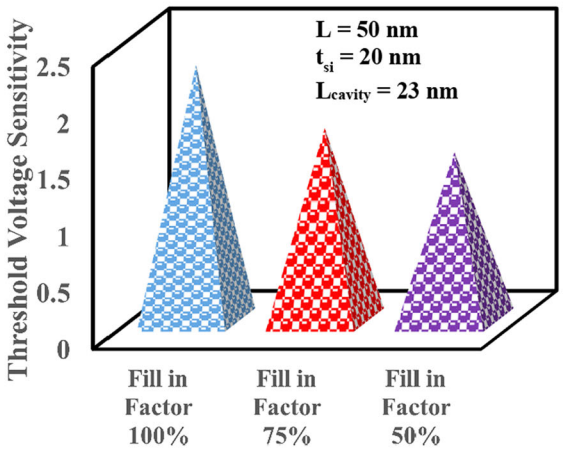
4.5 Impact of fill in factor on sensitivity

Biomolecules congregate in the nanogap at random, and fully filled nanogaps are rarely observed. Biomolecules that are partially filled have different surface profiles within the nanogaps and give varied capacitances as a result. Figure 15a illustrates three distinct scenarios for drain current to biomolecule fill in area in the nanocavity for biomolecule dielectric constant K = 4. The drain ON current and OFF current increases with decreasing fill in factor of the biomolecules as depicted in Fig. 15a. The OFF current is shown as inset in Fig. 15 (a). The threshold voltage sensitivity calculated using Eq. (1), decreases with decreasing fill in factor as shown in Fig. 15b. Similarly, the I_{ON} and I_{OFF} current sensitivities decreases with decreasing fill in factor but this decrement is more prominent for the case of I_{OFF} current sensitivity as shown in Fig. 15c, d. Therefore, when the fill-in factor is high, the device is more sensitive to biomolecules for all the sensitivity parameters.

5 Conclusion

In an effort to combat COVID-19, we successfully designed a biosensor that uses Dual Metal Dual Layer Gate-All-Around Nanowire FET for detection of SARS-CoV-2 Spike protein and DNA. The manuscript demonstrates that the Dual Metal and Dual Layered Gate All around structure has a persuasive effect on the sensing ability of the proposed biosensor. The combination of layered structure and dual metal work function engineering allows greater Gate control over the channel and improved analog performance. Comparative sensitivity analysis of the proposed DMDL–GAA–NW–FET biosensor device with normal Gate All Around NWFET device demonstrates the significance of the proposed Gate structural alteration in GAA-NW-FET. The proposed device's sensitivity to S-protein and DNA biomolecules of the SARS-CoV-2 virus has been evaluated in terms of drain ON current, subthreshold current, I_{ON}/I_{OFF} ratio, g_m, and g_d. The greater variation in these electrical parameters enables the development of biosensor that is incredibly sensitive. The DMDL-GAA-NW-FET achieves a V_{TH} shift of 225.74 mV as compare to GAA-NW-FET biosensor which produces a V_{TH} shift of 93.56 mV while varying the dielectric constant (K) from 1 to 4. The simulation findings show that DMDL-GAA-NW-FET is 7 times more sensitive to threshold voltage and 2.4 times more sensitive to I_{ON} current than GAA-NW-FET when it comes to immobilizing biomolecules with K = 4 inside the cavity. Furthermore, the proposed sensor's sensitivity for variation in DNA charge density has been investigated and concluded that as the charge on biomolecules changes from negative to positive, sensitivity increases. Therefore, the proposed DMDL-GAA-NW-FET biosensor is more applicable for the detection of the SARS-CoV-2 virus. The proposed biosensor design is responding better in term of sensitivity for different neutral and charged biomolecules so it can also be used for detecting other biological diseases like breast cancer, Alzheimer, Ebola and various other viral infections.





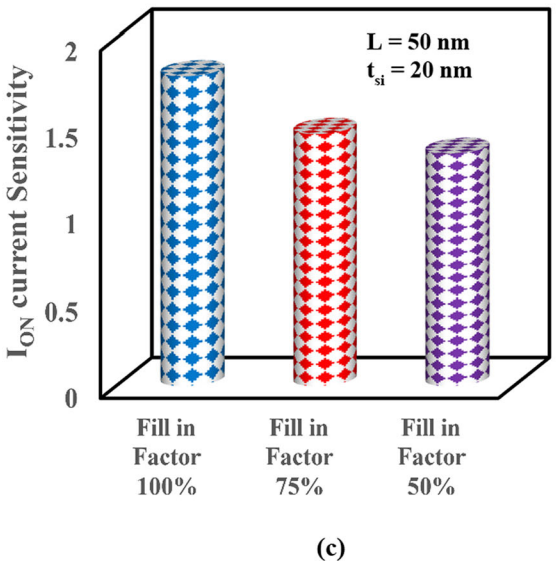


Fig. 15 **a** Drain Current versus Gate to Source Voltage, **b** Threshold Voltage Sensitivity, **c** I_{ON} Current Sensitivity, **d** I_{OFF} current Sensitivity for various fill in factor of the biomolecules into the nanocavity

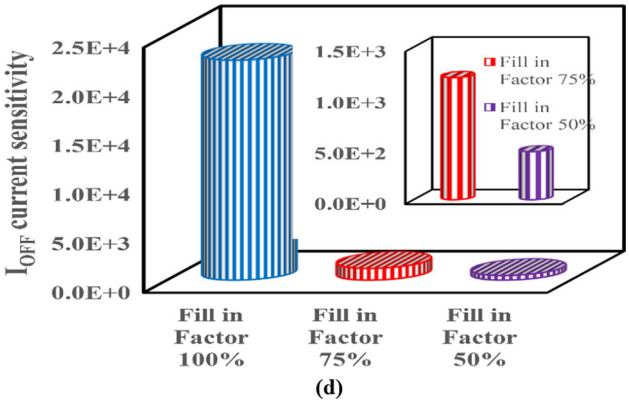


Fig. 15 continued

Funding The authors have not received any funding for this research.

Data availability Not Applicable.

Declarations

Conflict of interest Authors declare that there are no conflicts of interests.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Chakraborty A, Sarkar A (2017) Analytical modeling and sensitivity analysis of dielectric-modulated junctionless gate stack surrounding gate MOSFET (JLGSSRG) for application as biosensor. J Comput Electron 16:556–567. https://doi.org/10.1007/s10825-017-0999-2

Chen M, Cui D, Zhao Z et al (2022) Highly sensitive, scalable, and rapid SARS-CoV-2 biosensor based on In2O3 nanoribbon transistors and phosphatase. Nano Res 15:5510–5516. https://doi.org/10.1007/s12274-022-4190-0

Chen Y, Ren R, Pu H, et al (2017) Field-Effect transistor biosensor for rapid detection of ebola antigen. Sci Rep 7:. https://doi.org/10.1038/s41598-017-11387-7

Cucinotta D, Vanelli M (2020) WHO declares COVID-19 a pandemic. Acta Biomed 91:157–160. https://doi.org/10.23750/abm.v91i1.9397

Das A, Kanaujia BK, Deswal SS et al (2022a) Doping induced threshold voltage and ION/IOFF ratio modulation in surrounding gate MOSFET for analog applications. 2022 IEEE international conference of electron devices society kolkata chapter (EDKCON). Kolkata, India, pp 1–6

Das A, Rewari S, Kanaujia BK, Gupta RS (2022b) Recent technological advancement in surrounding gate MOSFET for biosensing applications—a synoptic study. SILICON 14:5133–5143. https://doi.org/10.1007/s12633-021-01288-w

Das A, Rewari S, Kanaujia BK et al (2023a) Numerical modeling of a dielectric modulated surrounding-triple-gate germanium-source MOSFET (DM-STGGS-MOSFET)-based biosensor. J Comput Electron. https://doi.org/10.1007/s10825-023-02008-w

Das A, Rewari S, Kanaujia BK et al (2023b) Analytical investigation of a triple surrounding gate germanium source metal–oxide–semiconductor field-effect transistor with step graded channel for biosensing applications. Int J Numer Model Electron Networks, Devices Fields. https://doi.org/10.1002/jnm.3106

Das A, Kanaujia BK, Nath V, Rewari S (2020) Impact of reverse gate oxide stacking on gate all around tunnel FET for high frequency analog and rf applications. https://doi.org/10.1109/INDICON49873.2020.9342175

Dastjerdi S, Malikan M, Akgöz B, et al (2022) On the deformation and frequency analyses of SARS-CoV-2 at nanoscale. Int J Eng Sci 170:. https://doi.org/10.1016/j.ijengsci.2021.103604

Dixit A, Samajdar DP, Bagga N (2021) Dielectric modulated GaAs1-xSb XFinFET as a label-free biosensor: Device proposal and investigation. Semicond Sci Technol 36:. https://doi.org/10.1088/1361-6641/ac0d97

Fathi-Hafshejani P, Azam N, Wang L et al (2021) Two-dimensional-material-based field-effect transistor biosensor for detecting COVID-19 Virus (SARS-CoV-2). ACS Nano 15:11461–11469. https://doi.org/10.1021/acsnano.1c01188

Getnet M, Chaujar R (2022) Sensitivity analysis of biomolecule nanocavity immobilization in a dielectric modulated triple-hybrid metal gate-all-around junctionless NWFET biosensor for detecting various diseases. J Electron Mater 51:2236–2247. https://doi.org/10.1007/s11664-022-09466-1

Gilson MK, Honig BH (1986) The dielectric constant of a folded protein. Biopolymers 25:2097–2119. https://doi.org/10.1002/bip.360251106

Goel A, Rewari S, Verma S, Gupta RS (2020) High-K spacer dual-metal gate stack underlap junctionless gate all around (HK-DMGS-JGAA) MOSFET for high frequency applications. Microsyst Technol 26:1697–1705. https://doi.org/10.1007/s00542-019-04715-6

Goel A, Rewari S, Verma S et al (2021) Dielectric modulated junctionless biotube FET (DM-JL-BT-FET) bio-sensor. IEEE Sens J 21:16731–16743. https://doi.org/10.1109/JSEN.2021.3077540

Gorbalenya et al (2020) Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndromerelated coronavirus

Gu B, Park TJ, Ahn JH et al (2009) Nanogap field-effect transistor biosensors for electrical detection of avian influenza. Small 5:2407–2412. https://doi.org/10.1002/smll.200900450

Han K, Zhang Y, Deng Z (2020) A simulation study of gate-all-around nanowire transistor with a core-substrate. IEEE Access 8:62181–62190. https://doi.org/10.1109/ACCESS.2020.2983724

Hardenbrook NJ, Zhang P (2022) A structural view of the SARS-CoV-2 virus and its assembly. Curr Opin Virol 52:123–134. https://doi.org/10.1016/j.coviro.2021.11.011

Im H, Huang XJ, Gu B, Choi YK (2007) A dielectric-modulated field-effect transistor for biosensing. Nat Nanotechnol 2:430–434. https://doi.org/10.1038/nnano.2007.180

Ishikawa FN, Chang HK, Curreli M et al (2009) Label-free, electrical detection of the SARS virus n-protein with nanowire biosensors utilizing antibody mimics as capture probes. ACS Nano 3:1219–1224. https://doi.org/10.1021/nn900086c

Kasturi P, Saxena M, Gupta RS (2005) Modeling and simulation of STacked Gate Oxide (STGO) architecture in Silicon-On-Nothing (SON) MOSFET. Solid State Electron 49:1639–1648. https://doi.org/10.1016/j.sse.2005.07.004

Kaul A, Rewari S, Nand D (2022) Macaroni Channel-Nanowire-Field Effect Transistor (MC-NW-FET) for Gate Induced Drain Leakage (GIDL) Reduction Application. In: Proceedings of 2022 IEEE International Conference of Electron Devices Society Kolkata Chapter, EDKCON 2022. pp 35–38

Kinsella JM, Ivanisevic A (2007) Biosensing: taking charge of biomolecules. Nat Nanotechnol 2:596–597. https://doi.org/10.1038/nnano.2007.313

Kumar A, Gupta N, Tripathi MM, Chaujar R (2020) Analysis of structural parameters on sensitivity of black phosphorus junctionless recessed channel MOSFET for biosensing application. Microsyst Technol 26:2227–2233. https://doi.org/10.1007/s00542-019-04545-6

Kumar S, Singh B, Singh Y (2021) Analytical Model of Dielectric Modulated Trench Double Gate Junctionless FET for Biosensing Applications. IEEE Sens J 21:8896–8902. https://doi.org/10.1109/JSEN.2021.3056385

Kumari M, Singh NK, Sahoo M, Rahaman H (2022) 2-D analytical modeling and simulation of dual material, double gate, gate stack engineered, junctionless MOSFET based biosensor with enhanced sensitivity. SILICON 14:4473–4484. https://doi.org/10.1007/s12633-021-01223-z

Li C, Liu F, Han R, Zhuang Y (2021) A Vertically stacked nanosheet gate-all-around FET for biosensing application. IEEE Access 9:63602–63610. https://doi.org/10.1109/ACCESS.2021.3074906

Liu G, Rusling JF (2021) COVID-19 Antibody Tests and Their Limitations. ACS Sensors 6:593–612. https://doi.org/10.1021/acssensors.0c02621

Lou H, Member S, Zhang L et al (2012) A junctionless nanowire transistor with a dual-material gate. IEEE Trans Electron Devices 59:1829–1836

Lu R, X Z, J L, et al (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet https://ww:1–3

Maiti S, De A, Sarkar SK (2022) Analytical modelling of symmetric gate underlap quadruple gate multichannel junctionless MOSFET biosensor. SILICON 14:6921–6932. https://doi.org/10.1007/s12633-021-01415-7

Nag P, Sadani K, Mukherji S et al (2020) Diagnosing COVID-19: the disease and tools for detection. ACS Nano 14:5135–5142

Najmzadeh M, Bouvet D, Grabinski W et al (2012) Accumulation-mode gate-all-around si nanowire nMOSFETs with sub-5 nm cross-section and high uniaxial tensile strain. Solid State Electron 74:114–120. https://doi.org/10.1016/j.sse.2012.04.021

Pal A, Sarkar A (2014) Analytical study of dual material surrounding gate MOSFET to suppress short-channel effects (SCEs). Eng Sci Technol an Int J 17:205–212. https://doi.org/10.1016/j.jestch.2014.06.002

Pelletier AN, Sekaly RP, Tomalka JA (2022) Translating known drivers of COVID-19 disease severity to design better SARS-CoV-2 vaccines. Curr Opin Virol 52:89–101. https://doi.org/10.1016/j.coviro.2021.11.012

Priyadarshani KN, Singh S, Mohammed MKA (2022) Gate-all-around junctionless FET based label-free dielectric/charge modulation detection of SARS-CoV-2 virus. RSC Adv 12:9202–9209. https://doi.org/10.1039/d1ra08587e

Prof Nanshan Chen M, Prof Min Zhou M, Xuan Dong P et al (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 395:507–513

Quest Diagnostics (2020) SARS-CoV-2 RNA, Qualitative Real-Time RT-PCR. 1–27

Rashid S, Bashir F, Khanday FA, Beigh MR (2022) Dielectrically modulated III-V compound semiconductor based pocket doped tunnel FET for label free biosensing applications. IEEE Trans Nanobioscience. https://doi.org/10.1109/TNB.2022.3178763

Rony MW, Bhowmik P, Myler HR, Mondol P (2017) Short channel effects suppression in a dual-gate gate-all-around Si nanowire junctionless nMOSFET. In: Proceedings of 9th international conference on electrical and computer engineering, ICECE 2016. pp 538–541

Sarkar CK (2016) Technology computer aided design: Simulation for VLSI MOSFET

Seo G, Lee G, Kim MJ et al (2020) Rapid detection of COVID-19 causative virus (SARS-CoV-2) in human nasopharyngeal swab specimens using field-effect transistor-based biosensor. ACS Nano 14:5135–5142

Shao W, Shurin MR, Wheeler SE et al (2021) Rapid detection of SARS-CoV-2 antigens using high-purity semiconducting single-walled carbon nanotube-based field-effect transistors. ACS Appl Mater Interfaces 13:10321–10327. https://doi.org/10.1021/acsami.0c22589

Sharma S, Nath V, Deswal SS, Gupta RS (2022) Analytical modelling and sensitivity analysis of Gallium Nitride-Gate Material and, dielectric engineered- Schottky nano-wire fet(GaN-GME-DE-SNW-fet) based label-free biosensor. Microelectronics J 129:. https://doi.org/10.1016/j.mejo.2022.105599

Singh KN, Dutta PK (2020) Analytical modeling of a high-K underlap dielectric- and charge-modulated silicon-on-nothing FET-based biosensor. J Comput Electron 19:1126–1135. https://doi.org/10.1007/s10825-020-01511-8

Singh S, Raman A (2018) Gate-all-around charge plasma-based dual material gate-stack nanowire FET for enhanced analog performance. IEEE Trans Electron Devices 65:3026–3032. https://doi.org/10.1109/TED.2018.2816898

Talley K, Ng C, Shoppell M, et al (2008) On the electrostatic component of protein-protein binding free energy. PMC Biophys 1:. https://doi.org/10.1186/1757-5036-1-2

Tayal S, Majumdar B, Bhattacharya S, Kanungo S (2022) Performance analysis of the dielectrically modulated junction-less nanotube field effect transistor for biomolecule detection. IEEE Trans Nanobioscience. https://doi.org/10.1109/TNB.2022.3172702

Warsame C, Valerini D, Llavori I, et al (2022) Modal analysis of novel coronavirus (SARS COV-2) using finite element methodology. J Mech Behav Biomed Mater 135:. https://doi.org/10.1016/j.jmbbm.2022.105406

Weissleder R, Lee H, Ko J, Pittet MJ (2020) COVID-19 diagnostics in context. Sci Transl Med 12:. https://doi.org/10.1126/scitranslmed.abc1931

Wierzbicki T, Li W, Liu Y, Zhu J (2021) Effect of receptors on the resonant and transient harmonic vibrations of Coronavirus. J Mech Phys Solids 150:. https://doi.org/10.1016/j.jmps.2021.104369

Wu C rong, Yin W chao, Jiang Y, Xu HE (2022) Structure genomics of SARS-CoV-2 and its Omicron variant: drug design templates for COVID-19. Acta Pharmacol Sin 43:3021–3033. https://doi.org/10.1038/s41401-021-00851-w

Xie Y, Du D, Karki CB et al (2020) Revealing the Mechanism of SARS-CoV-2 Spike Protein Binding with ACE2. Comput Sci Eng 22:21–29. https://doi.org/10.1109/MCSE.2020.3015511

Yadav S, Rewari S (2022) Trench Gate JAM dielectric modulated nanowire FET (TG-JAM-DM-NWFET) Biosensor. Proc 2022 IEEE Int Conf Electron Devices Soc Kolkata Chapter. EDKCON 2022:23–28. https://doi.org/10.1109/EDKCON56221.2022.10032912

Yang H, Rao Z (2021) Structural biology of SARS-CoV-2 and implications for therapeutic development. Nat Rev Microbiol 19:685–700. https://doi.org/10.1038/s41579-021-00630-8

Ye S, Yamabe K, Endoh T (2021) Ultimate vertical gate-all-around metal–oxide–semiconductor field-effect transistor and its three-dimensional integrated circuits. Mater Sci Semicond Process 134:. https://doi.org/10.1016/j.mssp.2021.106046

Zhu N, Zhang D, Wang W et al (2020) A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382:727–733. https://doi.org/10.1056/nejmoa2001017

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.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

eDRAM-CIM: Reconfigurable Charge Domain Compute-In-Memory Design With Embedded Dynamic Random Access Memory Array Realizing Adaptive Data Converters

Shanshan Xie, *Graduate Student Member, IEEE*, Can Ni, Aseem Sayal, Pulkit Jain, *Senior Member, IEEE*, Fatih Hamzaoglu, *Senior Member, IEEE*, and Jaydeep P. Kulkarni, *Senior Member, IEEE*

***Abstract*— This article presents a compute-in-memory (CIM) architecture for a large-scale machine learning (ML) accelerator, which employs 1*T*1*C* embedded dynamic random access memory (eDRAM) bitcells as charge domain circuits for convolution neural network (CNN) multiply-accumulation-averaging (MAV) computations. By repurposing existing 1*T*1*C* eDRAM columns to construct an adaptive data converter, dot-product, averaging, pooling, and rectified linear unit (ReLU) activation on the memory array, the eDRAM-CIM design eliminates the need for an extra dedicated hardware accelerator, which significantly reduces the hardware implementation cost and increases the reconfigurability of the CIM computation circuit. In the eDRAM-CIM design, 8 b digital inputs from image pixel values are converted to analog domain directly on eDRAM columns. The dot-product is computed without disturbing the kernel weights, which are stored in the eDRAM array, preventing data duplication and additional intra-memory data movement. In addition, the convolution results are transferred back to the digital domain using an in-eDRAM adaptive dynamic range successive-approximation (SAR) analog-to-digital converter (ADC) utilizing a narrow range of dot-product distribution to reduce ADC latency and energy. A 16 Kb eDRAM-CIM prototype, implemented in the 65 nm CMOS process, demonstrates the concept and the functionality of the test-chip using the CIFAR-10 dataset with 8 b input and 8 b signed weight, achieving 90.77% accuracy, 4.71 GOPS throughput, 4.76 TOPS/W energy efficiency, and 8.26 GOPS/mm^2. A scalability analysis is developed to show that the presented eDRAM-CIM approach, which was adopted in an advanced eDRAM technology node, shows promising energy efficiency (11.21 TOPS/W) and throughput (22.2 TOPS), suggesting its potential for application in large-scale energy-efficient CIM designs.**

***Index Terms*— Analog computation, compute-in-memory (CIM), dot-product, embedded dynamic random access memory (eDRAM), in-memory computation, machine learning (ML), ML accelerator.**

Manuscript received 17 March 2023; revised 21 June 2023, 29 August 2023, and 15 October 2023; accepted 16 October 2023. This article was approved by Associate Editor Kathryn Wilcox. This work was supported by the Intel Corporation. *(Corresponding author: Shanshan Xie.)*

The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA, also with Google LLC, Sunnyvale, CA 94089 USA, and also with the Intel Circuit Research Laboratory, Hillsboro, OR 97124 USA (e-mail: sxie@utexas.edu; can5@utexas.edu; jaydeep@austin.utexas.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSSC.2023.3326094.

Digital Object Identifier 10.1109/JSSC.2023.3326094

I. Introduction

MACHINE learning (ML) brings together data and computer science to provide solutions for different applications, such as image classification [1], face verification [2], and speech recognition [3]. To be more specific, convolution neural networks (CNNs) have been extensively explored and applied as a deep neural network (DNN) approach to handle complicated ML problems [4], [5]. However, due to the growth in the ML area, the complexity, and model size of DNN are also experiencing unprecedented growth. For example, a basic network, like LeNET-5 requires a 341 k number of MAC operations, while a more sophisticated network requires a 15.5 G number of MAC operations [6]. And this results in a massive amount of data movement, data transfer latency, and computation/memory access energy cost from OFF-chip memory to the ON-chip processing cores to complete the required MAC computations in modern large-scale ML accelerators. As shown in Table I, a 32-bit integer multiply takes only 3.1 pJ energy, while DRAM memory access can consume 2000 pJ energy [6], which is 650× larger than the multiply computation energy cost. Therefore, significant research is needed to reduce memory access.

To mitigate this "Memory Wall" bottleneck of latency and energy overheads, compute-in-memory (CIM) designs are being explored in different memory technologies. The CIM designs leverage the storage elements in the memory array and embed the computations inside the memory array to reduce the OFF-chip data movement between the central processing unit (CPU) and the memory block. Among different memory technologies, embedded nonvolatile magnetic [7], [8], [9], [10], and resistive [11], [12], [13], [14] as well as standalone flash memories suffer from poor endurance and lower speeds and cannot be used for programmable DNNs requiring fast and frequent model updates. Similarly, commodity dynamic random access memory (DRAM) cannot be leveraged for high-speed and custom CIM designs due to cost-sensitive processes, limited metal layers, and dense floorplan requirements. These constraints often lead to compute-near-memory designs, which have limited throughput benefits [15]. Among the prevalent semiconductor memories, embedded-DRAM (eDRAM) which integrates the DRAM bitcell monolithically along with high-performance logic transistors and

0018-9200 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
COMPUTATION ENERGY COST AND MEMORY ACCESS ENERGY COST [6]

Operation	Energy(pJ)
Computation Energy Cost	
Integer Add (32b)	0.1
Integer Multiply (32b)	3.1
Floating Point Add (32b)	0.9
Floating Point Multiply (32b)	3.7
Memory Access Energy Cost	
8KB SRAM (64b)	10
1MB SRAM (64b)	100
DRAM	2000

interconnects can enable custom CIM designs by offering the densest embedded bitcell, low pJ/bit access energy, high-endurance, high-performance, and high-bandwidth [16], [17]; all desired attributes for ML accelerators. Therefore, in this article, an eDRAM array using $1T1C$ (T = Transistor, C = Capacitor) bitcells is presented, *for the first time*, as an efficient custom CIM engine by repurposing $1T1C$ eDRAM columns as charge domain circuits to perform DNN computations and support large-scale ML accelerators. The key contributions of this work are as follows.

1) The eDRAM-CIM architecture fully reconfigures existing eDRAM columns as charge domain circuits for required computations (data conversion, DNN computations, and weight storage), which greatly minimizes peripheral circuit area and power overhead.
2) The majority of DNN computations, such as data conversion, dot-product, pooling, rectified linear unit (ReLU) activation and multiply-accumulation-averaging (MAV) operations between 8-bit inputs and 8-bit signed/unsigned weights are supported directly inside the eDRAM array.
3) Weight stationary methodology is proposed to avoid extra weight duplication, reduce memory access cost, extra control logic, and requirement for a preinitialized array. In addition, $1T1C$ eDRAM bitcells are reconfigured for the MAV computation along with additional multiplexers (MUXs) to minimize the area overhead due to additional circuits.
4) An in-eDRAM adaptive dynamic-range successive-approximation (SAR) analog-to-digital converter (ADC) is proposed using a narrow range of dot-product distribution to minimize the ADC latency/energy by reconfiguring the $1T1C$ eDRAM bitcells and sense amplifiers (SAs) (components that are inherently presented in eDRAM arrays). The ADC dynamic range is configured to clip the quantization of extremely small or large analog input voltages, which appear infrequently, for better power efficiency and throughput with marginal DNN accuracy change.
5) The eDRAM-CIM engine is demonstrated on a 65-nm technology node with the CIFAR-10 dataset on the ResNet-20 and achieves 4.71 GOPS throughput, 8.26-GOPS/mm^2 and 4.76-TOPS/W average energy efficiency.
6) A scalability analysis is performed showing that the presented eDRAM-CIM engine when adopted to an advanced eDRAM technology (22-nm tri-gate CMOS process and 38.4-MB eDRAM array size) is able to achieve 11.12 TOPS/W as average energy efficiency and 22.2 TOPS as throughput when doing 60% clipping ADC.

The rest of this article is organized as follows. Section II introduces the background of this work and the motivation behind bitcell selection, compute-in-eDRAM design and adaptive dynamic range design; Section III presents the concept of the eDRAM-CIM design; Section IV describes the circuit implementation details of the proposed eDRAM-CIM CNN engine. Section V presents the 65 nm CMOS test-chip measurement results. Section VI concludes the article.

II. BACKGROUND

A. Convolution Neural Network

A CNN is a feed-forward trainable algorithm that is usually used in visual image analysis [18]. The network consists of an input layer, convolution (Conv) layers, pooling layers, activation layers, fully connected layers, and classification outputs. In Conv layer (Fig. 1), every input image is represented in the form of H(height) $\times$ W(width) $\times$ C(channel) arrays of pixel values, X_{IN}. These values are fed through multiple trainable $F \times F \times C$ kernels, which are used to extract features, such as edge, gradients, and colors. Note that the depth of the input array matches the depth of the kernels for the convolution computation. The filters are shifted across the input array to perform element-wise multiplications between each filter value and the corresponding input value, along with summation operations among all $F \times F \times C$ input-weight multiplied values. These operations are known as MAV computation [19], which dominate the computational cost in a CNN. Next, an activation function, such as ReLU, tanh, and sigmoid, will be applied to the output (Y_{OUT}), introducing nonlinearity into the neural network. After that, the pooling layer is applied to downsample the matrix and reduce the size of the output Y_{OUT}, thus reducing the number of parameters in the computation process. There are two common pooling layers: max pooling and average pooling. Max pooling generates the maximum number among a 2-by-2 matrix, and average pooling takes the mean value of a 2-by-2 matrix as shown in Fig. 1. In the last step, fully connected layer is connected to compile the previous layer data into a classification result.

B. Motivation for Compute-In-eDRAM Approach

Several compute-in/near-memory approaches are reported using static random access memory (SRAM) [19], [20], [21], [22] and DRAM [23], [24], [25]. For example, in the compute-in-SRAM approach, the multiply-and-accumulate (MAC) computations are enabled by activating multiple word lines connected to a common 6T SRAM column [20]. This approach is susceptible to bitcell variations causing inaccurate bitline current summation and potential bit-flips which requires the use of larger 8T/10T bitcells with a decoupled read port [19], [21], [22]. However, such a requirement limits its applicability for large CIM designs because of the increased area overhead. On the other hand, in commodity DRAM, charge

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

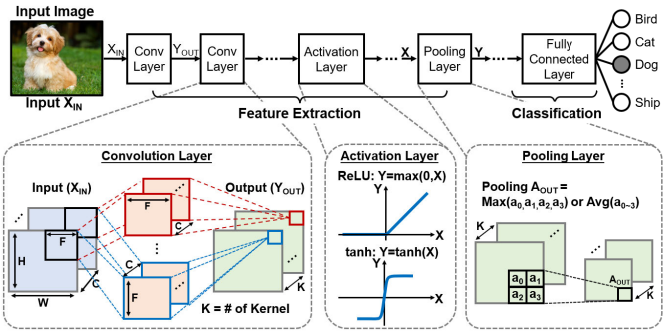


Fig. 1. CNN structure with convolution layer, pooling layer, and activation layer details.

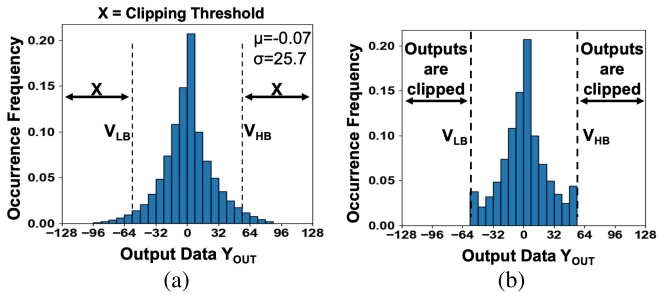


Fig. 2. (a) Convolution layer output distribution before output is clipped. (b) Convolution layer output after clipping is applied.

share methodology is used to perform 1-bit AND/OR operations among multiple bitcell storage capacitors [23], [24], [25]. However, since commodity DRAM is optimized for high bit density and low cost, adding extra CIM circuit into the existing DRAM technology is prohibitively expensive. eDRAM, on the other side, is logic-compatible and can support CIM circuits by utilizing the available high-performance transistors. The inherent capacitor inside the $1T1C$ eDRAM bitcell can be repurposed for charge share operations and can be used to effectively perform analog CIM computations. The existing $1T1C$ eDRAM columns in the normal mode of operation can be reconfigured as charge domain circuits in the CIM mode, which greatly minimizes peripheral circuit area and power overheads. In addition, configuring a part of eDRAM as a CIM engine (for ADC, DAC, weights) and the remaining part as regular memory (for inputs, gradients) can help to meet variable storage needs during a DNN inference/training step.

C. Motivation for Adaptive Dynamic Range In-eDRAM ADC

Despite the promising approach in various computation approaches inside memory, data conversion area/power overhead and conversion latency still remain elusive. In a conventional CIM approach, [19] presented an integrating ADC using a charge-sharing-based integrator, StrongARM latch-type SA, and logic block. This design replicates the 10T SRAM columns for charge-share and integration purposes, mitigating the process and temperature variation effect on the results. Sinangil et al. [26] developed a 4-bit flash ADC in the SRAM-CIM design, where 15 reference voltage levels and 15 comparators are used. Si et al. [27] presented a weight bitwise low-mac aware readout scheme to improve energy efficiency. In this design, the multibit MAC is read out through the switch-capacitor voltage amplifier. Moreover, the discussion on convolution data distribution strategies was also presented in [28] and [29], where an 11-level flash ADC was employed to constrain the quantization range, covering a majority of the data (−60, 60). Compared with these prior works, this work acknowledges the significance of incorporating convolution output distribution and the ability to support 127 uniform quantization levels in the SAR ADC design. Additionally, this work emphasizes minimizing the ADC design area overhead by leveraging the reuse of the existing $1T1C$ eDRAM bitcells. Therefore, the boundary level in the in-eDRAM ADC design can be easily adjusted by writing different bitcell values into the in-eDRAM ADC array. The distribution of outputs from convolution layers is depicted in Fig. 2(b), showing limited distribution at the extreme values, with most data concentrated around zero. A $1T1C$ in-eDRAM adaptive dynamic range ADC is proposed to address this, which utilizes the narrow distribution of convolution layer outputs. This approach involves trimming the extreme values and mapping them to the respective ends, mitigating the energy and area overhead associated with the ADC while improving computation latency.

III. eDRAM-CIM Architecture Overview

The concept of eDRAM-CIM computation is illustrated in Fig. 3(a). In eDRAM subarray, part of the array is used as input image storage and kernel weight storage, while the other part of the array can be configured as a CIM engine for data conversion and MAVs computations. All of these operations and functionalities are realized with the existing $1T1C$ eDRAM bitcells in order to minimize the extra dedicated circuit. The block diagram in Fig. 3(b) shows the details of the overall eDRAM-CIM design architecture. An 8-bit digital input (X_{IN}) and its 1's complement data (X_{IN}') are loaded to the eDRAM-based DAC. The digital-to-analog conversion is performed in two iterations and each iteration converts 4-bit digital data. In the first iteration, the converted analog voltage on the bitline is multiplied by a factor of 1 and in the second iteration, the next converted voltage is multiplied by a factor of 16. These two voltages are combined to generate the final differential analog voltages (V_a and V_{a_bar}) representing the 8-bit digital inputs. V_a and V_{a_bar} are differential analog voltages, which are centered around $1/2V_{DD}$, representing positive (above $1/2V_{DD}$) value and negative (below $1/2V_{DD}$) value. Multiple $1T1C$ eDRAM columns are reconfigured to realize multiple DAC pairs for different inputs (X_{INs}). The converted analog voltage from the DAC ($V_{DAC,i}$) can be represented in (1), where i represents the ith input in each operation and $W_{i,b7}$ represents the most significant bit (sign bit) of each weight (0: positive, 1: negative)

$$V_{DAC,i} = \begin{cases} V_{a,i}, & \text{if } W_{i,b7} = 0 \\ V_{a_bar,i}, & \text{otherwise.} \end{cases} \quad (1)$$

Next, an MAV operation is performed by first reading all 8 b weights in a row and sampling the weights onto the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

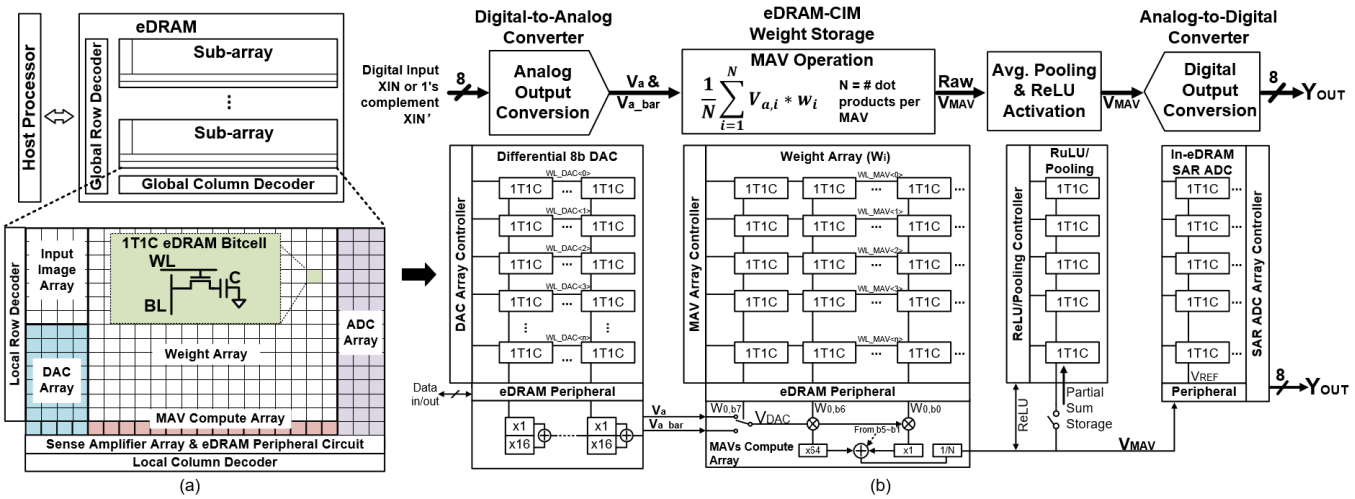


Fig. 3. Overall block diagram of eDRAM-CIM architecture (a) with details operations and (b) performing majority of DNN computations: digital-to-analog conversion (DAC), MAV, average pooling, ReLU, and ADC.

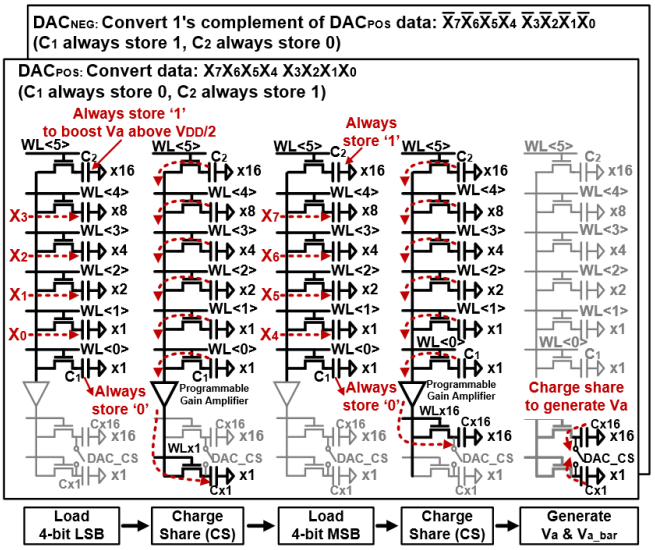


Fig. 4. In-eDRAM digital-to-analog converter (DAC) circuit diagram and data flow.

MAVs compute array as shown in Fig. 3(b). While the weights are read from the weight array, the bitcells are refreshed automatically with the intrinsic write-after-read process in the eDRAM read operation because after the sensing in the read operation is completed the data is restored to the bitcell through the sensing latch circuit [30]. The multiplications between V_a/V_{a_bar} and the weights are realized using the MUX. The MAV computation is shown in the following equation:

$$V_{\mathrm{MAV}} = \frac{1}{N} \sum_{i=0}^{F\times F\times C-1} V_{\mathrm{DAC,i}} \times W_{\mathrm{i}}. \tag{2}$$

Equation (2), V_{MAV} is the MAV computation result in the analog domain, N is the number of dot products in each MAV operation, F is the width and height of a filter, C is the number of channels, $F \times F \times C$ is the filter size and the total number of dot products in each convolution computation, $V_{\mathrm{DAC,i}}$ is one of the analog voltages (V_a or V_{a_bar}) from the differential DAC, and W_{i} is the corresponding kernel weight in analog voltage. If the filter size is larger than N, the partial dot product computation results can be stored in the partial sum storage eDRAM array, as shown in Fig. 3. In addition, the eDRAM-CIM also supports in-eDRAM ReLU activation and max/average pooling ON-chip to further reduce the intra-memory and OFF-chip data movement. In the last step, the MAV computation results (V_{MAV}s) are converted back to 8-bit digital output Y_{OUT} through the in-eDRAM SAR ADC and the reference is generated through $1T1C$ eDRAM columns.

IV. Circuits for eDRAM-CIM Design

A. In-eDRAM DAC

Fig. 4 gives an example of the in-eDRAM DAC operation assuming 8-bit digital input data is X_7–X_0. First, the lower 4 bits (X_3–X_0) of the 8-bit digital input, which is stored in different portions of the array, are first loaded into a $1T1C$ DAC column. The number of bitcells for each bit position is binary scaled. For example, LSB, X_0, is stored on 1 bitcell, while the MSB, X_3, is loaded onto 8 bitcells. In this case, each bitcell and bitcell capacitance are identical to improve the reconfigurability and scalability of the eDRAM-CIM design. Each DAC operation consists of two DAC columns ($\mathrm{DAC_{POS}}$ and $\mathrm{DAC_{NEG}}$). The top 16 eDRAM bitcells in $\mathrm{DAC_{POS}}$ always store data "1" and those in $\mathrm{DAC_{NEG}}$ always store data "0" to generate the differential voltage. After data is loaded to the column, the charge share operation is performed by activating all WLs (WL⟨0⟩–WL⟨5⟩) together, and thus the analog voltage ($V_{\mathrm{a_iter1}}$) for the first 4 bits of input data is generated. Equation (3) shows the analog voltage conversion in the $\mathrm{DAC_{POS}}$ for the first iteration (iter1)

$$V_{\mathrm{a_iter1}} = V_{\mathrm{DD}} \times \frac{16\times 1 + 8\times X_3 + 4\times X_2 + 2\times X_1 + 1\times X_0}{32}. \tag{3}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

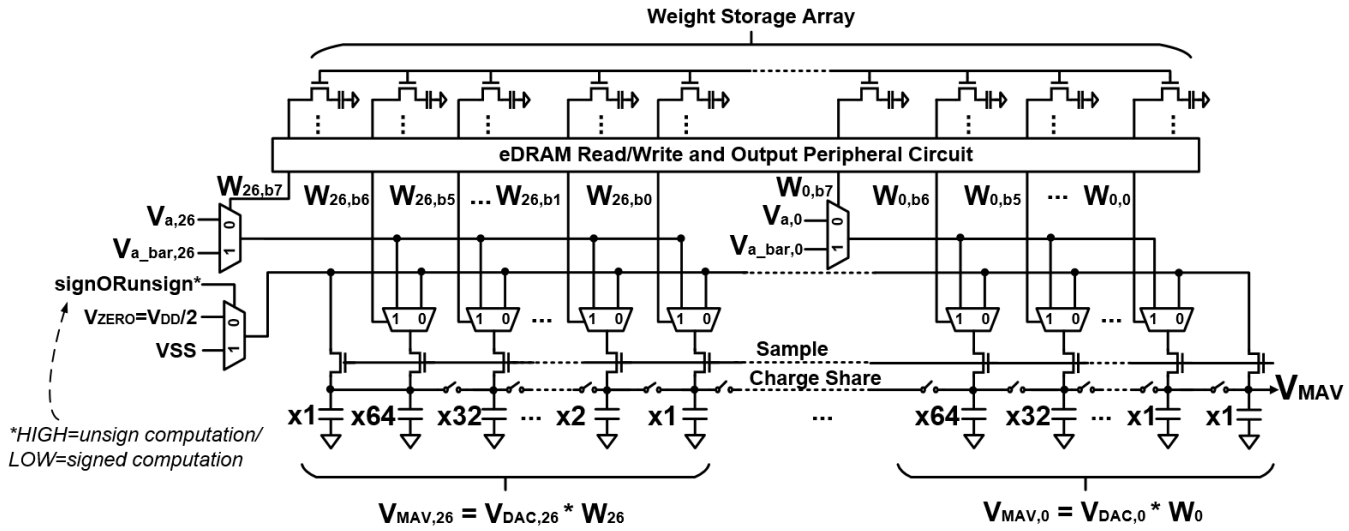


Fig. 5. In-eDRAM MAV computation detailed circuit diagram, where signORunsign defines signed or unsigned CIM computations, V_{ZERO} defines the zero reference voltage $V_{DD}/2$, V_a and V_{a_bar} are the analog voltage from the in-eDRAM DAC, and the weights are stored in the main eDRAM array.

Similar steps are performed for higher 4 bits, but the charge shared value (V_{a_iter2}) is sampled on a $16C$ capacitor (C_{X16}) to reflect its scaling factor in the second iteration (iter2). The analog voltage conversion for the second iteration is shown in the following equation:

$$V_{a_iter2} = V_{DD} \times \frac{16 \times 1 + 8 \times X_7 + 4 \times X_6 + 2 \times X_5 + 1 \times X_4}{32}. \tag{4}$$

After V_{a_iter1} and V_{a_iter2} are generated, final charge share across C_{X16} and C_{X1} is performed and the converted analog voltage is generated. Equation (5) shows the math behind the DAC_{POS} operation. Note that the "17" in the denominator could introduce a constant gain error in the conversion. However, such a system error can be calibrated and has a minimum effect on the test accuracy

$$V_a = \frac{V_{a_iter1} + 16 \times V_{a_iter2}}{17}. \tag{5}$$

For negative DAC columns, data conversion operations are similar to the data loading and charge share concepts. However, the data that are loaded into the bitcells are different. Instead of loading the input data X_7–X_0, the complementary data $\overline{X_7}$–$\overline{X_0}$ is written into the $1T1C$ eDRAM DAC_{NEG} column. Equation (6) shows the charge share operation in DAC_{NEG} for the first iteration where the top 16 eDRAM bitcells are always storing "0" and the bottom single bitcell is storing a constant "1"

$$V_{a_bar_iter1} = V_{DD} \times \frac{8 \times \overline{X_3} + 4 \times \overline{X_2} + 2 \times \overline{X_1} + 1 \times \overline{X_0} + 1}{32}. \tag{6}$$

Because of the complementary data that are stored in the DAC pair (DAC_{POS} and DAC_{NEG}), the generated analog voltages V_a and V_{a_bar} are centered around $V_{DD}/2$.

B. In-eDRAM Multiply-Accumulate-Average Operations

MAV operation details are shown in Fig. 5. The operation is performed by first reading all 8-bit weights in the entire row from the eDRAM array. The initial MUX takes in the two differential analog voltages (V_a or V_{a_bar}) as input from the DAC. The most significant bit, W_7, representing the sign bit, is used to select between V_a or V_{a_bar}. At the same time, the magnitude bit W_6–W_0 is used to select between analog voltage from the DAC and zero reference voltage $V_{DD}/2$ to perform dot product computations. Once the selection is made, the corresponding voltages after the MUX are sampled on the binary scaled eDRAM capacitors depending on the weight bit position. The analog voltage of each weight is represented in (7). Note that the extra capacitor, which is always sampling the zero reference voltage, is to make the denominator 128 instead of 127

$$W_i = \frac{64 \times W_{i,b6} + 32 \times W_{i,b5} + \ldots + 1 \times W_{i,b0}}{128}. \tag{7}$$

In this design, $1T1C$ eDRAM bitcells are reconfigured as the sampled capacitor for the dot product calculation, thus significantly reducing the area overhead in the design because the bitcells can be reused as other non-CIM applications as well. After the voltage is sampled on the $1T1C$ eDRAM capacitors, WLs are activated to trigger the charge share between all the bitcells and the final MAV computation result V_{MAV} is generated as shown in the following equation:

$$V_{MAV} = \frac{V_{MAV,0} + V_{MAV,1} + \ldots + V_{MAV,26}}{27}. \tag{8}$$

Before passing the final analog MAV results to the ADC, ReLU can be triggered by enabling the comparator shown in Fig. 6. The SA compares the V_{MAV} with zero reference voltage $V_{DD}/2$ and controls the analog MUXs, S_1 and S_2, to pass either the zero voltage or the MAV voltage to the ADC.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

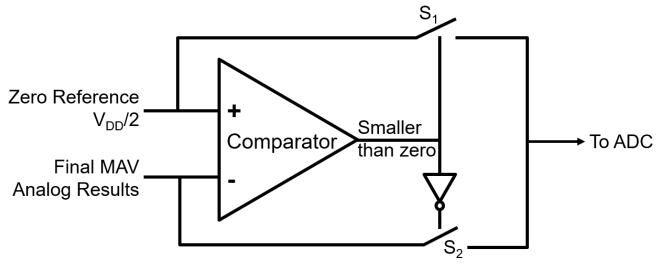


Fig. 6. ReLU computation circuit using a comparator and analog MUX.

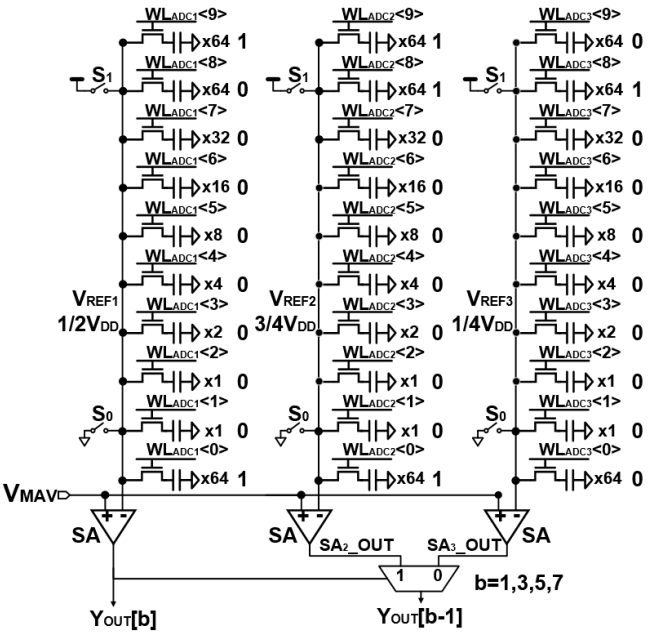


Fig. 7. In-eDRAM ADC array circuit with V_{REF} generation circuit and SA comparison between V_{REF} and V_{MAV}.

C. Adaptive Dynamic Range In-eDRAM ADC

In the last step, an SAR ADC is realized by comparing V_{MAV} with a reference voltage (V_{REF}), as shown in Fig. 7. To perform the data conversion, first, each $1T1C$ eDRAM bitcells will be written with corresponding data. For example, in Fig. 7, to generate reference 1 (V_{REF1}, $1/2V_{\mathrm{DD}}$), the top and bottom 64 $1T1C$ bitcells are loaded with data "1" and the rest of bitcells are loaded with data "0" shown in the first column. Next, all the WLs are turned on to perform charge share operation and generate the reference voltage, as represented in the following equation, where C is the eDRAM bitcell capacitor:

$$V_{\mathrm{REF1}} = V_{\mathrm{DD}} \times \frac{64C \times 1 + 64C \times 1}{256}. \tag{9}$$

The same concept is applied to V_{REF2}–V_{REF3}. After all the bitcell capacitors are charge shared, the first column generates $1/2V_{\mathrm{DD}}$ as V_{REF1}, the second column generates $3/4V_{\mathrm{DD}}$ as V_{REF2} and the third column generates $1/4V_{\mathrm{DD}}$ as V_{REF3}. At the same time, SAs are triggered to compare the MAV computation result (V_{MAV}) and the three different reference voltages. By generating three reference voltages simultaneously, 8-bit SAR ADC only takes four cycles to complete the data conversion. In contrast, an 8-bit SAR ADC traditionally needs eight cycles to complete the conversion. Such a design is critical because the capacitor leakage will affect the conversion results if the conversion time is extensive.

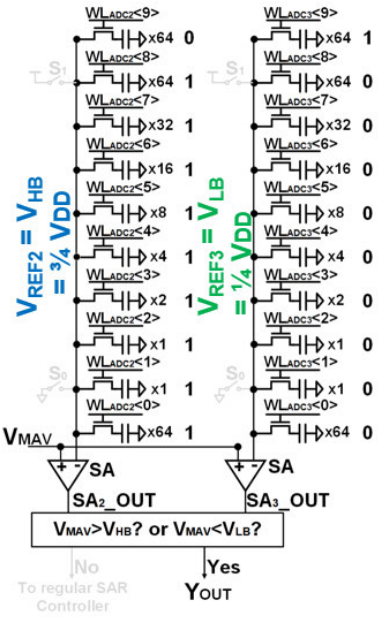


Fig. 8. In-eDRAM ADC column configurations for generating upper boundary, V_{HB} and lower boundary V_{LB}.

Furthermore, to mitigate the ADC energy and area overheads, a $1T1C$ eDRAM-based adaptive dynamic-range ADC is devised by leveraging the narrow V_{MAV} distribution, as discussed in Fig. 2(b). In order to utilize the narrow data distribution, the in-eDRAM ADC columns are configured to generate two boundaries, V_{HB} and V_{LB}, and the column configurations are shown in Fig. 8. These two columns' configurations are subject to change depending on the desired boundary reference voltage. The controller first compares the V_{MAV} with the upper and lower boundaries and if the MAV results are out of the boundary, the controller will skip the ADC operation immediately to avoid extra energy and computation latency. If the MAV results are inside the boundary, the controller will continue the normal SAR ADC operation.

V. Measured Results

A. Measurement Setup and Test-Chip Summary

Fig. 9 shows the overall measurement setup of the test-chip implemented in 65-nm CMOS process technology. The test-chip supports 8-bit input, signed/unsigned 8-bit weight, and 8-bit output in CNN architecture and occupies a total area of 0.57 mm^2, shown in Fig. 10. In the total macro area, the eDRAM weight array takes 0.22 mm^2, the MAV computation, which is responsible for averaging, takes 0.08 mm^2, DAC array takes 0.12 mm^2 and the SAR ADC array takes 0.09 mm^2. The bitcell capacitor used in the $1T1C$ eDRAM array design is selected to be 13 fF to ensure >2 ms retention time in hot temperature (80 °C). In addition, due to the limitation of the technology, metal-oxide-meal (MOM) capacitor is used from M3 to M7 layer to design the bitcells.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

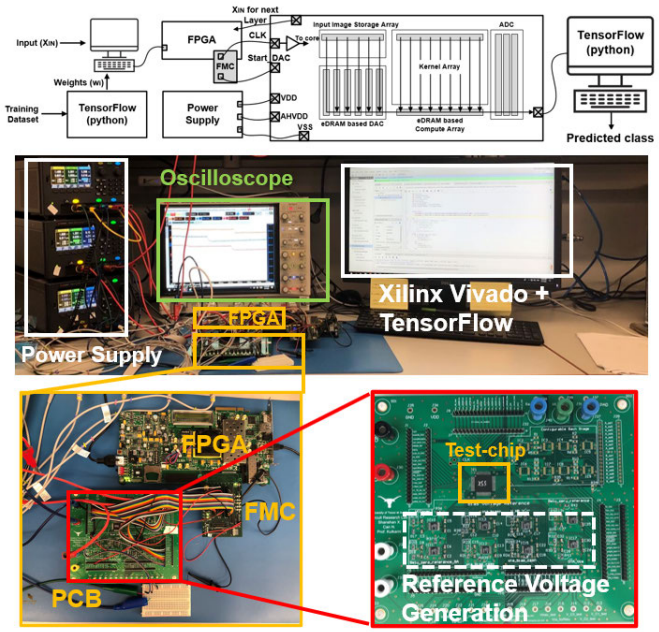


Fig. 9. eDRAM-CIM experiment setup with power supply, FPGA, oscilloscope, test board PCB, and Xilinx Vivado.

The overall CNN inference and training methodology consists of the following steps.

1) Input training using Keras/TensorFlow [31] application programming interface (API) in Python.
2) Floating point trained weight is quantized into signed 8-bit data.
3) Hardware communication interface between the test-chip on the printed circuit board (PCB) and the Virtex-7 FPGA VC707 board is developed using FMC XM105 debug card. The connection is made using jumper wires.
4) CIFAR10 validation dataset image pixel inputs and trained 8-bit kernel weights are fed to the test-chip through the FPGA board.
5) ON-chip data conversion, MAVs, ReLU, and pooling operations are performed for convolution layers.
6) Fully connected layers are performed OFF-chip using Keras/TensorFlow in Python.

Due to the absence of ON-chip global memory, the input activation and weight values are loaded onto the eDRAM array through a scan chain interface. The 8-bit SAR ADC outputs are directly transferred out to the chip using eight digital I/O pads.

B. Test-Chip Characterization

1) In-eDRAM DAC: The experimental demonstration of the in-eDRAM DAC operation is shown in Fig. 11. In the first step, lower 4-bit LSBs are loaded to the corresponding bitcell through BL by turning on WL⟨0⟩–WL⟨5⟩ sequentially. WL⟨5⟩ turns on twice, shown as point A in Fig. 11, to make sure that 16 capacitors are fully charged to ensure correct charge share voltage. After loading the data, all the bitcells in the column are charge shared by turning on all the WLs. At the same time, the WL×1 turns on to sample the charge-shared voltage to C_{X1} capacitor. In the second round, the charge-shared voltage is sampled on the C_{X16} capacitor. After C_{X1} and C_{X16} are charge shared, V_{a} and $V_{\text{a_bar}}$ are generated, which are centered around $V_{\text{DD}}/2$. Fig. 12 shows the differential nonlinearity (DNL) of in-eDRAM DAC is within 1 LSB at lower input values. The analog voltage of positive DAC is saturated at around 1 V because of the nonrail-to-rail operational amplifier design.

2) In-eDRAM ADC: The in-eDRAM ADC transfer curves in both nominal (nonclipping) and clipping configurations are shown in Fig. 13(a) and (b), respectively. In the clipping configuration, the ADC saturates the quantization of low and high V_{MAV} at 64 and 191 without undergoing subsequent successive approximation cycles, although it incurs 1 or 2 extra cycles to compare V_{MAV} with V_{LB} and V_{HB}. The V_{MAV} values within the bounded range are quantized using the remaining nonclipping 128 levels. This approach shortens the conversion latency and saves power by reducing the active charge redistribution in the capacitive DAC, contributing to ADC power consumption. The DNL results are shown at the bottom of each transfer curve.

C. Energy and Accuracy Trade-Off

Data clipping in in-eDRAM SAR ADC design helps to skip the subsequent computation and comparison and thus decreases the computation energy. However, such a technique will also incur accuracy loss in the classification process because values at the two ends are saturated to the boundaries. Fig. 14 evaluates the normalized ADC energy and accuracy trade-off. As the clipping threshold increases, accuracy drops down. The normalized ADC energy is the ratio between ADC energy with clipping and ADC energy without clipping configuration. To optimize the design and not sacrifice too much accuracy drop, a 60% optimum clipping threshold is chosen in the measurement. Fig. 15 shows that, at 60% optimum clipping threshold, accuracy drops by 1.94% using CIFAR-10 dataset with ResNet-20. At the same time, throughput is increased by 1.14×, and ADC energy is reduced by 1.14×.

D. Classification Accuracy

To understand the classification accuracy, software accuracy, measurement inference accuracy, and simulation accuracy with different variations are shown in Fig. 16. Software 8-bit INT inference accuracy is obtained by quantizing the floating point weights to provide baseline accuracy data for the test-chip using ResNet-20. In addition, variation analysis is performed to capture the accuracy degradation due to the transistor/capacitor variations and retention time leakage on the $1T1C$ eDRAM bitcells. The transistor and capacitor variation is captured by Monte Carlo simulation over 10 K samples. To address the retention time leakage issue on the bitcell capacitor, the entire computation is purposely slowed down to sample and charge share at the end of the retention time without frequent refresh. As shown in Fig. 16, 92.29% accuracy is obtained as a baseline. In the measurement, the test chip achieves 90.77% accuracy in nominal CIM mode and 89.86% in clipping CIM mode. Accuracy drops down concerning nonclipping mode because the two extreme values are trimmed and saturated at the upper and lower bounds,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Technology	65nm CMOS
Supply Voltage	1~1.2V
Input Clock Freq.	50~200MHz
Bitcell Size	22.08μm²
Macro Area	0.57mm²
eDRAM Capacitor Type	Metal-Oxide-Metal (MOM)
Unit eDRAM Capacitor Size	13fF
Input/Weight/Output Precision	8b/8b/8b

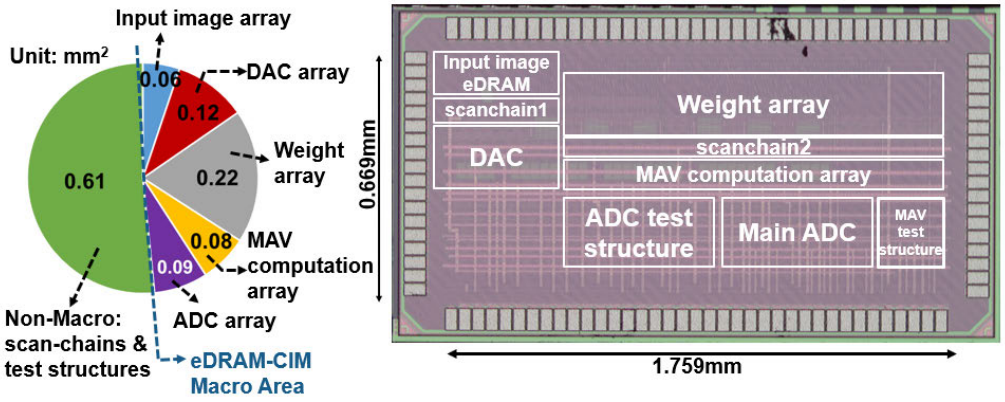


Fig. 10. Test-chip measurements summary, area breakdown, and die micrograph with functional blocks highlighted.

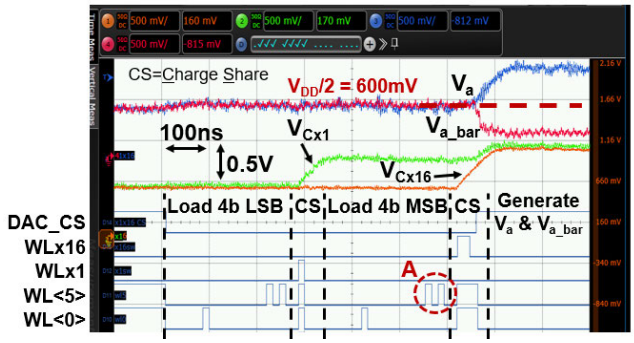


Fig. 11. eDRAM DAC demonstration with oscilloscope waveforms.

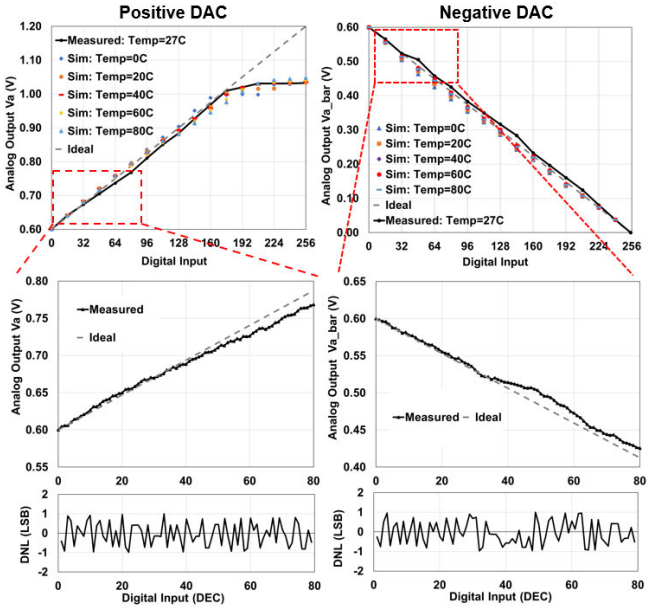


Fig. 12. In-eDRAM DAC input to output characteristic and DNL.

thus creating a small amount of data deviation. In the variation simulation, the first ideal dataset represents the accuracy of no transistor and capacitor variations. When transistor and capacitor variations are involved, accuracy is dropped by 1.61%. The next case takes the retention time into consideration. When the computation is performed at the end of the retention time, the inference accuracy is dropped by 5.93% compared to the ideal case. However, in the realistic eDRAM-CIM

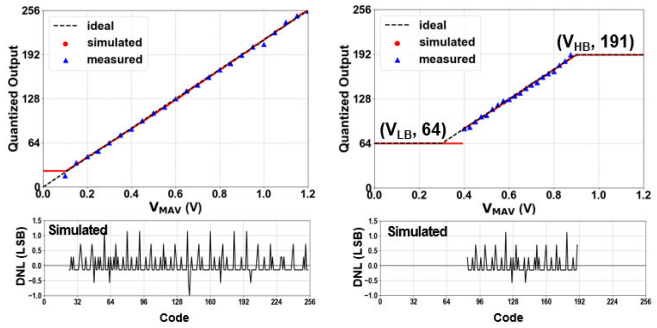


Fig. 13. (a) In-eDRAM SAR ADC DNL without clipping feature and (b) in-eDRAM SAR ADC DNL with clipping.

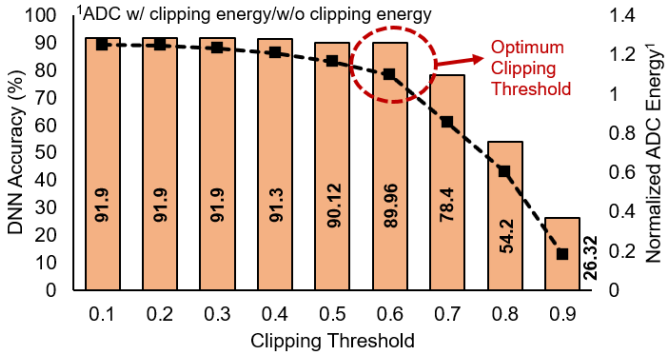


Fig. 14. DNN classification accuracy and normalized energy variation with the in-eDRAM ADC clipping threshold.

computation, the charge share is performed right after the data are written into the memory array. Moreover, because of the write-back mechanism inherent in the $1T1C$ eDRAM array, the charges of the bitcells are refreshed after the bitcell read operation. To illustrate, weights are read from the weight storage array during the MAV computation and transferred to the computation array. Consequently, the refresh operation is executed whenever the MAV computation takes place.

E. Test-Chip Summary

The eDRAM-CIM design achieves an average throughput of 4.71 (4.037) GOPS with (without) clipping mode on the CIFAR10 dataset while operating at 1–1.2 V and 50 MHz experimental frequency and attains 4.76 TOPS/W average

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

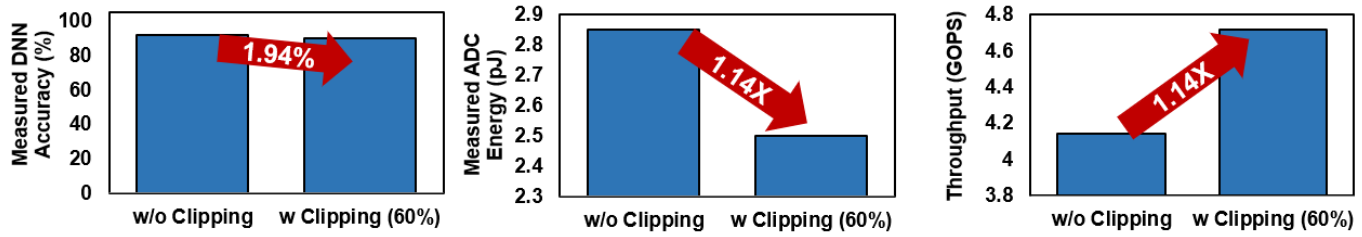


Fig. 15. ADC accuracy, throughput, and energy trade-off with respect to clipping operation.

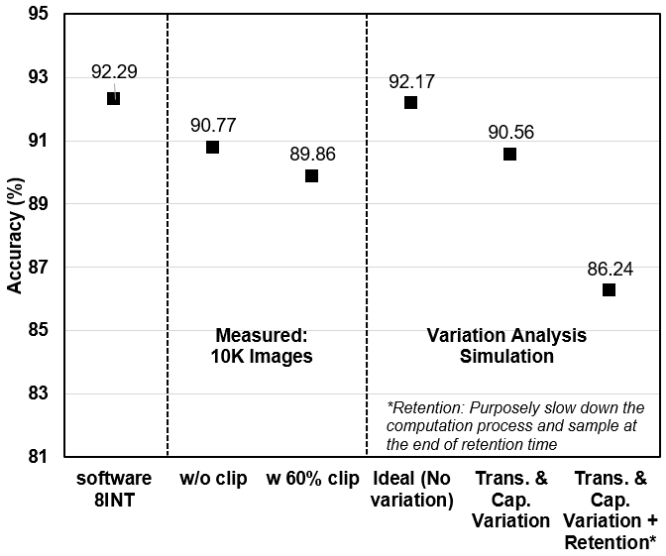


Fig. 16. DNN accuracy of eDRAM-CIM design in software, measurement, and variation analysis cases.

energy efficiency. The throughput is calculated as follows. There are 32 storage elements for partial sum storage in the $1T1C$ eDRAM array. The eDRAM array can perform 32 MAV operations at the same time. Thus, 32×32 operations are counted toward total operations. In addition, each MAV operation is considered two operations (average and accumulation). Concurrently, total operations equal $32 \times 32 \times 2$. Regarding latency, the design takes 87 (99) cycles to complete 32×32 MAV computations with (without) clipping feature. The total power consumption measured on the chip is 0.99 mW, which results in the energy efficiency number shown in Table II. As discussed in Section V-A, $1T1C$ eDRAM bitcells are limited by the 65 nm technology and the available MOM capacitor. Therefore, a scalability analysis is performed for a large capacity, 22-nm Tri-gate eDRAM array featuring a dense bitcell and supporting high operating frequency, as shown in the summary table, Table II. With the scalability analysis, the $1T1C$ eDRAM capacitor is designed with capacitor-over-bitline (COB) technology [32] and has the same 13 fF capacitance. By assuming 30% utilization of a 128 MB (77 mm^2) [33] eDRAM memory for CIM usage, the GOPS/mm^2 number is 1305. With a larger array size and advanced technology, the throughput increases to 22.2 TOPS since more computation can be performed concurrently. Energy efficiency increases to 37.29 TOPS/W with a higher clock frequency design and increased parallelism in large-capacity $1T1C$ eDRAM array.

TABLE II
TEST-CHIP MEASUREMENTS SUMMARY

		Test-chip Summary	Estimated
Technology		65nm CMOS	22nm tri-gate CMOS
Supply Voltage (V)		1~1.2	1.05
Total eDRAM Size		16Kb	[2]38.4MB
Bit-cell Size (um^2)		22.08	0.029
Unit eDRAM Capacitor Size (fF)		13	13
eDRAM Capacitor Type		Metal-oxide-metal (MOM)	Capacitor-over-bitline (COB)
Input Clock Frequency		[4]50 MHz (measured) 200MHz (simulation)	1.6GHz
[3]eDRAM-CIM Macro Area (mm^2)		0.57	[2]17
GOPS/mm^2		8.26	1305
MAV Operation Latency (ns)		10 (2 cycles)	1.25 (2 Cycles)
Input, Weight, Output Precision		8b, 8b, 8b	8b, 8b, 8b
Dataset		CIFAR-10	CIFAR-10
Model		ResNet-20	ResNet-20
Software Classification Accuracy (INT8)		92.29%	N/A
Measured Classification Accuracy for 5000 Images	**w/o ADC Clipping**	90.77%	
	with 60% ADC Clipping	89.86%	
[1]Throughput @ 1.1V	**w/o ADC Clipping**	4.037 GOPS	19.5 TOPS
	with 60% ADC Clipping	4.71 GOPS	22.2 TOPS
[5]Energy Efficiency (TOPS/W) @ 1.1V		4.76	37.29
Energy/MAV (fJ/MAV) with 60% ADC Clipping @ 1.1V		30.6	[6]0.21

[1]Assume 1MAV = 2 operations
[2]Assume 30% utilization of 128MB (77mm^2) [32] eDRAM memory for CIM usage
[3]Main design area (not including others structure, e.g., scanchain, test structure)
[4]Limited by jumper wire bandwidth
[5]Average Energy Efficiency and excluded scan-chains power
[6]Increased parallelism in large-capacity eDRAM array

F. Comparison With Prior Energy-Efficient CNN Accelerators

Table III presents a comparison with prior CIM works for different types of memories [19], [27], [34], [35], [36], [37], [38], [39] which support multibit input and weight integer operands using CIFAR-10 dataset. The eDRAM-CIM design performs the entire MAV computation within the 16 Kb $1T1C$ eDRAM array, which can support 8-bit input and signed 8-bit weight. Compared with prior methods, the eDRAM-CIM approach repurposes the existing $1T1C$ eDRAM bitcells to achieve the computations, significantly minimizing the area overhead due to the dedicated accelerator arithmetic circuits. Additionally, the figure of merit is computed based on input

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
COMPARISON WITH PRIOR CIM WORKS SUPPORTING MULTIBIT INPUT AND WEIGHT INTEGER OPERANDS AND USING CIFAR-10 DATASET

	This work	VLSI'22 [34]	TCAS'21 [35]	ISSCC'21 [36]	ISSCC'23 [37]	ISSCC'20 [38]	ISSCC'20 [27]	ISSCC'20 [39]	ISSCC'18 [19]
Technology	**65nm**	65nm	65nm	65nm	28nm	28nm	28nm	22nm	65nm
Memory Cell Structure	**1T1C**	2T1C eDRAM	4T2C eDRAM	3T1C eDRAM	3T2C eDRAM	6T SRAM	6T + Local Computing SRAM	1T1R SLC ReRAM	6T SRAM
Array Size	**16Kb**	32Kb	16Kb	8Kb	9600Kb	64Kb	64Kb	2Mb	128Kb
Input Precision (bit)	**8**	8	5	4	4	8	8	4	8
Weight Precision (bit)	**8**	8	3	4	5	8	8	4	8
Supply Voltage (V)	**1~1.2**	1	0.5	1.1	1	0.85~1.0	0.7~0.9	0.8	1
Dataset	**CIFAR-10**	CIFAR-10							
Model	**ResNet-20**	ResNet-20	VGG-like CNN	ResNet-18	ResNet-20	ResNet-20	ResNet-20	-	SVM
Measured Accuracy	**90.77%**	92.02%	82.8% (Simulated)	91.2%	89.5%	91.91%	92.02%	-	83.27%
Throughput (GOPS)	**[1]4.71 GOPS**	22 GOPS	-	-	19.5 TOPS	-	-	-	4
Energy Efficiency (TOPS/W)	**[1]4.76**	7.39	17.8	32.5	13.2	7.3 [2](1.35)	14.08 [2](2.61)	28.93 [2](3.31)	3.125
[3]FoM	**304.6**	472	267	520	264	86.4	167	53	201.6

[1]measured at 1.1V
[2]Scaled to 65nm, assume energy $\propto$ (Tech.)^2 [19]
[3]FoM = input precision x weight precision x energy efficiency

precision, weight precision, and energy efficiency in 65 nm. Compared with prior approaches, the eDRAM-CIM design achieves 304 FoM in 65-nm CMOS process.

VI. CONCLUSION

This article presents a CIM design using eDRAM memory array by repurposing $1T1C$ eDRAM bitcell and its intrinsic bitcell capacitor to perform the charge share operation and realize the MAV computation. This design supports 8-bit input, signed/unsigned 8-bit weight, and 8-bit output for the operations. With the CIFAR10 dataset evaluation, the design achieves 90.77% accuracy, 8.26 GOPS/mm^2 and 4.76 TOPS/W average energy efficiency. Compared with prior CIM design, this design significantly reduces area overhead and mitigates extra circuitry in order to realize the computation operations, which is confirmed in GOPS/mm^2 result. The results indicate that the eDRAM-CIM approach, both in 65 nm technology and in an advanced eDRAM technology, shows promising energy efficiency and throughput metrics suggesting its potential for deploying in large-scale CIM designs to mitigate the memory bottleneck challenges.

ACKNOWLEDGMENT

The authors would like to thank Clifford Ong for the technical discussions and Intel for funding support.

REFERENCES

[1] 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.

[2] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1891–1898.

[3] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[4] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[5] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[6] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[7] L. Wei et al., "A 7 Mb STT-MRAM in 22FFL FinFET technology with 4 ns read sensing time at 0.9 V using write-verify-write scheme and offset-cancellation sensing technique," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 214–216.

[8] Y.-D. Chih et al., "A 22 nm 32 Mb embedded STT-MRAM with 10 ns read speed, 1M cycle write endurance, 10 years retention at 150 °C and high immunity to magnetic field interference," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 222–224.

[9] Q. Dong et al., "A 1-Mb 28-nm 1T1MTJ STT-MRAM with single-cap offset-cancelled sense amplifier and in situ self-write-termination," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 231–239, Jan. 2019.

[10] Y. Shih et al., "Logic process compatible 40-nm 16-Mb, embedded perpendicular-MRAM with hybrid-resistance reference, μA sensing resolution, and 17.5-ns read access time," *IEEE J. Solid-State Circuits*, vol. 54, no. 4, pp. 1029–1038, Apr. 2019.

[11] P. Jain et al., "A 3.6 Mb 10.1 Mb/mm^2 embedded non-volatile ReRAM macro in 22 nm FinFET technology with adaptive forming/set/reset schemes yielding down to 0.5 V with sensing time of 5 ns at 0.7 V," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 212–214.

[12] C.-C. Chou et al., "A 22 nm 96KX144 RRAM macro with a self-tracking reference and a low ripple charge pump to achieve a configurable read window and a wide operating voltage range," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2020, pp. 1–2.

[13] M.-F. Chang et al., "Embedded 1 Mb ReRAM in 28 nm CMOS with 0.27-to-1 V read using swing-sample-and-couple sense amplifier and self-boost-write-termination scheme," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 332–333.

[14] C.-F. Lee, H.-J. Lin, C.-W. Lien, Y.-D. Chih, and J. Chang, "A 1.4 Mb 40-nm embedded ReRAM macro with 0.07 μm^2 bit cell, 2.7 mA/100 MHz low-power read and hybrid write verify for high endurance application," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2017, pp. 9–12.

[15] S. Aga, N. Jayasena, and M. Ignatowski, "Co-ML: A case for collaborative ML acceleration using near-data processing," in *Proc. Int. Symp. Memory Syst.*, Sep. 2019, pp. 506–517.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[16] G. Fredeman et al., "A 14 nm 1.1 Mb embedded DRAM macro with 1 ns access," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 230–239, Jan. 2016.

[17] C. Berry et al., "IBM z15: A 12-core 5.2 GHz microprocessor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 54–56.

[18] B. A. El-Rahiem, M. A. O. Ahmed, O. Reyad, H. A. El-Rahaman, M. Amin, and F. A. El-Samie, "An efficient deep convolutional neural network for visual image classification," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl.* Cham, Switzerland: Springer, 2019, pp. 23–31.

[19] A. Biswas and A. P. Chandrakasan, "Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 488–490.

[20] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.

[21] X. Si et al., "A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 396–398.

[22] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A 64-tile 2.4-Mb in-memory-computing CNN accelerator employing charge-domain compute," *IEEE J. Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, Jun. 2019.

[23] V. Seshadri et al., "Ambit: In-memory accelerator for bulk bitwise operations using commodity DRAM technology," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2017, pp. 273–287.

[24] S. Li, D. Niu, K. T. Malladi, H. Zheng, B. Brennan, and Y. Xie, "DRISA: A DRAM-based reconfigurable in-situ accelerator," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2017, pp. 288–301.

[25] F. Gao, G. Tziantzioulis, and D. Wentzlaff, "ComputeDRAM: In-memory compute using off-the-shelf DRAMs," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 100–113.

[26] M. E. Sinangil et al., "A 7-nm compute-in-memory SRAM macro supporting multi-bit input, weight and output and achieving 351 TOPS/W and 372.4 GOPS," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 188–198, Jan. 2021.

[27] X. Si et al., "A 28 nm 64 Kb 6T SRAM computing-in-memory macro with 8b MAC operation for AI edge chips," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 246–248.

[28] Z. Jiang, S. Yin, J.-S. Seo, and M. Seok, "C3SRAM: An in-memory-computing SRAM macro based on robust capacitive coupling computing mechanism," *IEEE J. Solid-State Circuits*, vol. 55, no. 7, pp. 1888–1897, Jul. 2020.

[29] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks," *IEEE J. Solid-State Circuits*, vol. 55, no. 6, pp. 1733–1743, Jun. 2020.

[30] B. Jacob, D. Wang, and S. Ng, *Memory Systems: Cache, DRAM, Disk*. San Mateo, CA, USA: Morgan Kaufmann, 2010.

[31] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.

[32] F. Hamzaoglu et al., "A 1 Gb 2 GHz embedded DRAM in 22 nm tri-gate CMOS technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 230–231.

[33] N. Kurd et al., "Haswell: A family of IA 22 nm processors," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 49–58, Jan. 2015.

[34] S. Xie, C. Ni, P. Jain, F. Hamzaoglu, and J. P. Kulkarni, "Gain-cell CIM: Leakage and bitline swing aware 2T1C gain-cell eDRAM compute in memory design with bitline precharge DACs and compact Schmitt trigger ADCs," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 112–113.

[35] C. Yu, T. Yoo, H. Kim, T. T. Kim, K. C. T. Chuan, and B. Kim, "A logic-compatible eDRAM compute-in-memory with embedded ADCs for processing neural networks," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 2, pp. 667–679, Feb. 2021.

[36] Z. Chen, X. Chen, and J. Gu, "A 65 nm 3T dynamic analog RAM-based computing-in-memory macro and CNN accelerator with retention enhancement, adaptive analog sparsity and 44TOPS/W system energy efficiency," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 240–242.

[37] S. Kim et al., "DynaPlasia: An eDRAM in-memory-computing-based reconfigurable spatial accelerator with triple-mode cell for dynamic resource switching," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2023, pp. 256–258.

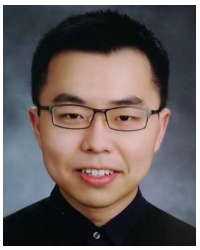
[38] J.-W. Su et al., "A 28 nm 64 Kb inference-training two-way transpose multibit 6T SRAM compute-in-memory macro for AI edge chips," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 240–242.

[39] C.-X. Xue et al., "A 22 nm 2 Mb ReRAM compute-in-memory macro with 121–28TOPS/W for multibit MAC computing for tiny AI edge devices," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 244–246.

Shanshan Xie (Graduate Student Member, IEEE) received the B.S. degree in electrical and computer engineering (ECE) from Worcester Polytechnic Institute (WPI), Worcester, MA, USA, in 2018. She is currently pursuing the M.S. and Ph.D. degrees in electrical and computer engineering from The University of Texas at Austin (UT Austin), Austin, TX, USA.

She was an Intern with Analog Devices, Wilmington, MA, USA, and Texas Instrument Corporation, Dallas, TX, USA, where she was involved in programmable gain instrumentation amplifier, ECG heart rate monitor, CAN isolation products. After graduation from UT Austin, she worked as a Circuit Research Scientist with Reliable and Resilient Circuits Team, Intel Laboratory, Hillsboro, OR, USA. Her research interests include mix-signal design for compute-in-memory techniques, machine learning accelerator, and annealing processor.

Dr. Xie is currently one of the TPC Members for the digital subcommittee of the IEEE CUSTOM INTEGRATED CIRCUITS CONFERENCE. She was a recipient of the 2020 Cadence Women in Technology Scholarship, from 2021 to 2022 SSCS Predoctoral Achievement Award and the 2023 ECE Graduate Achievement Award.



Can Ni received the bachelor's degree in engineering physics from the University of Alberta, Edmonton, AB, Canada, in 2019, and the master's degree in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2021.

He was an Intern with Qualcomm, San Diego, CA, USA, and currently with Apple Cellular Modem Group, San Diego. His research interests include compute-in-memory and neuromophic hardware.

Aseem Sayal received the bachelor's degree in electrical and electronics engineering from Delhi Technological University (formerly Delhi College of Engineering), Delhi, India, in 2013, and the M.S. and Ph.D. degrees in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2017 and 2020, respectively.

From 2013 to 2015, he worked as a Physical Design CAD Engineer at Qualcomm India Private Ltd., India. He was also an Intern at Apple Inc., Cupertino, CA, USA, and Google LLC, Sunnyvale, CA, USA, where he was involved in developing clock tree simulation methodology and average power estimation solutions for long-running arbitrary workloads, respectively. He is currently working as an ML Design Engineer with the Platforms Datacenter Chip Team, Google LLC. He has filed five patents and authored 15 articles in referred journals and conferences. His graduate research was focused on designing energy-efficient hardware accelerators for machine learning applications and developing EDA design methodologies for heterogeneously integrated secured ASICs.

Dr. Sayal was a recipient of the Chancellor Gold Medal and Meritorious Student Award from Delhi Technological University in 2013.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Pulkit Jain (Senior Member, IEEE) received the B.Tech. degree from IIT Kanpur, Kanpur, India, and the Ph.D. degree in EE from the University of Minnesota, Minneapolis, MN, USA.

He is a Memory Architect at AMD, Santa Clara, CA, USA, where he is driving the memory subsystem of adaptive SoCs for datacenter and automotive markets. Before this, he was at Intel Laboratory, Hillsboro, OR, USA, where he also drove design and new architectures centered around in-package memory and system caches.

Fatih Hamzaoglu (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2002.

After finishing the Ph.D., he joined Technology Development Group, Intel Corporation, Hillsboro, OR, USA, and since then, he has been working on memory technology developments such as SRAM, eDRAM, MRAM, and RRAM. Currently, he is an Intel Fellow and the Director of new in-package memory technologies, IP design and product integration. He has authored or coauthored more than 40 papers and inventor/co-inventor of more than 30 patents.

Dr. Hamzaoglu served for the VLSI Symposium Circuits Committee and the ISSCC Memory Subcommittee from 2013 to 2019.

Jaydeep P. Kulkarni (Senior Member, IEEE) received the B.E. degree from the University of Pune, Pune, India, in 2002, the M.Tech. degree from the Indian Institute of Science (IISc), Bengaluru, India, in 2004, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2009, all in electronics/electrical engineering.

From 2009 to 2017, he was with Intel Circuit Research Laboratory, Hillsboro, OR, USA, where he worked on energy-efficient integrated circuit technologies. He is currently an Assistant Professor in electrical and computer engineering at The University of Texas at Austin, Austin, TX, USA, and a fellow of the AMD Endowed Chair in Computer Engineering and the Silicon Labs Chair in Electrical Engineering. He has filed 35 patents and authored 85 articles in referred journals and conferences. His research is focused on machine learning hardware accelerators, in-memory computing, emerging nano-devices, hardware security, heterogeneous/3-D integration, and cryogenic computing.

Dr. Kulkarni is a member of the ACM. He received the Best M.Tech. Student Award from IISc, the Intel Foundation Ph.D. Fellowship Award, the SRC Best Paper and Inventor Recognition Awards, the Purdue Outstanding Doctoral Dissertation Award, seven Intel Divisional Recognition Awards, the 2015 IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS Best Paper Award, the SRC Outstanding Industrial Liaison Award, Micrometer Foundation Faculty Awards, and the Intel Rising Star Faculty Award. He has served as the Conference General Co-Chair for 2018 ISLPED, and is currently participating in the Technical Program Committees of Custom Integrated Circuits Conference (CICC), the International Conference on Computer-Aided Design (ICCAD), the Design Automation Conference (DAC), and the International Conference on Artificial Intelligence Circuits and Systems (AICAS) conferences. He is currently serving as the Chair for the IEEE Central Texas SSCS/CAS Joint Chapter. He is also serving as an Associate Editor for IEEE SOLID STATE CIRCUIT LETTERS and IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 06:25:28 UTC from IEEE Xplore. Restrictions apply.

Journal of the Brazilian Society of Mechanical Sciences and Engineering (2023) 45:625
https://doi.org/10.1007/s40430-023-04542-x

TECHNICAL PAPER

Effect of varying ambient conditions on the performance of air conditioner using evaporative cooler

S. K. Gupta[1] · B. B. Arora[1] · A. Arora[1]

Received: 3 October 2022 / Accepted: 12 October 2023
© The Author(s), under exclusive licence to The Brazilian Society of Mechanical Sciences and Engineering 2023

Abstract

The increasing usage of refrigeration and air conditioning worldwide is a serious concern as regard the greenhouse gas emissions which degrade the environment. Improving the energy efficiency could reduce this damage. The heat dissipation capacity of the condenser and performance of an air conditioner largely depend upon the temperature of the air entering the condenser. This paper evaluates the performance of a 1.5TR split air conditioner with and without evaporative cooling of condenser. The temperature reduction of the ambient air by evaporative cooling is computed for different weather conditions. The Box–Behnken design (BBD) methodology is applied to predict the interaction of parameters and optimize the results while analysis of variance (ANOVA) provides the fitness and validation of the model. It is observed that the proposed system significantly improves the performance of air conditioner. The maximum COP enhancement (9.37%) is achieved for the ambient temperature drop of 6.7 °C for the optimum operating conditions. The corresponding improvement in the refrigerating effect and savings in work input are obtained as 9.52% and 8.54%, respectively. The proposed system is estimated to pay back within 2.9 years. The sustainability indices indicate the environmental improvement by the proposed system.

Keywords Evaporative cooling · Box–Behnken design · ANOVA · COP · Sustainability index

1 Introduction

Most parts of India are very hot and humid from April to September. The ambient temperature and relative humidity during these months lie between 30–44 °C and 18.1–80.1%, respectively [1], making it uncomfortable for humans to work efficiently. Therefore, comfort-air conditioning is necessary to overcome the harmful effects of hot weather conditions.

For comfort-air conditioning, one may look for either direct evaporative cooling options or air conditioning based on vapor compression systems, depending on the prevailing level of humidity. Direct evaporative coolers are ineffective in providing comfort when the ambient humidity levels are high [2]. The vapor compression refrigeration-based air conditioning systems are suitable for humid climatic conditions. The increase in the use of split air conditioners (SAC) has put a great impact on total energy consumption worldwide. Refrigeration and air conditioning account for 20% of gross energy consumption all over the world which also produces about 7.8% of global greenhouse gas (GHG) emissions [3]. Global warming caused due to GHG emissions is a serious concern in today's scenario. One of the methods to address the environmental global warming issue is to increase the efficiency of SAC systems. The coefficient of performance (COP) is a measure of the energy efficiency of a refrigeration system. Several studies explored the methods for reducing the energy in these systems. Energy reduction could be achieved by replacing less efficient old air conditioners with more efficient units, installing an additional heat exchanger, or reducing the heat exchanger area [4–6].

In a vapor compression cycle, the cooling capacity and COP can be enhanced by liquid cooling below saturation temperature [7]. Various methods adopted for subcooling include internal heat exchanger (IHX), mechanical subcooling by providing auxiliary arrangement, etc. Generally, the liquid refrigerant gets subcooled, while flowing in the liquid line of a vapor compression refrigeration system (VCRS), by dissipating heat to the encompassing air

Technical Editor: Luben Cabezas-Gómez.

✉ S. K. Gupta
sunilgupta_me@yahoo.co.in

[1] Department of Mechanical Engineering, Delhi Technological University, New Delhi 110042, India

Published online: 09 November 2023

or during storage in the liquid receiver [8]. Liquid–vapor or internal heat exchangers are advantageous in refrigeration system performance enhancement [9]. Yau and Pean [10] employed a suction line heat exchanger (SLHX) in a split air-cooled ducted air conditioner of 29.3 kW capacity using R22 refrigerant. The influence of weather variation on the performance of the system was investigated. They concluded that COP was reduced by 2% for every 1℃ rise in the ambient temperature. Pottker et al. [7] studied the effect of condenser subcooling on the performance of a 3.5 kW split air conditioning system. They showed that subcooling maximized COP with a compromise between refrigerating effect and specific compressor work. Yang and Yeh [11] used a subcooler for a 400 kW vapor compression refrigeration system for improving system efficiency. The cooling capacity of the system was reported to increase by increasing the degree of subcooling. The optimal degrees of subcooling with the cost-saving aim ranged from 2 to 6℃ for R134a, R22, R410A, and R717. Low cooling water temperatures achieved significant improvement in the condenser performance.

Numerous theoretical and experimental studies are available in the literature which reports the use of environment-friendly evaporative cooling in air conditioning to lower the energy expenditure and consequently the operational cost. Yang et al. (2021) [12] employed a spray evaporative cooling with a 60 kW air conditioner condenser. For the ambient temperature between 31 and 48℃, the spray cooling decreased the air temperature by 6.7–13.1℃ and increased the relative humidity by 28.3–31.3% at the condenser entry. A larger temperature drop was realized for the higher outside temperature. It was concluded that in hot climates, a 22% cut in energy expenditure and about a 42.6% increase in COP could be obtained. In another attempt to enhance the 2650W capacity inverter SAC performance by reducing the ambient air temperature at the condenser inlet, Yang et al. [13] sprayed the condensate water from the evaporator through an atomization cooling element (ACE) installed on the outdoor unit fan. For complete evaporation of the condensate water, the condenser inlet air temperature was decreased by 6.02℃. At the ambient temperature of 35℃, the improvement achieved in cooling capacity, power input, and EER were 1.3%, (-)0.9%, and 2.1%, respectively. The corresponding variations at 43℃ were reported to be 8.1%, (-)9.5%, and 20%, respectively.

Alaa Ruhma Al-Badri [14] investigated the effects of inlet air DBT, RH, water subcooling, and air/water ratio on DEC performance. The performance was significantly increased with relatively dry air at high temperatures. Even in very humid climate regions, DEC could be applied by lowering the chilled water temperature and reducing the air-to-water mass flow ratio. Wang et al. [15] employed a split air conditioning system of 5.3 kW to 7 kW refrigerating capacity using evaporative cooling to decrease the condenser inlet air temperature by 2.4℃ to 6.6℃. The consequent increase of refrigerant liquid mass entering the evaporator enhanced the COP from 6.1% to 18%. The cost-effective air temperature at the condenser inlet was determined to be about 33.1℃. Sarntichartsak and Thepa [16] tested the performance of the R410A inverter air conditioner system (3.5 kW capacity) with an evaporatively cooling unit by varying the input conditions. The COP was improved by 18.32% at the lowest compressor frequency (30 Hz) and water flow rate of 200 l/h. Ketwong et al. [17] proposed an evaporatively cooled condenser for performance enhancement of 1 TR air conditioner in hot–dry and hot–humid climates. They reported that the energy efficiency ratios of the system with direct evaporative cooling for hot–dry and hot–humid weather conditions ranged from 3.40 to 4.22 and 3.30 to 3.94, respectively, compared to 3.01 for the conventional unit. Ahmet et al. [18] compared the performance of a split-type air conditioner with an evaporatively cooled condenser with a conventional type air-cooled condenser for different ambient conditions. The COP of the system with the evaporative cooling of the condenser increased by 10.2%–35.3%. The ambient relative humidity and temperature were found to impact the COP significantly. Hajidavalloo & Eghtedari [19] installed an evaporative cooler in the outdoor unit of a 1.5ton split air conditioner. Their experiments at different ambient temperatures (35 to 49℃) and RH (12 to 40%) showed that the reduction in power consumption and improvement in COP could be around 20% and 50%, respectively. Chauhan & Rajput [1] conducted an experiment on an air conditioner with a 6.5 kW capacity, incorporating a direct evaporative cooler in very hot and dry condition only (DBT > 34℃ and WBT = 20 to 23℃). They achieved maximum energy savings of 23.8% with an outdoor temperature of 43.3℃ and a relative humidity of 18.1%. However, no energy savings observed in humid conditions. Martínez et al. [20] used cooling pads of different thicknesses for increasing the energy efficiency with direct evaporative cooling of the condenser of a 2500W capacity air conditioner in the ambient conditions of Spain ($T_a < 30°C$). They achieved significant decrease in the condenser inlet temperature with 11.4% reduction in power consumption and 10.6% increase in COP.

The literature shows that the effect of varying ambient conditions on the performance of the air conditioning system is investigated in very narrow range of either temperature or relative humidity or both. This work presents a comprehensive analysis of the performance of a split air conditioner using an evaporative cooler in a wide range of ambient temperature (30–44℃) and relative humidity (20–80%). The novelty of this work lies in the integration of the condensate water channel to DEC section and placing a single axial fan in between the DEC section and the condenser. This configuration could save considerable power by eliminating

a separate fan for the cooler. Additionally, a sustainability index has been developed to compare the environmental impact of the proposed system in different ambient conditions. An optimization study is also carried out to find out the optimal conditions for obtaining maximum COP. Genetron Properties software is used to obtain thermodynamic data and the analyses are performed using Design Expert software. The proposed system is quite suitable for its implementation in the real environments of extremely hot–dry and moderately hot–humid regions.

2 System selection and methodology adopted

A split air conditioner (SAC) with refrigerant R410A is employed for the study. The conventional system consists of evaporator in the indoor and air-cooled condensing unit in the outdoor (Fig. 1). The condensing unit of the proposed system is integrated with a direct evaporative cooler (DEC) as shown in Fig. 2. The temperature reduction of ambient air in DEC is computed by a developed model. The cycle performance results without and with evaporative cooling are simulated using Genetron Properties software [21]. The Box–Behnken Design methodology is adopted to maximize the output by optimizing the input variables. The exergy analysis is used to obtain the sustainability indices which indicate the effect of different ambient conditions on the environment.

The heat absorbed by the evaporator from the cooling room is dissipated outside by the condenser. The condenser fan sucks the ambient air to pass through the condenser. The condenser releases the heat of refrigerant into the air. The heat loss to the surroundings is neglected. In the summer season, the high ambient temperature badly affects the

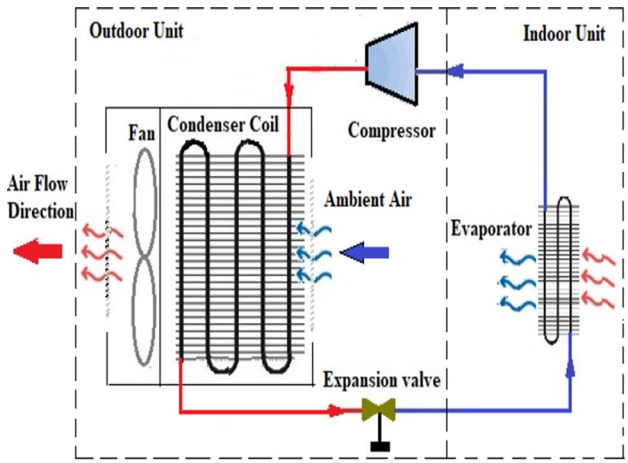


Fig. 1 Schematic diagram of conventional SAC

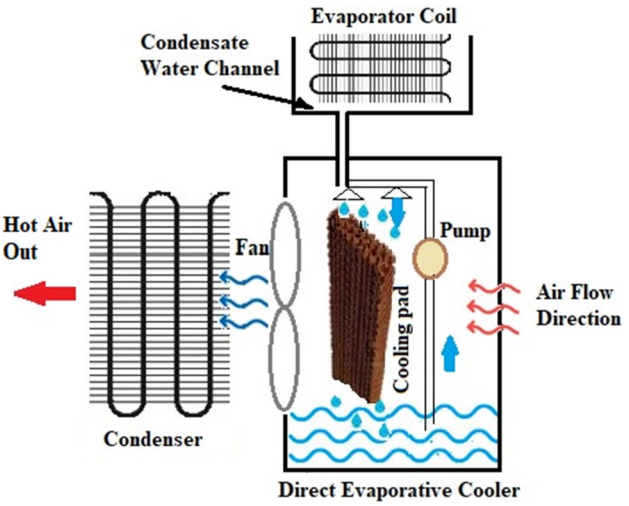


Fig. 2 Schematic diagram of DEC-cooled condenser

condenser performance. The increasing ambient temperature increases the condenser temperature, which results in a higher compression ratio, and consequently higher power consumption. During evaporative cooling, the water absorbs the sensible heat of the air and evaporates, making the surrounding air cooler. The air wet-bulb temperature is the saturation limit of water evaporation. With the decrease in the inlet air temperature, the heat released by the condenser increases. The evaporative cooling reduces the condenser pressure below saturation pressure, leading to subcooling [15].

2.1 DEC modeling

The water is continuously sprinkled to keep the cooling pad completely moistened during the DEC operation. The water flow rate is such that it maintains the water temperature at the inlet equal to outlet temperature, T_w across the cooling pad. In this way, the sensible heat is minimized which represents the true system behavior in real environment situations. Fixed thermal properties of air and water are assumed across the completely wet cooling pad. The water evaporates after gaining latent heat from the air which loses its sensible heat and is cooled. The modeling considers the assumptions that the air/water heat exchange due to conduction, heat loss to the surroundings, and changes in potential and kinetic energy are negligible. The following equation describes the energy and mass balance for the thermal exchange between air and water across the cooling pad (Fig. 3a) under steady-state conditions.

$$\sum \dot{m}_i h_i = \sum \dot{m}_o h_o \tag{1}$$

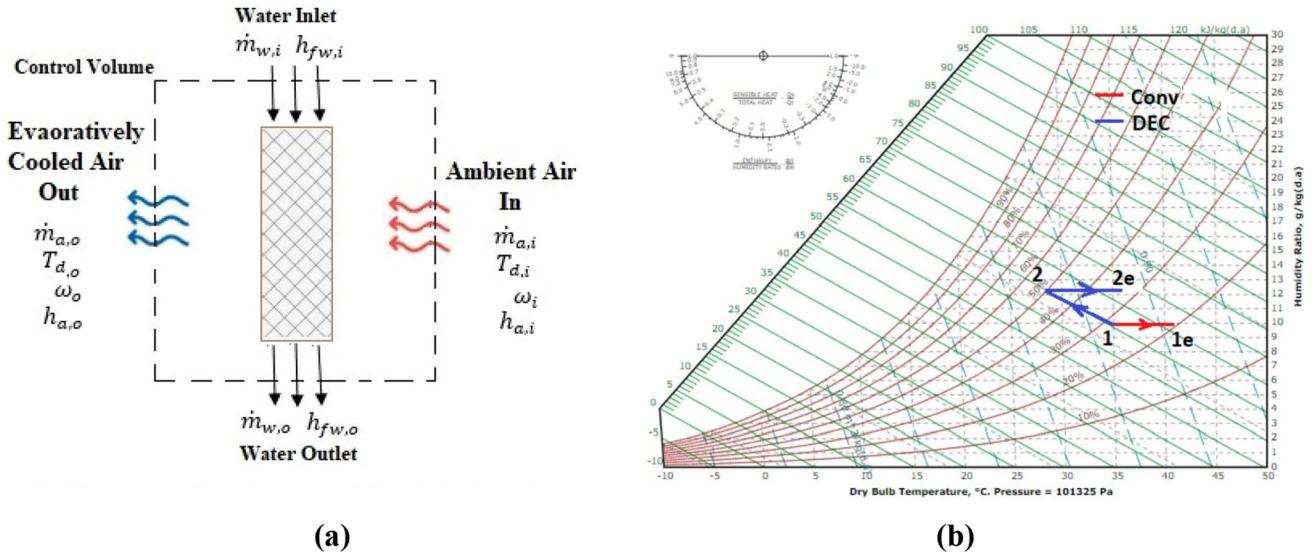


Fig. 3 **a** Heat and mass balance across the cooling pad, **b** Psychrometric changes in air through the conventional condenser and DEC-cooled condenser

The loss in sensible heat of the air in passing through the small depth dx of the cooling pad is obtained as [22]:

$$d\dot{q}_s = -\dot{m}_a.C_{p,a}.dT = -h_c.dF.(T_a - T_w) \tag{2}$$

Here, $h_c = 42.9 \times V^n$ and the pad configuration factor,

$$dF = \xi.B.H.dx$$

ξ is the pore surface coefficient for the pad material of width B and height H.

Also, the water evaporation adds latent heat into the air as given by Eq. (3),

$$d\dot{q}_l = d\dot{m}_w.L_w = h_m.dF.(\omega_s - \omega_a) \tag{3}$$

The convective heat transfer coefficient,h_c, and mass transfer coefficient,h_m, of moist air on the surface of water film are constant. The water loss due to evaporation is considered as the water consumption of the DEC. Assuming that there is no heat transfer from the surroundings, the sensible heat lost by the air equals the latent heat added in the form of water vapor.

$$d\dot{q}_s = d\dot{q}_l \tag{4}$$

Considering Lewis number, L_e=1, we get

$$\frac{h_c}{h_m} = C_{p,a} \tag{5}$$

By solving Eq. (1) to (5), the temperature of outlet air is obtained by Eq. (6)

$$T_{a,dec} = T_s + \left(T_a - T_w\right).exp\left(\frac{-h_c.\xi.\mathrm{B.L.}\delta}{\dot{m}_a.C_{p,a}}\right) \tag{6}$$

The cooling and humidification of air nearly follows an isenthalpic process. The water temperature, T_w, can be assumed approximately equal to wet-bulb temperature, T_s, of air [22]. The value of T_s for varying relative humidity is calculated from the psychrometric calculator at different dry-bulb temperatures (T_a). Then, evaporative cooling efficiency can be expressed in terms of pad material characteristics as:

$$\eta = 1 - \exp\left(-\alpha\frac{\delta}{V^{0.35}}\right) \tag{7}$$

where $\alpha = \frac{A.\xi}{\rho_a.C_{p,a}}$

In the above expression, δ is the thickness of the cooling pad and V is the mean air velocity. According to Wu et al. [22], the values of A, V, and n in the above equation are 25.2, 1 ms^{-1}, and 0.65, respectively, for the pad material. The cooling pad material is chosen such that it can retain water and allow the passage of the incoming air with minimum pressure drop.

2.1.1 Determination of process in conventionally cooled and evaporatively cooled condensers

In very hot–humid air, the high humidity level, or low wet-bulb depression (the difference between dry and wet-bulb temperatures), the DEC effectiveness reduces drastically [14]. Before analyzing the performance of DEC-SAC, the

effect of ambient air conditions (T_a *and RH*) on the evaporative cooler is investigated.

The process lines 1-1e and 2-2e on the psychometric chart (Fig. 3b) show the enthalpy changes through the conventional and evaporatively cooled condensers, respectively. The condenser dissipates the heat of the refrigerant vapor to the outside air and converts it into a liquid phase. The latent heat transfer by water evaporation to the air passing through the pad as represented by process lines 1 to 2, is an adiabatic process (constant enthalpy). Therefore, the DEC reduces the air temperature from 1 to 2 with a corresponding rise in specific humidity from ω_1 to ω_2. The cool air leaving the DEC is then flown over the condenser coil where the air temperature rises in response to the sensible heat gain from the refrigerant, as shown by the process line 2-2e. The air during the heat exchanges in the condensers of conventional SAC and DEC-SAC follows the process lines 1-1e and 1–2-2e, respectively.

Neglecting heat flux transferred from surroundings to the DEC, the air is cooled and humidified with constant enthalpy, i.e., the air loses a certain amount of sensible heat as well as gains an equal amount of latent heat of water evaporation.

2.1.2 Performance of DEC

The DEC performance is assessed for the design parameters appended in Table 1. The inlet water temperature T_w is usually slightly higher than the saturation temperature of air, so it is taken as 5℃ more than T_s. A water circulation pump of 12W power (commonly available in market) provides a constant flow rate of 1.2 m^3h^{-1}.

Figure 4 depicts the effect of T_a on the variation of temperature reduction ($\Delta T_{a,dec}$) for different saturation conditions, T_s according to variable relative humidity (20–80%). For the temperature rise from 30 to 44 °C, the drop in dry-bulb temperature increases from 3.5 to 13.7 °C. It is seen that a higher ambient temperature is favorable for a greater reduction in the cooling pad outlet temperature. In the DEC, the air exchanges heat with the water before passing through the condenser. A higher air temperature increases the water—air temperature differential at the pad and consequently the air-cooling effect improves. The temperature

Table 1 Design parameters for DEC

T_a(℃)[a]	(30–44)
RH(%)[a]	(20–80)
T_w(℃)[b]	(T_s + 5°C)
MFR $\left(\frac{\dot{m}_w}{\dot{m}_a}\right)$[c]	1
Saturation efficiency, ε[b]	0.65

[a][1], [b][22], [c][14]

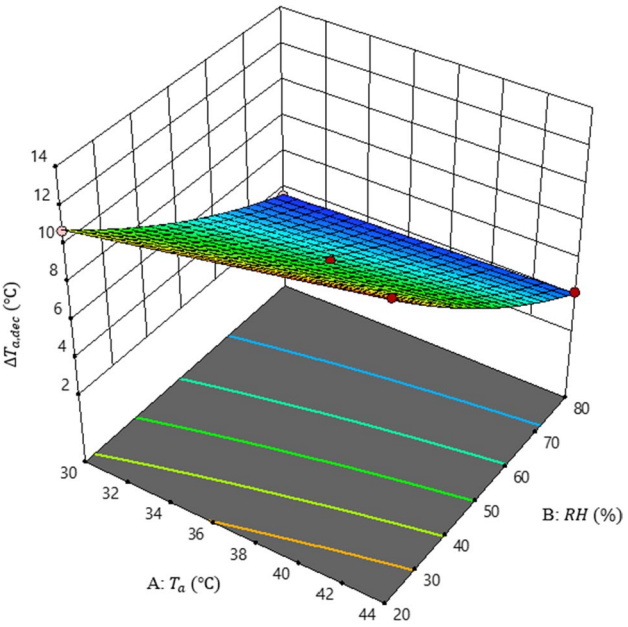


Fig. 4 Effect of ambient conditions on evaporative cooling degree

reduction is more at higher ambient temperature than that at lower temperature.

It is also seen in the figure that a higher humidity level does not have much effect on the decrease in dry-bulb temperature. This is so because as the air approaches saturation condition, the sensible heat reduction of air falls due to lower water evaporation rate. As a result, $\Delta T_{a,dec}$ decreases for higher *RH* (80%) compared with lower *RH* (20%) in the whole range of inlet air temperature.

2.2 Refrigeration cycle analysis

The refrigeration cycles without and with evaporative cooling on the temperature-enthalpy curve can see in Fig. 5.

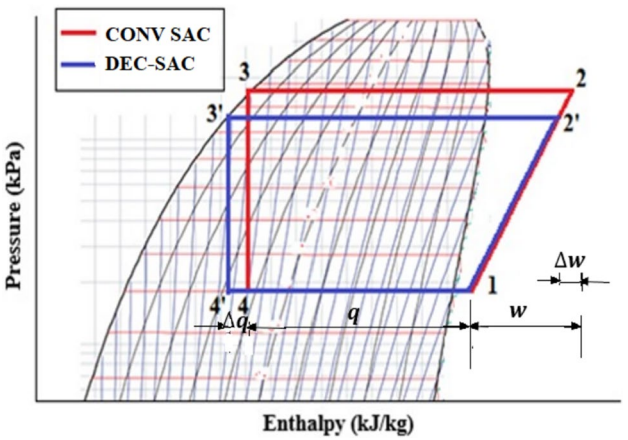


Fig. 5 P–h curve for CONV SAC and DEC-SAC

The evaporator and condenser interact with the surrounding room air and ambient air, respectively. The conventional air-cooled and evaporatively cooled cycles are represented by 1–2-3–4-1 and 1–2'-3'-4'-1, respectively. The ideal condenser and evaporator have been considered with unity efficiency and without heat and pressure losses. The performance of the thermodynamic cycle is assessed by carrying out energy and exergy analysis as follows.

The coefficient of performance (COP) is the prime thermodynamic parameter used to compare the efficiency of different refrigerants. The COP of the cycle without evaporative cooling is given by Eq. (8).

$$COP = \frac{q}{w} \tag{8}$$

where $'q'$ is the amount of cooling produced and $'w'$ is the compressor work.

The COP' of the cycle with evaporative cooling can be represented by Eq. (9).

$$COP' = COP\left[\frac{1 + \Delta q/q}{\Delta w/w}\right] \tag{9}$$

Enhancement in cooling effect due to evaporative cooling, $\Delta q/q$ is expressed by Eq. (10)

$$\frac{\Delta q}{q} = \frac{(h_4 - h_{4\prime})}{(h_1 - h_4)} \tag{10}$$

The lower compression ratio due to subcooling leads to a reduction in power consumption. The reduction in the specific work of compression relative to that without evaporative cooling is expressed by Eq. (11),

$$\frac{\Delta w}{w} = \frac{(h_2 - h_{2'})}{(h_2 - h_1)} \tag{11}$$

2.2.1 Exergy analysis and sustainability index

The exergy analysis evaluates the effect of irreversibility in the process. This analysis is useful to identify the components wasting more energy as compared to other components.

Referring to Fig. 5, the contribution of system components to waste energy or exergy destruction ($\dot{E}_D$) is assessed according to Eqs. (12) to (15)[23].

$$\dot{E}_{D,e} = \dot{m}_r\left\{\left(h_{4'} - h_1\right) - T_0\left(s_{4'} - s_1\right)\right\} + \dot{Q}_e\left[1 - \left(\frac{T_0}{T_R}\right)\right] \tag{12}$$

$$\dot{E}_{D,c} = \dot{W}_c + \dot{m}_r\left\{\left(h_1 - h_{2'}\right) - T_0\left(s_1 - s_{2'}\right)\right\} = \dot{m}_r T_0(s_{2'} - s_1) \tag{13}$$

$$\dot{E}_{D,k} = \dot{m}_r\left\{\left(h_{2'} - h_{3'}\right) - T_0\left(s_{2'} - s_{3'}\right)\right\} - \dot{Q}_k\left[1 - \left(\frac{T_0}{T_k}\right)\right] \tag{14}$$

$$\dot{E}_{D,v} = \dot{m}_r T_0(s_{4'} - s_{3'}) \tag{15}$$

The total exergy destruction in the system is obtained by Eq. (16).

$$\dot{E}_{D,tot} = \dot{E}_{D,e} + \dot{E}_{D,c} + \dot{E}_{D,k} + \dot{E}_{D,v} \tag{16}$$

The emission of greenhouse gases (mainly CO_2) causes global warming, which is responsible for climate change. The direct and indirect contribution of carbon dioxide poses a potential threat to the environment. Thus, Low-carbon energy options are helpful to lower the impact on the environment.

The sustainability index (SI) defined as the ratio of energy input ($\dot{w}_c$) and the exergy destruction ($\dot{E}_{D,tot}$)[24], is used to compare the environmental degradation by different degree of evaporative cooling. Therefore, sustainability index is given by Eq. (17) as,

$$\text{Sustainability index} = \frac{1}{(1 - \psi)} \tag{17}$$

$$\text{Where } \psi \text{ is Exergy ratio} = \frac{\dot{w} - \dot{E}_{D,tot}}{\dot{w}}$$

A high index indicates a lower environmental impact due to the reduced waste energy, consequent upon the lower greenhouse gas emissions.

2.3 Box–Behnken design methodology

DoE (Design of Experiment) is a technique used to reduce the cost and time involved in experimentation while ensuring logical responses and acceptable products. Design of experiments also helps to design and develop steady processes to withstand different environmental conditions and resources. The Box–Behnken design (BBD) is applied to obtain numerical values following the quadratic model. BBD is considered to be a very useful design method. This technique outlines the interactions between the individual factors in the polynomial relation, which supports the projected outcomes within prescribed limits. BBD is not only beneficial to identify the independent parameters and their relationships, but it also optimizes the effective parameters.

The numerical optimization technique is used in this work to optimize the input variables of the SAC system. The Design Expert software helps decide the combination of independent inputs for producing an optimized solution.

2.3.1 Analysis of variance (ANOVA)

The analysis of variance (ANOVA) is applied to establish the empirical correlation between multiple inputs and outputs. The ANOVA is a statistical technique to estimate the sizeable difference in the means of two or more groups. ANOVA evaluates the impact of one or more factors by comparing the means of different experiments. The numerical simulation data fits into a quadratic polynomial equation to represent all the linear interactions of the factor shown in Eq. 9.

$$Y = \beta_0 + \beta_1 \sum_{i=1}^{3} X_i + \beta_2 \sum_{i=1}^{3} X_i^2 + \beta_3 \sum_{ij=1}^{3} X_i X_j + \varepsilon \quad (18)$$

where Y is the output parameter, β_0 is a constant, β_1 and β_2 are the linear and second-order reactions of parameter X_i, β_3 is linear interaction of X_{ij}, and ε represents errors.

2.4 Model verification

Figure 6 compares the present model with the experimental results of Çag & Ali, [18] under two similar ambient conditions of warm–dry and cold–wet climates. The refrigerant in both cases is R410A. The deviations of 3.68% and 1.96% in *COP* enhancement for warm–dry and cold–wet conditions, respectively, show a good agreement between simulated and experimental values.

3 Results and discussion

3.1 Model description

A 5.25 kW cooling capacity split air conditioner (SAC) with R410A as working substance is used for the analysis. The operating variables for split air conditioner are displayed in Table 2. The investigations are carried out at evaporator temperature of 3 to 12 °C [25]. The ambient temperature and

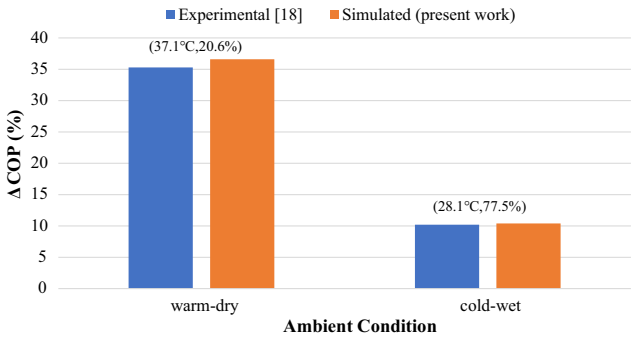


Fig. 6 Comparison of COP enhancement in Present work and Çag & Ali [18] for different ambient conditions (T_a, RH)

Table 2 Design variables for SAC

Room temperature, T_R [a]	18–27℃
Superheat at the evaporator outlet[b]	5℃
Compressor isentropic efficiency[b]	70%
Compressor volumetric efficiency[b]	95%
$(T_R - T_e)$ [c]	15℃
$(T_k - T_a)$ [c]	15℃

[a][25], [b]as per product catalog, [c][26]

relative humidity are considered in the range of 30 to 44 °C and 20% to 80%, respectively [1]. The air inlet temperature is assumed to bear a difference of 15°Cwith both evaporator and condenser [26]. Fixed values for air density and specific heat are assumed as 1.3 kg m^{-3} and 1.0 kJ $kg^{-1}℃^{-1}$, respectively. The influence of evaporative cooling on COP can be determined if the actual values of condenser inlet air temperature drop ($\Delta T_{a,dec}$) for different ambient conditions are known. For this purpose, Eq. (6) of the proposed model has been used.

3.2 Effect of evaporative cooling on the performance of SAC

3.2.1 Refrigerating effect

The subcooled liquid at lower enthalpy enters the evaporator, thereby increasing the refrigerating effect per unit mass. So, the refrigerant mass velocity for the subcooled cycle reduces per unit capacity. Figure 7a–c represents the enhancement in refrigerating effect, decrease in the work consumption, and the improvement in COP as a result of evaporative cooling. A large increase in refrigerating effect results from the increase in evaporative cooling degree. The behavior of refrigerating effect predicts the volume of mass flow rate and consequently the size of the air conditioner. A higher refrigerating effect results in lowering of mass flow rate and accordingly a smaller system for the same cooling capacity as that for the system without evaporative cooling.

3.2.2 Compression work

The refrigerant vapor-specific volume at the compression suction is identical for both cycles, i.e., without (conventional) and with (DEC) evaporative cooling. However, less mass flows in the subcooled cycle than the non-subcooled cycle. Thus, the compressor handles less vapor volume for a unit capacity. Because of the lower vapor volume, the subcooled cycle needs smaller compressor displacement than that required without evaporative cooling. Figure 7b shows the percentage decrease in the specific work consumption as a result of evaporative cooling.

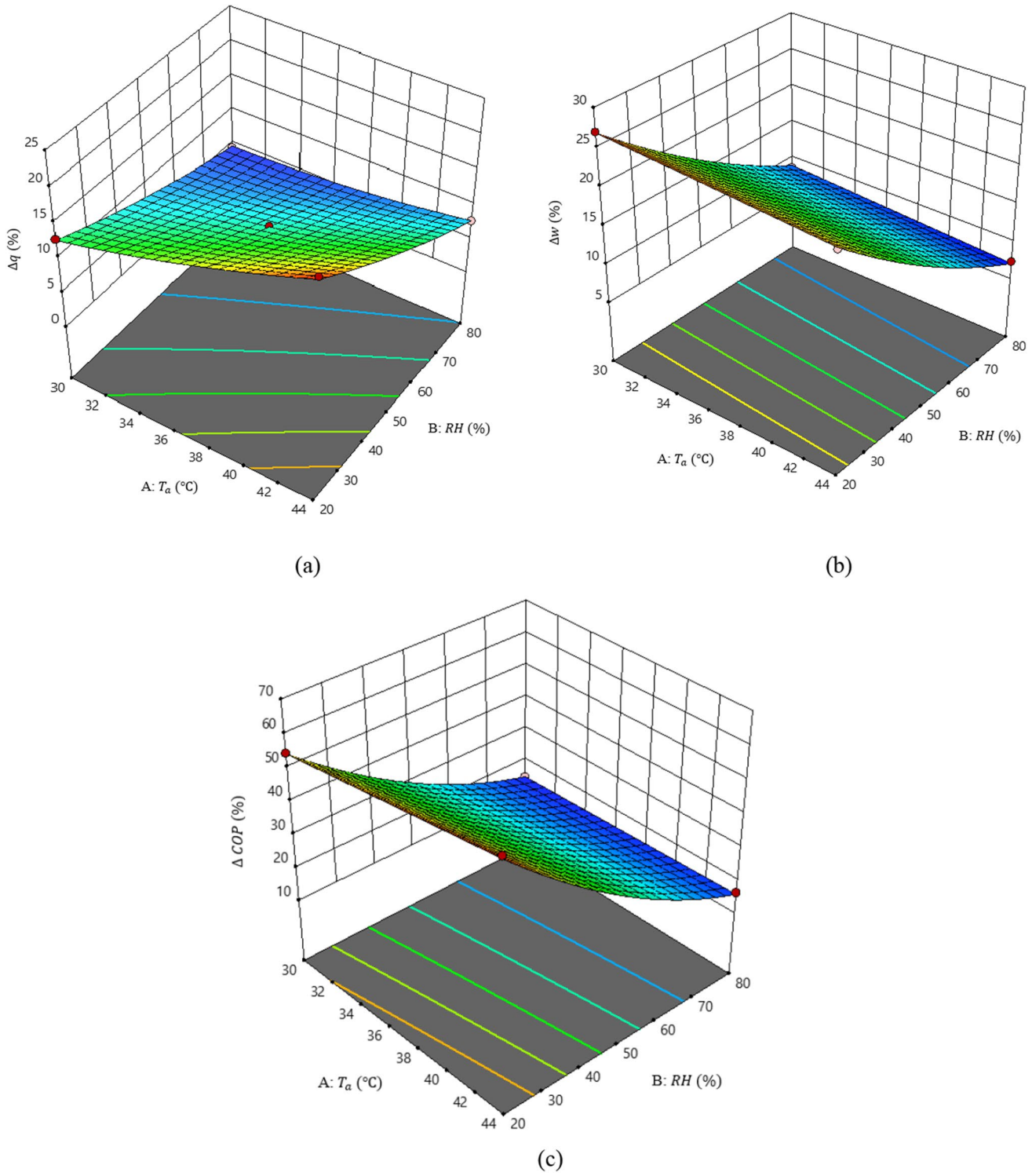


Fig. 7 **a** Enhancement in refrigerating effect, **b** reduction in specific work input, and **c.** improvement in COP due to evaporative cooling in different ambient conditions for DEC-SAC

Increased evaporative cooling reduces the saturation condenser temperature. Consequently, the compression work decreases due to the lower compression ratio as the evaporator temperature is fixed. The increase is steeper at higher degrees of evaporative cooling. The reduction in input work of DEC-SAC as compared to the conventional SAC is maximum for the highest ambient temperature (44℃) and lowest RH (20%).

3.2.3 Coefficient of performance (COP)

Figure 7c displays the improvement in COP with the increase in evaporative cooling. The cooling capacity is maintained at 5.25 kW to only improve the COP. As the evaporative cooling increases, a reduction in the two-phase area takes place to make room for the subcooled liquid that is formed in the condenser. In the subcooled zone, the air-refrigerant temperature differential and the coefficient of heat transfer are smaller than in the liquid–vapor zone. The saturation temperature rise increases the temperature difference between air and refrigerant in the two-phase area which compensates for the drop in condenser heat transfer effectiveness. The relative decrease in the compressor work is obtained as mentioned before. Meanwhile, the decrease in condenser outlet temperature increases the refrigerating effect with a larger enthalpy variation in the evaporator. The above results establish that the evaporative cooling improves the performance of the system.

The computed values of increase in refrigerating effect (Δq), reduction in specific work (Δw), and enhancement in COP for varying relative humidities are displayed in Table 3.

The highest enhancement in refrigerating effect (13.05%) occurs when RH is 20% with improvement of 11.58% in the input work. Also, the lowest enhancement in refrigerating effect is 4.69% when the RH is maximum in the considered range, i.e., 80%, obviously due to smallest temperature drop of ambient air and consequently condenser temperature. The corresponding decrease in input work is 4.5%. The highest and lowest improvements in COP are 13.08% (20%RH) and 4.71% (80%RH). These results confirm that evaporative cooling is beneficial in enhancing COP.

3.3 Model fitness

Results obtained from ANOVA are displayed in Table 4 for significant terms. The model F values imply that the model is significant. The large F values of 44.45, 52.55, and 44.67 corresponding to improvement in refrigerating effect, work input, and COP with very low p values confirm the good fit of the model for DEC-SAC.

P values less than 0.0500 indicate that model terms are significant. In this case, A (T_a) and B (RH) are significant model terms. The outcomes depict a higher confidence level (greater than 95%) for R^2 and R_a^2 values to accept the model [27]. Predicted R^2 (0.7841) are in reasonably good agreement with Adjusted R^2 (0.8619) values for the COP. Similar is the case for other responses. The model is acceptable as the variations between R_a^2 and R_p^2 are less than 0.1% [28]. Adeq. Precision measures the signal-to-noise (S/N) ratio. For validating the fit, the S/N ratio is satisfactory (> 4) for the present analysis [29]. The low values of SD and mean ratio (coefficient of variation) suggest that the model is accurate.

Diagnostic plots (Fig. 8a–c) suggest that the model has a good fit. The normal probability plot for residuals with the accumulation of residuals about the central line denotes the applicability of the analysis. The above results suggest that the Box–Behnken design applies effectively to the present model. Intending to investigate the parameters

Table 3 Calculated values of the performance parameters for condenser inlet air temperature, T_a=30℃ and evaporator temperature, T_e=6℃

$RH(\%)$	20	30	40	50	60	70	80
$\Delta T_{a,dec}$(°C)	10.8	9.4	8.1	6.9	5.7	4.7	3.5
$\Delta q(\%)$	13.05	11.46	9.95	8.54	7.12	5.91	4.69
$\Delta w(\%)$(-)	11.58	10.27	9.08	7.89	6.64	5.57	4.5
$\frac{COP_{DEC}-COP_{CV}}{COP_{CV}}(\%)$	13.08	11.50	9.98	8.58	7.15	5.93	4.71

Table 4 ANOVA model analysis

Source	Δq		Δw		$\frac{COP_{DEC}-COP_{CV}}{COP_{CV}}$	
	F value	p value	F value	p value	F value	p value
Model	44.45	< 0.0001	52.55	< 0.0001	44.67	< 0.0001
A	35.14	< 0.0001	40.85	< 0.0001	35.47	< 0.0001
B	53.75	< 0.0001	64.24	< 0.0001	53.88	< 0.0001
Lack of fit	1.63		0.9352		1.63	
R^2	0.8811		0.8811		0.8816	
R_a^2	0.8612		0.8612		0.8619	
R_p^2	0.7831		0.7831		0.7841	
Adq.Pres	20.9655		20.9655		21.0223	
SD (σ^2)	1.16		0.8828		1.16	
Mean	9.57		8.68		9.58	

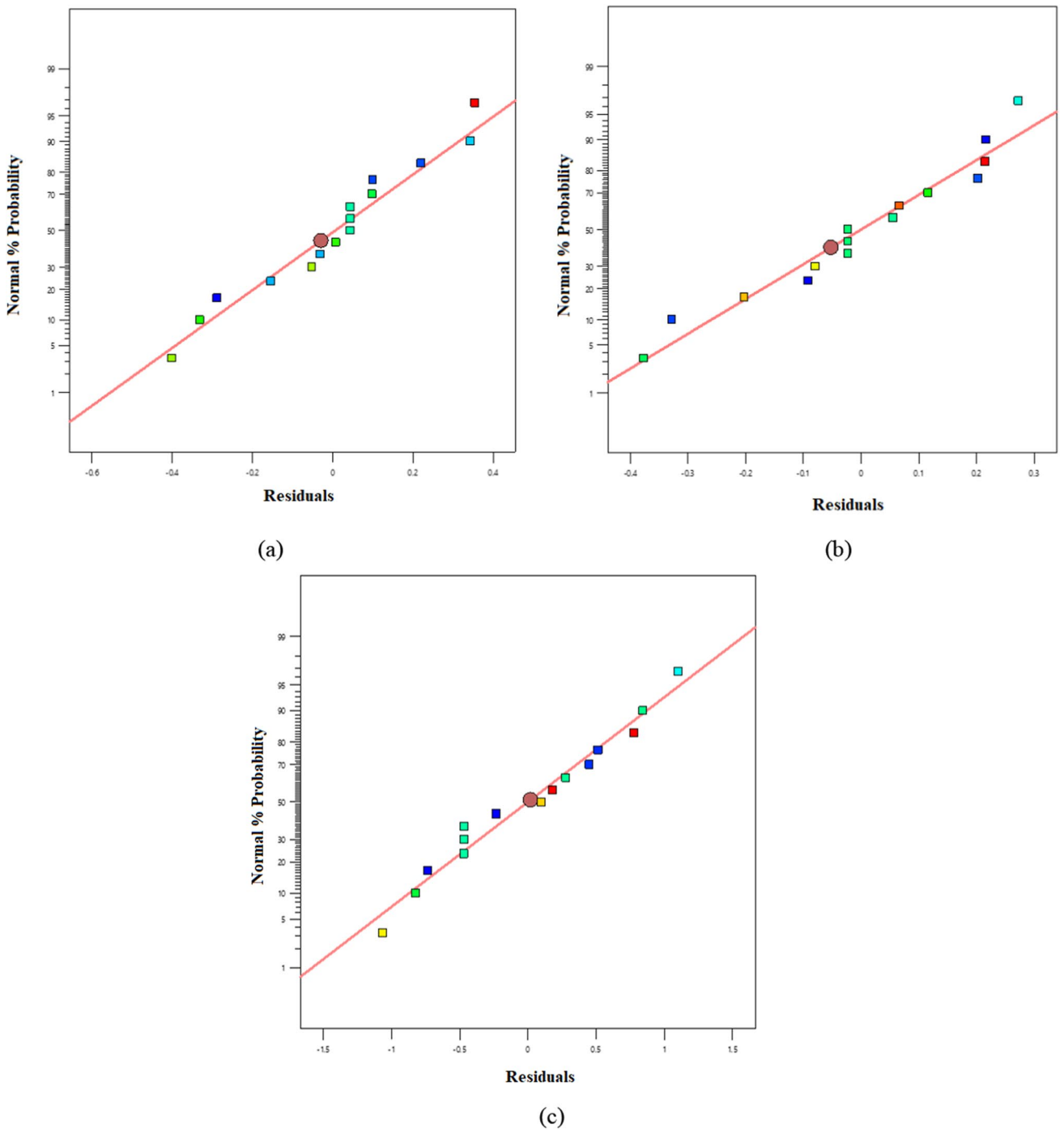


Fig. 8 Residuals of **a** Δq, **b** Δw, and **c** ΔCOP for DEC-SAC

affecting the refrigerating effect, specific work, and COP of the DEC-SAC system, the BBD stipulated simulations result in improving the cycle for varying outside temperature, relative humidity, and evaporator temperature.

3.4 Optimization using Box–Behnken design methodology

Box–Behnken design, inbuilt in Design Expert software, optimizes the three input variables -$\boldsymbol{T_a}$, RH and $\boldsymbol{T_e}$- to

Table 5 Optimization criteria

Parameter	Goal	Lower limit	Upper limit
A:T_a(℃)	maximize	30	44
B: RH%	is fixed = 50	20	80
C:T_e(℃)	minimize	3	12
q(kJ/kg)	maximize	135.36	175.7
w(kJ/kg)	minimize	28.13	51.93
COP	maximize	2.72	6.02

maximize the objective functions COP and refrigerating effect, and minimize the work consumption simultaneously. The COP measures the performance of the cycle and provides a basis for effective energy utilization. The refrigerating effect decides the capacity of the system. The input compression work predicts the size of the compressor. For higher input work requirements, the size of the compressor becomes larger which is linked to the cost of the SAC system. The optimum response values are obtained using BBD technique. Equal importance is given to all control factors and the responses. Table 5 represents the optimization criteria and Table 6 results for DEC-SAC.

3.5 Sustainability index

A high value of the sustainability index indicates low degradation of the environment. Figure 9a, b shows the effect of ambient temperature and relative humidity (at fixed T_e), respectively, on the sustainability indices for non-DEC and DEC-SAC. It can be seen that SI increases for DEC with the rise in ambient temperature. The index of DEC-SAC is more than that of non-DEC at higher T_a (> 38℃). It is probably due to increase in evaporative cooling degree with the rise in ambient temperature which results in more temperature differential between the condenser and inlet air. From Fig. 9b, it is noticed that index of DEC-SAC increases with RH at high T_e (12℃) but reverse is the case when T_e is low (3℃).

3.6 Water requirement for the proposed system

From the above discussion, it is evident that DEC-SAC is best suited in hot–dry climates. Referring to Eq. (19), water evaporation rate for the extreme situation (high $T_{a,}$ and low RH) is computed as 8.5 L/h.

$$d\dot{m}_w = \frac{h_c.\xi.B.H.\delta.(T_a - T_w)}{L_w} \times 3600\text{kgh} - 1 \tag{19}$$

Also, the condensate water produced from the indoors is approximated as 0.65 L/h according to Eq. (20) [13].

$$\dot{m}_{condensate} = \rho V_a(\omega_{in} - \omega_{out})\text{kgh} - 1 \tag{20}$$

where V_a is the circulating air quantity ($m^3.h^{-1}$), and ω_{in} and ω_{out} are the moisture contents of the indoor air and supply air, respectively. As ω_{in} increases with the rise in relative humidity, the quantity of condensate water significantly increases which further reduces the consumption of make-up water. Thus, with very low requirement of fresh water (about 40–60 L per day), the proposed system could be well suited for the very hot and/or humid countries like India, Saudi Arabia, Brazil etc., excluding some regions which lack even adequate drinking water. Also, the annual energy consumption (≈ 2kWh) for pumping such a small quantity of water can be ignored against the annual energy reductions (≈ 224kWh) offered by the optimized system. According to an estimate, the annual energy savings with the proposed system in India can be around 4.85 TWh [30]. The energy

Table 6 Optimized results for DEC-SAC

T_a (℃)	RH (%)	T_e (℃)	($\Delta T_{a,dec}$) (℃)	q (kJ/kg)	w (kJ/kg)	COP
34.5	50	7.3	7.2	160.8	37.8	4.27

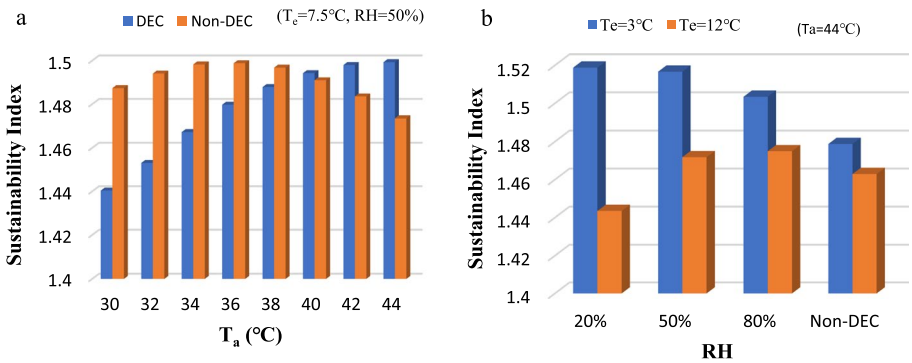


Fig. 9 Comparative sustainability index with and without evaporative cooling: **a** T_a varying, **b** RH varying

savings will reduce equivalent CO_2 emissions and degradation of the environment.

3.7 Payback of DEC-SAC system

The payback period of the proposed DEC-SAC system is computed for the following parameters:

$T_a = 34.5$ ℃, $RH = 50\%$ (optimized conditions).

Saturation efficiency of cooling pad = 0.65 [22],

Annual operation of air conditioner = 1600 h [25], and.

The unit cost of electrical energy = 0.1$.

The total cost of modification comprises structure and piping cost = $56,

Cost of pump = $5 and the cost of cooling pad = $2,

Approximate cost of running the 12W pump for 1600 h = $2.

The total cost of DEC is approximately $65 assuming the negligible cost of water. The total annual electrical energy savings are 224 kWh. Therefore, the cost of the DEC system could be recovered within 2.9 years.

4 Conclusions

The following findings can be summarized from the above study:

1. The performance of DEC-SAC improves with an increase in evaporative cooling.
2. The performance maximizes with optimized evaporative cooling of 6.7 °C. For the optimum conditions of DEC-SAC, the enhancement in COP corresponding to conventional SAC is found to be 9.37%. The improvements in refrigerating effect and input work are observed as 14 kJ kg^{-1} and (-)140W.
3. Sustainability index increases with the rise in ambient temperature.
4. The cost of proposed system is recoverable in 2.9 years.

From the above results, it can be concluded that the DEC improves the thermodynamic performance of SAC in all weather conditions. The proposed system is most effective when the ambient temperatures are high and relative humidity is low. The proposed system is suitable to implement in most of the hot–humid regions, leaving a few parts of the world where any kind of water is scarce.

Funding No funding was received for conducting this study.

Declarations

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

References

1. Chauhan SS, Rajput SPS (2017) Experimental analysis of an evaporative – vapour compression based combined air conditioning system for required comfort conditions. Appl Therm Eng 115:326–336. https://doi.org/10.1016/j.applthermaleng.2016.12.072
2. Fouda A, Melikyan Z (2011) A simplified model for analysis of heat and mass transfer in a direct evaporative cooler. Appl Therm Eng 31(5):932–936. https://doi.org/10.1016/j.applthermaleng.2010.11.016
3. Dupont JL, Domanski P, Lebrun P, ZF (2019) 38th Informatory Note on Refrigeration Technologies - The Role of Refrigeration in the Global Economy. 2019. https://doi.org/10.18462/iif.NItec38.06.2019.
4. Krarti M, Howarth N (2020) Transitioning to high efficiency air conditioning in Saudi Arabia: A benefit cost analysis for residential buildings. J Build Eng 31:101457. https://doi.org/10.1016/j.jobe.2020.101457
5. Khosravi A, Koury RNN, Machado L (2018) Thermo-economic analysis and sizing of the components of an ejector expansion refrigeration system Analyse thermo-économique et dimensionnement des composants d ' un système frigorifique à détente par éjecteur. Int J Refrig 86:463–479. https://doi.org/10.1016/j.ijrefrig.2017.11.007
6. Dixit M, Arora A, Kaushik SC (2017) Thermodynamic and thermoeconomic analyses of two stage hybrid absorption compression refrigeration system. Appl Therm Eng 113:120–131. https://doi.org/10.1016/j.applthermaleng.2016.10.206
7. Pottker G, Hrnjak P (2015) Effect of the condenser subcooling on the performance of vapor compression systems. Int J Refrig 50(217):156–164. https://doi.org/10.1016/j.ijrefrig.2014.11.003
8. Selbaş R, Kizilkan Ö, Şencan A (2006) Thermoeconomic optimization of subcooled and superheated vapor compression refrigeration cycle. Energy 31(12):2108–2128. https://doi.org/10.1016/j.energy.2005.10.015
9. Klein SA, Reindl DT, Brownell K (2000) Refrigeration system performance using liquid-suction heat exchangers. Int J Refrig 23(8):588–596. https://doi.org/10.1016/S0140-7007(00)00008-6
10. Yau YH, Pean HL (2014) The performance study of a split type air conditioning system in the tropics, as affected by weather. Energy Build 72(2014):1–7. https://doi.org/10.1016/j.enbuild.2013.12.010
11. Yang MH, Yeh RH (2015) Performance and exergy destruction analyses of optimal subcooling for vapor-compression refrigeration systems. Int J Heat Mass Transf 87:1–10. https://doi.org/10.1016/j.ijheatmasstransfer.2015.03.085
12. Yang H, Pei N, Fan M, Liu L, Wang D (2021) Experimental study on an air-cooled air conditioning unit with spray evaporative cooling system Étude expérimentale d ' une unité de conditionnement d ' air à condenseur à air avec système de refroidissement évaporatif par aspersion. Int J Refrig 131:645–656. https://doi.org/10.1016/j.ijrefrig.2021.06.011
13. Yang H, Pei N, Liu L, Fan M, Qin Y (2021) Experimental study on the effect of condensate water on the performance of split air conditioning system. Energy Rep 7:840–851. https://doi.org/10.1016/j.egyr.2021.01.037
14. Al-badri AR, Al-waaly AAY (2017) The influence of chilled water on the performance of direct evaporative cooling. Energy Build 155:143–150. https://doi.org/10.1016/j.enbuild.2017.09.021
15. Wang T, Sheng C, Nnanna AGA (2014) Experimental investigation of air conditioning system using evaporative cooling condenser. Energy Build 81:435–443. https://doi.org/10.1016/j.enbuild.2014.06.047

16. Sarntichartsak P, Thepa S (2013) Modeling and experimental study on the performance of an inverter air conditioner using R-410A with evaporatively cooled condenser. ATE 51(1–2):597–610. https://doi.org/10.1016/j.applthermaleng.2012.08.063
17. Ketwong W, Deethayat T, Kiatsiriroat T (2021) Case Studies in Thermal Engineering Performance enhancement of air conditioner in hot climate by condenser cooling with cool air generated by direct evaporative cooling. Case Stud. Therm. Eng. 26:101127. https://doi.org/10.1016/j.csite.2021.101127
18. Çag A, Ali S (2022) Engineering Science and Technology , an International Journal Performance testing and optimization of a split-type air conditioner with evaporatively-cooled condenser. https://doi.org/10.1016/j.jestch.2021.09.010.
19. Hajidavalloo E, Eghtedari H (2010) Performance improvement of air-cooled refrigeration system by using evaporatively cooled air condenser. Int J Refrig 33(5):982–988. https://doi.org/10.1016/j.ijrefrig.2010.02.001
20. Martínez P, Ruiz J, Cutillas CG, Martínez PJ, Kaiser AS, Lucas M (2016) Experimental study on energy performance of a split air-conditioner by using variable thickness evaporative cooling pads coupled to the condenser. Appl Therm Eng 105:1041–1050. https://doi.org/10.1016/j.applthermaleng.2016.01.067
21. Ankit Sethi EDVB (2010) Genetron Properties 1.4.1. Honeywell International, Inc. - Buffalo Research Laboratory 20 Peabody St, Buffalo, NY, 14210, 2010, Available: https://www.honeywell-refrigerants.com/europe/genetron-refrigerants-modeling-software-download/.
22. Wu JM, Huang X, Zhang H (2009) Theoretical analysis on heat and mass transfer in a direct evaporative cooler. Appl Therm Eng 29(5–6):980–984. https://doi.org/10.1016/j.applthermaleng.2008.05.016
23. Arora A, Arora BB, Pathak BD, Sachdev HL, Arora A (2007) Exergy analysis of Vapour Compression Refrigeration system with R-22, R-407C and R-410A'
24. Marc A, Rosen MK, Dincer I (2008) "ole of exergy in increasing efficiency and sustainability and reducing environmental impact. Energy Policy 36(1):128–137. https://doi.org/10.1016/j.enpol.2007.09.006
25. BEE Govt. of India, "Schedule - 19." Bureau of Energy Efficiency, Govt. of India, New Delhi, India, pp. 1–12, 2015. Available: https://www.beestarlabel.com/content/files/inverter ac schedule final.pdf.
26. Lateef A, Talib M, Assad W, Mechanics T, College-basra ET (2020) Thermal and exergy analysis of optimal performance and refrigerant for an air conditioner split unit under different Iraq climatic conditions. Therm. Sci. Eng. Prog. 19:100595. https://doi.org/10.1016/j.tsep.2020.100595
27. Ramachander J, Gugulothu SK, Sastry GRK, Kumar Panda J, Surya MS (2021) Performance and emission predictions of a CRDI engine powered with diesel fuel: A combined study of injection parameters variation and Box-Behnken response surface methodology based optimization. Fuel 290:120069. https://doi.org/10.1016/j.fuel.2020.120069
28. Khurana H, Majumdar R, Saha SK (2022) Response Surface Methodology-based prediction model for working fluid temperature during stand-alone operation of vertical cylindrical thermal energy storage tank. Renew Energy 188:619–636. https://doi.org/10.1016/j.renene.2022.02.040
29. Borikar SA et al (2021) Case Studies in Thermal Engineering A case study on experimental and statistical analysis of energy consumption of domestic refrigerator. Case Stud Therm Eng 28:101636. https://doi.org/10.1016/j.csite.2021.101636
30. F and CC Ministry of Environment and Government of India, "India Cooling Action Plan (ICAP): Operationalizing Space Cooling Recommendations," 2019. Available: http://ozonecell.nic.in/wp-content/uploads/2021/10/India-Cooling-Action-Plan-Booklet-SP.pdf.

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.

Efficient Design and Implementation of Matrix Multiplication

Shivam Kumar
Department of Electronics and Communication
Delhi Technological University
Delhi, India
shivamkumar2k21vls17@dtu.ac.in

Jeebananda Panda
Department of Electronics and Communication
Delhi Technological University
Delhi, India
jpanda@dce.ac.in

2023 1st International Conference on Circuits, Power and Intelligent Systems (CCPIS) | 979-8-3503-1576-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CCPIS59145.2023.10291365

***Abstract*—The semiconductor industry plays a crucial role in the design and manufacture of integrated circuits (ICs) used in a wide range of electronic devices. Matrix multiplication is of significant importance in various fields and applications as it plays a fundamental role in linear algebra, solving systems of linear equations, data analysis and machine learning, computer graphics and computer vision, network theory, graph algorithms, etc. This paper gives a thorough investigation into how the Wallace tree multiplier, Vedic multiplier, and parallel prefix adders might be combined to enhance matrix multiplication performance. These techniques contribute to achieving significant speed improvements and reduced and optimized resource utilization. Firstly, Ripple Carry adders, Kogge Stone adders, and Han Carlson adders have been designed and analyzed. After that, the Wallace tree multiplier and Vedic multiplier are designed using these adders. By combining both multiplier and adder, matrix multiplication designs, analyses the performance data, and interprets the results obtained from the experiments. Using the ISE Design Suite tools in Verilog, all circuits are created and simulations are run. The XC6SLX150T are the devices used for synthesis.**

***Index Terms*— VLSI (Very Large Scale Integration), DSP (Digital Signal Processing), WTM (Wallace Tree multiplier), VM (Vedic Multiplier), PPA (Parallel Prefix Adder), HCA (Han Carlson Adder), KSA (Kogge Stone Adder), RCA (Ripple Carry Adder)**

INTRODUCTION

Multipliers are essential arithmetic units in the VLSI (Very Large-Scale Integration) industry, particularly in DSP, communication systems, image processing, and various other applications. They are responsible for performing multiplication operations, which are fundamental to numerous computational tasks. The VLSI industry continually works on developing advanced multiplier architectures to enhance speed, power efficiency, and area utilization. Emerging technologies such as approximate computing and approximate multipliers contribute to achieving a balance between accuracy and computational efficiency. Overall, multipliers are indispensable components in the VLSI industry, enabling efficient multiplication operations in various applications and driving ongoing research and development.

The requirement for effective and optimized matrix multiplication algorithms grows as the size and complexity of matrices continue to increase. There is a growing need for effective and optimized methods for matrix operations as the need for high-performance computing keeps rising. In this work, to create a comprehensive framework for high-performance matrix multiplication and multiplication operations in digital circuit designs, investigate the integration of parallel prefix adders alongside the Wallace tree multiplier and Vedic multiplier.

The Wallace tree multiplier and Vedic multiplier are the most effective multipliers. The Wallace tree multiplier[1] reduces the number of partial products and offers faster multiplication speeds by utilizing a combination of carry-save adders and carry-propagate adders. Similarly, the Vedic multiplier[5] leverages simple arithmetic operations, inspired by ancient Indian mathematics principles, to achieve high- speed multiplication with fewer logic gates compared to conventional multipliers. We suggest using parallel prefix adders in the architecture of matrix multiplication and the multiplier. Carry propagation is parallelized via parallel prefix adders, also known as carry-look-ahead adders, which significantly increase speed. Parallel prefix adders reduce the critical route time by efficiently allocating and computing the carries in parallel, which improves overall performance.

For the purpose of creating an optimized framework for matrix multiplication and multiplication operations in digital circuit designs, this thesis investigates the synergistic integration of the Wallace tree multiplier, Vedic multiplier, and parallel prefix adders. All the design and simulation are performed in Verilog using the ISE Design Suite tool. The selected board is XC6SLX150T.

BACKGROUND AND RELATED WORK

The literature contains a variety of research proposals that concentrate on the Wallace Tree Multiplier, Vedic Multiplier, and Parallel Prefix Adder. There are papers present that discuss several Vedic sutras and their application in multiplier design. They discuss the design methodology, implementation details, and optimization techniques used to achieve high-performance multipliers. The performance metrics considered include area utilization, power consumption, and speed.

979-8-3503-1576-9/23/$31.00 ©2023 IEEE
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:05:02 UTC from IEEE Xplore. Restrictions apply.

A. Wallace Tree Multiplier

The idea of "tree reduction," which entails reducing the quantity of partial products that must be merged to produce the finished result, is the foundation for the Wallace tree multiplier. The circuit is therefore faster and more efficient than standard multipliers.

The following steps are required to operate a Wallace tree multiplier:

- Partial products generation
- Reduction of partial products
- Final addition

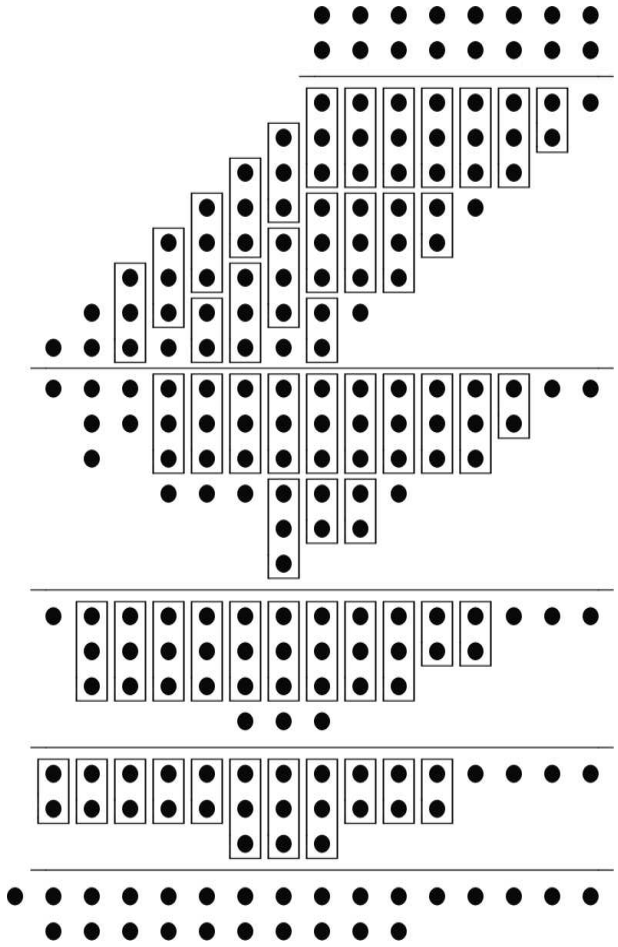


Fig. 1. Graphical representation of 8x8 bits WTM

In order to generate partial products, the Wallace Tree Multiplier multiplies both input values by passing through AND gates. By adding them with half adders and full adders, these partial products are lowered even further Figure.1. It continues to reduce until the final two rows are received. to perform an addition more quickly. Using adder (KSA or HCA or RCA), the final two rows are added. The process flow diagram for creating the Wallace Tree Multiplier is shown in Figure 2.

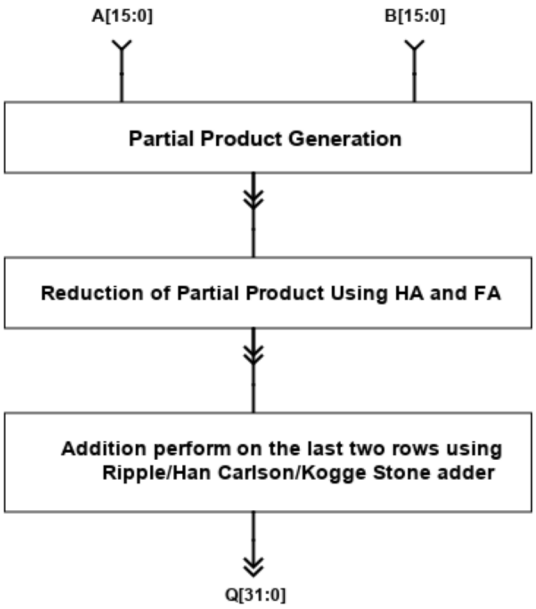


Fig. 2. Process flow diagram of WTM

B. Vedic Multiplier

The Vedic multiplier is a sort of multiplier derived from the Vedic mathematics of ancient India. With this approach, the terms are multiplied both vertically and crosswise, and the results are then added according to Figure 3.

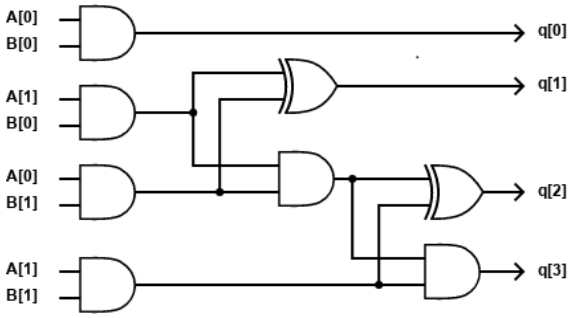


Fig. 3. 2-bit gate level netlist of Vedic Multiplier

The basic steps of the Vedic multiplier algorithm for multiplication are as follows:

- Depending on place values, divide the multiplicand and multiplier into their appropriate halves.
- Perform cross-multiplication between the corresponding parts of the multiplicand and the multiplier. In order to do this, numbers must be multiplied both vertically and diagonally, and the partial products must be arranged in a particular way.
- Vertical additions are used to add up the cross-multiplication step's partial products.
- Any necessary carries from each vertical addition are carried forward.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:05:02 UTC from IEEE Xplore. Restrictions apply.

- The final output is created by combining the results.

In order to construct an N-bit Vedic Multiplier, the input values A and B are split into two equal parts. These parts are referred to as AH, AL, BH, and BL, representing the Most Significant Bits (MSB) and Least Significant Bits (LSB) respectively. This division allows the creation of an N x N size Vedic multiplier, as depicted in Figure 4. are utilized as inputs for four Vedic Multipliers, each of size N/2. These outputs are then fed into four additional Vedic Multipliers, this time with a size of N/4. The N/4 size Vedic Multipliers are further split into four Vedic Multipliers of size N/8, following a similar pattern. This continues until the splitting reaches a block size of 2x2, as depicted in the block diagram shown in the figure. Finally, the outputs of the four N/2 x N/2 multipliers are inputted into an adder with (N + N/2) bits.

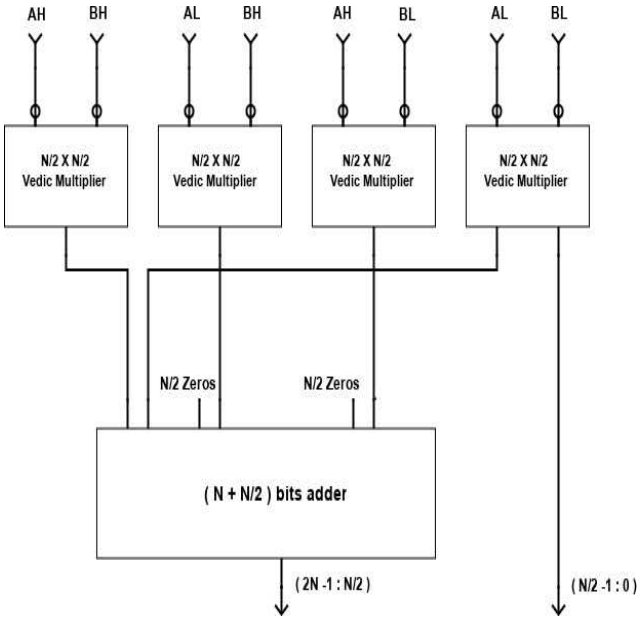


Fig. 4. Block diagram of N bits multiplier

C. Parallel Prefix Adder

Parallel Prefix Adders (PPA) are fast adders that are derived from Carry Look Ahead Adders. Each stage in the parallel prefix adder performs parallel operations on different subsets of the input bits. The final carry-out is then calculated by prefixing the carry-out bits from each stage. The parallel prefix adder works by performing parallel computations within each group to calculate the carries. These computed carries are then passed to the next group, where they are combined with the corresponding inputs to compute the next level of carries. This process continues until the final carry-out bits are obtained.

Parallel Prefix adder can broadly split into three stages. The first stage is the processing stage, in which the propagate and generate signals are calculated using each bit of the pair-bit signal.

$$Pi = Ai \text{ xor } Bi$$

$$Gi = Ai \text{ and } Bi$$

The second step generates carries by utilizing each bit individually and in parallel. This is utilized for carry generation and carries propagation in the intermediate stage logically and is referred to as the carry generation stage.

$$CPi:j = Pi:k+1 \text{ and } Pk:j$$

$$CGi:j = Gi:k+1 \text{ or } (Pi:k+1 \text{ and } Gk:j)$$

In last step, sum and carry are calculated and this step is known as post processing stage.

$$Si = Pi \text{ xor } Ci-1$$

$$Ci-1 = (Pi \text{ and } Ci) \text{ or } Gi$$

The two main parallel prefix adders are explored in this paper. Kogge Stone adder and Han Carlson adder are most efficient and faster adders in parallel prefix adder.

I. Kogge Stone Adder

The Kogge-Stone adder is a type of parallel prefix adder that efficiently computes the carry bits in parallel and combines them in a prefix manner to perform fast addition of binary numbers. By parallelizing carry computations and reducing the ripple carry effect, it decreases the critical path delay. It requires extra area when compared to other Parallel adders, but it has lesser fan-outs at each phase as shown in Figure 5.

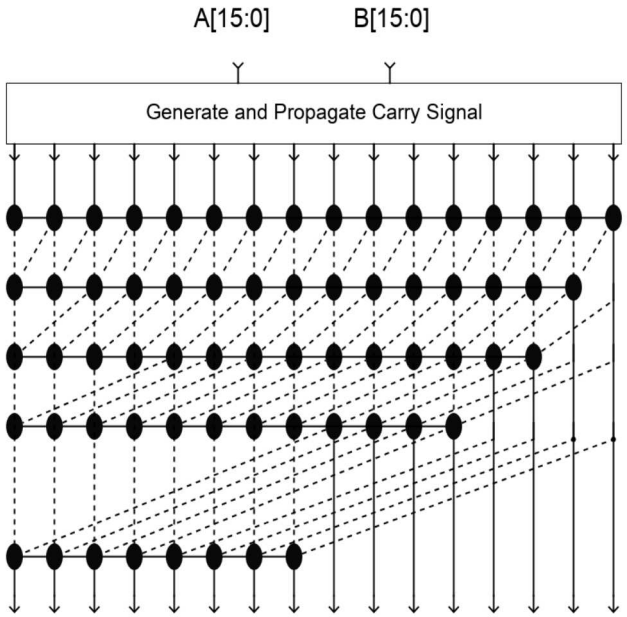


Fig. 5. 16 bit Kogge Stone adder

II. Han Carlson Adder

Han Carlson adder is like Kogge Stone adder. The adder has the benefit of using fewer cells and being shorter. To calculate the odd numbered prefixes, the Han-Carlson adder starts with a Brent-Kung stage, moves through Kogge-Stone stages, and finishes with another Brent-Kung stage as shown in Figure 6. As a result, the complexity is reduced at the cost of an extra stage merging the carry. Due to fewer cells, the speed is fastest in all parallel prefix adder.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:05:02 UTC from IEEE Xplore. Restrictions apply.

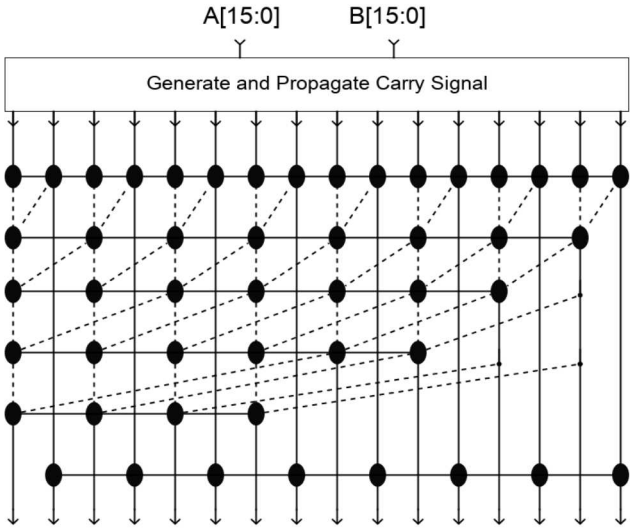


Fig. 6. 16 bit Han Carlson adder

PROPOSED MATRIX MULTIPLICATION MODEL

Matrix multiplication performance can vary greatly depending on the matrix multiplication method used, with effects on things like computational complexity, memory access patterns, parallelizability, and cache utilization. The effectiveness of matrix multiplication can also be impacted by the adder that is selected in a matrix multiplier. The critical path delay of the matrix multiplier, which affects the total latency of the multiplication operation, can be influenced by the adder choice. The critical route length can be shortened with faster adders with less latency, which will speed up multiplication. Lower latency can be achieved with adder designs that have reduced carry propagation delays and optimized logic.

The "left" matrix and the "right" matrix are the two matrices needed to multiply two matrices. The size of the left matrix, denoted as A, is characterized by its dimensions of m x n, where m represents the number of rows and n represents the number of columns. The right matrix, referred to as B, has dimensions of n x p, where p represents the number of columns. The resulting matrix, denoted as C, will have dimensions of m x p, where m represents the number of rows from matrix A and p represents the number of columns from matrix B. The dot product of the i-th row of matrix A and the j-th column of matrix B is used to calculate each element C[i][j] in the resulting matrix.

The i-th row of matrix A and the j-th column of matrix B's dot product are formed by multiplying and adding the respective elements. Element A[i][k] from the i-th row of matrix A is multiplied by element B[k][j] from the j-th column of matrix B in this operation, where k is a number between 1 and n. The last element of matrix C[i][j] is obtained by adding the generated products.

The mathematical formula for the element C[i][j] is:

$$C[i][j] = A[i][1] * B[1][j] + A[i][2]* B[2][j] +... + A[i][n]*B[n][j]$$

In this paper, matrix multiplication is performed using the Wallace Tree Multiplier and Vedic Multiplier. For the addition purpose, Ripple Carry Adder, Kogge Stone Adder, and Han Carlson Adder are used and compared for their performance.

RESULTS

A comparative analysis of 32-bits Ripple Carry Adder, Kogge Stone Adder and Han Carlson Adder are performed. The factors that are considered for comparison are area and delay. Through this detailed analysis, we aim to provide insights into the performance characteristics of different adders and assist in making informed design choices for multiplier implementations.

TABLE I
COMPARITIVE ANALYSIS OF DELAY AND AREA OF ADDERS

Adders	Delay(ns)	Area (LUT)
Ripple Carry Adder	20.8	49
Kogge Stone Adder	14.26	139
Han Carlson Adder	12.101	105

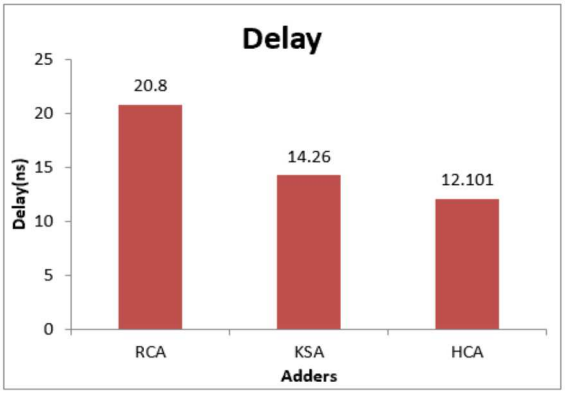


Fig. 7. Delay comparison graph for Adders

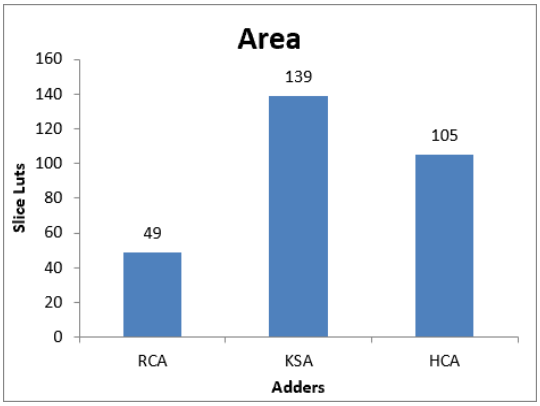


Fig. 8. Area comparison graph for Adders

Table 1 illustrates the performance analysis of the Ripple

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:05:02 UTC from IEEE Xplore. Restrictions apply.

Carry Adder, Kogge Stone Adder and Han Carlson Adder. From table 1, performing addition using Han Carlson adder can reduce delay up to 40%. In terms of area, the ripple carry adder utilized space better than the parallel prefix adder.

The simulation result and device utilization for matrix multiplication using different architectures is attached here. The inputs A and B are split into four 16-bit numbers, which correspond to four numbers of a 2x2 matrix. The first 16 MSB bits i.e. A[63:48] correspond to a[1][1], A[47:32] correspond to a[1][0], A[31:16] correspond to a[0][1] and A[0:15] correspond to a[0][0] of matrix a. In the same way, input B splits.

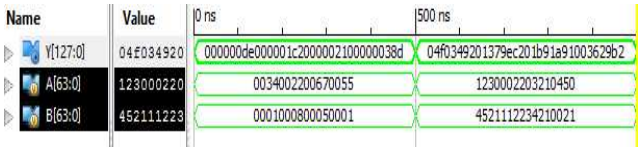


Fig. 9. Simulation of matrix multiplication using VM and RCA

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice LUTs	4864	92152	5%
Number of fully used LUT-FF pairs	0	4864	0%
Number of bonded IOBs	256	396	64%

Fig. 10. Device Utilization of Matrix Multiplication using VM and RCA

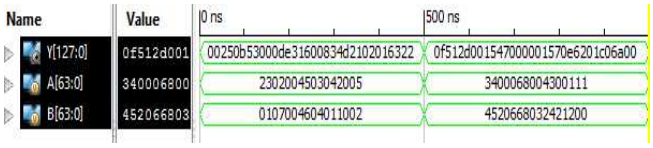


Fig. 11. Simulation of matrix multiplication using VM and KSA

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice LUTs	5523	92152	5%
Number of fully used LUT-FF pairs	0	5523	0%
Number of bonded IOBs	256	396	64%

Fig. 12. Device Utilization of Matrix Multiplication using VM and KSA

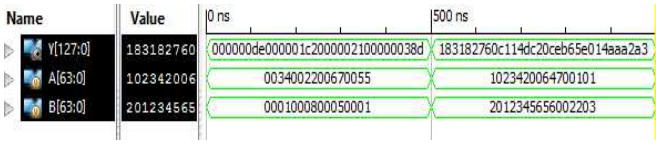


Fig. 13. Simulation of matrix multiplication using VM and HCA

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice LUTs	5100	92152	5%
Number of fully used LUT-FF pairs	0	5100	0%
Number of bonded IOBs	256	396	64%

Fig. 14. Device Utilization of Matrix Multiplication using VM and HCA

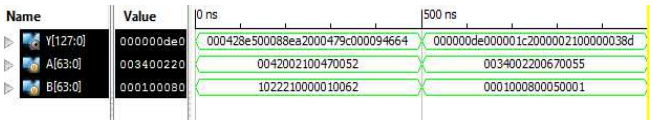


Fig. 15. Simulation of matrix multiplication using WTM and RCA

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice LUTs	4009	92152	4%
Number of fully used LUT-FF pairs	0	4009	0%
Number of bonded IOBs	256	396	64%

Fig. 16. Device Utilization of Matrix Multiplication using WTM and RCA

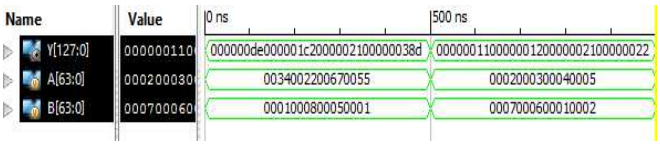


Fig. 17. Simulation of matrix multiplication using WTM and KSA

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice LUTs	4972	92152	5%
Number of fully used LUT-FF pairs	0	4972	0%
Number of bonded IOBs	256	396	64%

Fig. 18. Device Utilization of Matrix Multiplication using WTM and KSA

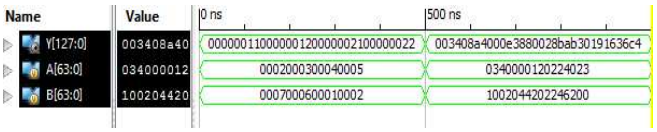


Fig. 19. Simulation of matrix multiplication using WTM and KSA

Device Utilization Summary (estimated values)			[-]
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice LUTs	4528	92152	4%
Number of fully used LUT-FF pairs	0	4528	0%
Number of bonded IOBs	256	396	64%

Fig. 20. Device Utilization of Matrix Multiplication using WTM and KSA

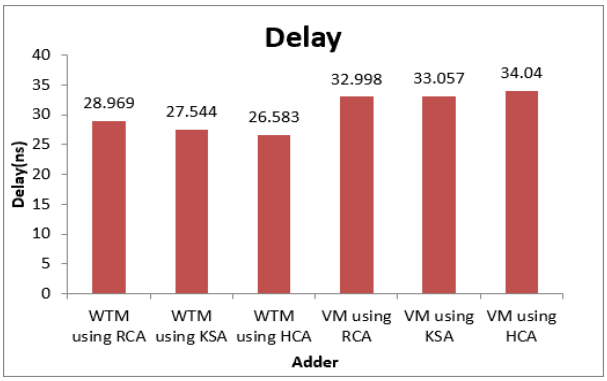


Fig. 21. Delay comparison graph for Matrix Multiplication

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:05:02 UTC from IEEE Xplore. Restrictions apply.

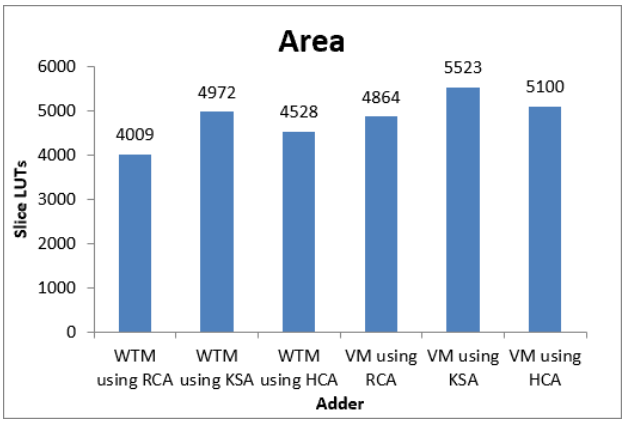


Fig. 22. Area comparison graph for Matrix Multiplication

TABLE II
COMPARITIVE ANALYSIS OF MATRIX MULTIPLICATION USING DIFFERENT MULTIPLIERS

Matrix Multiplication Method	Delay(ns)	Area (LUTs)
Wallace Tree Multiplier and RCA	28.969	4009
Wallace Tree Multiplier and KSA	27.544	4972
Wallace Tree Multiplier andHCA	26.583	4528
Vedic Multiplier and RCA	32.998	4864
Vedic Multiplier and KSA	33.057	5523
Vedic Multiplier and HCA	34.04	5100

Table 2 illustrates the performance analysis of the matrix multiplication using different multiplier and adders. Matrix multiplication using Wallace Tree Multiplier and Han Carlson Adder shows to have the least delay as compared to other methods. The matrix multiplication using Wallace Tree Multiplier and Ripple Carry Adders are shown to have the least area in terms of LUTs.

CONCLUSION AND FUTURE SCOPE

This paper has presented a comprehensive investigation into the optimization of matrix multiplication through the utilization of the Wallace tree multiplier, Vedic multiplier, and parallel prefix adders. This efficiency in improving the performance of matrix multiplication operations has been revealed by the performance evaluation and analysis, which has been quite helpful.

The best superior 32 bits adder is first designed. The Han Carlson adder provides higher performance in terms of speed, according to examination and comparison with the Kogge-Stone adder and ripple carry adder. The Han Carlson adder combines the advantages of the Han-Carlson parallel prefix structure and the carry-save adder, resulting in a reduction of critical path delay and efficient carry propagation. The Han Carlson adder reduces the delay by around 40% compared to the Ripple Carry adder. In terms of area, due to their added circuitry for parallel processing and carry propagation, the HAN-Carlson adder and Kogge-Stone adder often demand more hardware resources than the ripple carry adder. The performance of 16x16 multipliers is then designed and compared after that. Vedic and Wallace tree multipliers are successfully designed with Han Carlson, Kogge Stone, and Ripple Carry adders, and their performance is compared. According to the results, the Wallace tree multiplier and Han-Carlson adder performed faster than the other configuration. Faster multiplication operations were made possible by the Wallace tree multiplier's parallel architecture, and overall performance was improved because the Han- Carlson adder considerably decreased the propagation delay of carry bits.

Matrix Multiplication is implemented using the designed multipliers and adders. By leveraging the strengths of the Wallace tree multiplier, Vedic multiplier, and parallel prefix adders, the integrated framework offers a powerful approach for efficient matrix multiplication, paving the way for advancements in various computational domains. The performance evaluation demonstrates that the combination of the Wallace tree multiplier and the Han-Carlson adder achieves the fastest matrix multiplication results compared to other techniques.

The future scope of this paper includes several prospective directions for additional study and advancement. To further optimize the calculation process and lower the computational complexity, future studies can go deeper into algorithmic improvements. Research in the future can concentrate on selecting the multiplier and optimizing the optimization approaches for specific applications.

REFERENCES

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

[2] T. Gupta and J. B. Sharma, “Han–Carlson adder based high-speed Vedic multiplier for complex multiplication,” Microsystem Technologies, vol. 24, no. 9, 2018,pp. 3901–3906.

[3] A. Gupta and K. Suneja, "Hardware Design of Approximate Matrix Multiplier based on FPGA in Verilog," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 496-498.

[4] A. Raju and S. K. Sa, ”Design and performance analysis of multipliers using Kogge Stone Adder,” 2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Tumkur, India, 2017, pp. 94-99.

[5] P. Gulati, H. Yadav and M. K. Taleja, ”Implementation of an efficient multiplier using the vedic multiplication algorithm,” 2016 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2016, pp. 1440-1443.

[6] S. Lad and V. S. Bendre, ”Design and Comparison of Multiplier using Vedic Sutras,” 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019, pp. 1-5.

[7] M. Kivi Sona and V. Somasundaram, “Vedic multiplier implementation in VLSI,” Mater. Today, vol. 24,2020, pp. 2219–2230.

[8] P. Mehta and D. Gawali, "Conventional versus Vedic Mathematical Method for Hardware Implementation of a Multiplier," 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, Bangalore, India, 2009, pp. 640-642, doi: 10.1109/ACT.2009.162.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 07,2023 at 07:05:02 UTC from IEEE Xplore. Restrictions apply.

Accepted Article

Review

Emerging Trends in Microfluidics Based Devices

Shipra Solanki[1,2]

Chandra M Pandey[2]

Rajinder K Gupta[2]

Bansi D Malhotra[1]

[1]Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India

[2]Department of Applied Chemistry, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India
Correspondence: Prof. Bansi D Malhotra,
Department of Biotechnology,
Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India
E-mail: bansi.malhotra@gmail.com

Additional Correspondence: Dr. Chandra M Pandey,
Department of Applied Chemistry,
Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India
E-mail: cmp.npl@gmail.com

Keywords: Biosensor, Field-deployable systems, Lab-on-a-disc, Microfluidics, Organ-on-a-Chip, Point-of-care diagnostic, Soil-on-a-chip

Abbreviations: **AFP**, Alpha-fetoprotein; **AS**, Aptasensor; **BOC**, Body-on-a-Chip; **CEA**, Carcinoembryonic antigen; **FDS**, Field-deployable systems; **GO,** Graphene oxide; **IA,** Immunoassay; **LAMP**, Loop-mediated isothermal amplification; **LOAD**, Lab-on-a-disc; **MFs**, Microfluidics; **MWCNT,** Multiwall carbon nanotube; **NA**, nucleic acid; **NAD**, Nucleic acid diagnostics; **OOC,** Organ-on-a-chip; **PCR**, Polymerase chain reaction; **PDMS**, Polydimethylsiloxane; **PoCDs,** Point-of-care devices**; PoCT**, Point-of-care testing; **PSA**, Prostate-specific antigen; **RCs**, Root-chips; **SP**, Smartphones; **µPADs,** MFs paper-based analytical devices; **MFD**, Microfluidics based devices; **LOC**, Lab-on-chip; **SCA**, Single-cell

Received: 14/10/2019; Revised: 28/01/2020; Accepted: 31/01/2020

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/biot.201900279.

This article is protected by copyright. All rights reserved.

Accepted Article

analysis; **MFC**, Microfluidic chip; **LOD**, Limit of detection; **CMOS**, Complementary metal oxide semiconductor; **PET**, polyethylene terephthalate; **HDA**, Helicase-dependent amplification; **SDA**, Strand displacement amplification

Abstract

The major challenge for scientists and engineers today is to develop technologies for the improvement of human health in both developed and developing countries. However, the need for cost-effective, high-performance diagnostic techniques is very crucial for providing accessible, affordable, and high-quality healthcare devices. In this context, microfluidic-based devices (MFDs) offer a powerful platform for automation and integration of complex operations onto a single chip. The distinct advantage of MFDs lies in precise control of the sample quantities and flow rate of samples and reagents that enable quantification and detection of analytes with high resolution and sensitivity. With these excellent properties, Microfluidics (MFs) has been used for various applications in healthcare, along with other biological and medical areas. This review focuses on the emerging demands of MFs in different fields, i.e., biomedical diagnostics, environmental analysis, food and agriculture research, etc. in the last about three years. It also aims to reveal new opportunities in these areas and future prospects of commercial MFDs.

1. Introduction

There is a huge demand for emerging technologies aimed at next-generation microfluidics-based devices (MFDs) for different applications [1]. The vision of sample-in and answer-out platforms has encouraged the researchers to work extensively in this direction, which resulted in enormous development in the field of MFDs. Recent advances in emerging technologies such as paper-based assays [2], smart-phone (SP) based technologies, continuous environment monitoring devices and on-site monitoring systems are leading to rapid and miniaturized analysis systems [3]. Currently, SP based technologies have emerged as the futuristic smart devices for affordable personalized health care and are highly promising interfaces for other areas like food quality monitoring etc [4, 5]. The development of low cost and disposable paper- based MFDs, integrated with novel assay formats, have reduced the time and complexity of diagnostic tests [6-8].

MFDs have evolved in many technological aspects whether in the form of fabrication techniques (from photographic technologies to soft lithography) [9, 10] or chip materials used (silicon to polymers and finally to paper-based devices) or integration of operations (from

This article is protected by copyright. All rights reserved.

Accepted Article

simple chemical assays to complex molecular reactions mimicking complex living systems) [11]. The different materials and microfabrication techniques associated with the fabrication of MFDs, along with the fundamental principles and electrokinetic phenomenon, separations, and detection, have already been discussed in several reviews [12-15]. MFs has thus established itself as a standalone discipline of science that finds application in the development of many other disciplines such as biomedical diagnostics, optofluidic technologies, chromatography, etc [16-18]. These MFDs provide improved performance, speed and throughput, reduced reagents/sample consumption and can be used for parallel and integrated analysis, which may probably satisfy the market demand for high-throughput and large-scale applications [19]. Point-of-care testing (PoCT), with ideal characteristics like robustness, quick response, user-friendly, sensitivity, and specificity, is the largest area of application for MFDs with the maximum number of publications [20]. MFs have been considered to comprise of four parents: molecular analysis, biodefence, molecular biology, and microelectronics. However, the modular nature of this technology to fabricate point-of-care devices (PoCDs) for many more areas of biotechnology, medical diagnostics, analytical chemistry, environmental monitoring, and agricultural improvement are yet to be explored [21, 22].

This review emphasizes the emerging trends in the field of MFDs. We focus on a broad range of MFs applications, including molecular diagnostics, PoCT, environmental analysis, clinical and forensic analysis that have been developed in the last about three years **(Fig. 1)**. Besides this, efforts have been made to provide examples of the recent developments on the organ-on-a-chip for clinical or biomedical science applications.

2. Biomedical Diagnostics

Biomedical diagnostics is considered to be the major application area of MFs technologies. These devices are being used for manipulating and transporting a small quantity of fluids and biological samples through micro-channels. Further, the integration of MFs with different sensors or sample-pre-treatment modules helps to reduce cross-contamination, thus making it an appropriate candidate for designing the desired PoCDs. These features provide faster diagnosis, bring down the cost, and enable the use of PoCDs by common people. Different biological samples like urine, blood, saliva, sweat, stool, cerebral, and amniotic fluids can be used for measurements by PoCDs. These PoCDs are mostly used for single disease detection and the currently fabricated PoCDs focus on screening multiple diseases simultaneously [23]. Broadly, biomedical diagnostics can be classified into three categories:

This article is protected by copyright. All rights reserved.

Accepted Article

(1) nucleic acid (NA)-based detection that utilizes sequence-specific NA hybridization; (2) immunoassays (IAs), based on specific immuno-complex formation, and (3) cell-based methods that involve separation and sorting, labeling and target amplification of cell via cell culturing [24].

2.1. MFs Based Nucleic Acid Detection

Nucleic acid diagnostics (NAD) relies on the specific gene sequence detection of the target pathogen. The traditional techniques such as DNA sequencing, hybridization, polymerase chain reaction (PCR) have certain limitations as they have high operational costs, require lengthy analytical procedures, sophisticated instruments along with trained human resources, and hence are complicated [25, 26]. With recent developments in MFs technology and new amplification methods, the PoCT for NAD has dramatically improved human health and resulted in the prevention of the outbreaks of various infectious diseases. The main objective of MFs implementations for NAD is to develop an easy-to-use PoCDs, and this can be achieved by employing four basic steps, i.e. (i) Development of MFs platform, (ii) Lysis of the target cells, (iii) Purification and amplification of the target DNA, and (iv) Detection of target analyte using different transducers [27].

MFs has been widely explored in PCR based methods for NAD [28] **(Fig. 2 (A))**. Fully integrated lab-on-a-disc (LOAD) consisting of magnetic bead–mediated DNA extraction, multiplex PCR pre-amplification, and multiplex species-specific real-time PCR have been reported [29]. The fabricated LOAD device was used to identify multiple pathogens using a single device. However, the primary concern with the developed system is the reproducibility and the complicated DNA extraction processes. In another work, a total integrated genetic analysis microsystem was developed, which could perform pre-treatment of the samples, the isothermal amplification of DNA, and detection using colorimetric technique [30]. Loo et al. developed an integrated LOAD platform that has loop-mediated isothermal amplification (LAMP) reaction and can be used for real-time fluorescence detection [31]. However, this platform is unable to detect multiple pathogens in one run. Other LOAD based platforms reported till date and have been summarized in **Table ST1** [32-34].

Most of the above-discussed LOAD platforms use dedicated chips and cartridges that are used for sample-to-answer detection of pathogens. However, the primary concern with the LOAD is the requirement of an external fluidic pump that not only increases the cost but also

This article is protected by copyright. All rights reserved.

Accepted Article

makes the device bulkier [35, 36]. Moreover, the major challenge for these devices is the incorporation of convective PCR (cPCR) amplification and amplicons detection into a single portable PoCDs. Now-a-days, SP based PoCDs are also being explored as they have a simple optical attachment (camera) that can be used as a transducer for NAD [37, 38]. This SP based flexible technology has been used to develop a cPCR amplification system using consumer-class quadcopter drones technology "Lab-on-a-drone" for rapid field deployment of NAD [39, 40]. Simple, cost-effective, and reusable platform based on composite polydimethylsiloxane (PDMS)-on-glass microchannel functionalized with peptide-NA probes have been used to construct MFs streaming potential analyzer for NAD [41]. In some other studies, the cPCR and multiplex fluorescence-based detection using SP has been utilized for the fabrication of simple and portable PoCDs for NAD [42, 43]. These reported battery-operated devices can amplify the target DNA in less than 30 min and are user-friendly and portable [44]. PoCDs based on synthetic NA (Aptasensor), has been recently reported **(Fig. 2(B))** [45], which has been discussed in the supplementary information.

Despite many advantages, PoCT for NAD is currently limited to centralized laboratories since the NA testing requires elaborated sample processing steps starting from the release, isolation, and concentration of NA. Moreover, cell-free plasma is required for blood sample analysis, which is generally separated from the whole blood using centrifugation, adding an extra step to NAD and thus burdening the POC application. This drawback may perhaps be overcome by developing an automated MFs platform that gives a complete 'sample to answer' assay. Further, for the commercialization of a versatile and cost-effective PoCDs for NAD, efforts should be made towards electrochemical detection with multiple sample analyses, using microarrays and manipulation of the samples using MFs systems [41].

2. MFs in Cell Diagnostics

For the analysis of single-cell, several challenges like size (micrometers) of the cell, the presence of desired macromolecules in low quantity, etc. need to be addressed. The tricky step in the single-cell analysis (SCA) is the development of a reproducible method for the smooth transfer of desired cells to specific locations for analysis. In this context, MFs technique plays an important role in SCA with the capacity to integrate various experimental procedures into a single micro-device [46]. MFs based SCA has been widely used in drug discovery [47], genetic and protein analysis [48, 49], intracellular signaling [50] multi-drug

This article is protected by copyright. All rights reserved.

Accepted Article

resistance [51], etc. The most popular application of MFs based SCA is in cancer research ranging from the studies of cellular interaction to cancer diagnosis and treatment [52-54]. Recently circulating tumor cell identification and capturing have attracted much interest in cancer research, especially in designing novel diagnostics techniques for the treatment of cancer patients [55] **(Fig. 2 (C))**.

One of the most promising therapeutics is regenerative medicine using stem cell therapy, and to better understand stem cell differentiation MFs tools that can be extremely useful which include systems that can discriminate various biochemical signals in controlled environments [56]. To study a microenvironment, MFDs can either be used for screening a varied range of conditions for high throughput or to reconstruct the physiological environment [57]. Several MF methods based on adhesion, surface receptor expression, size, and electrical characteristics have been reported to separate stem cells [58]. However, MF techniques for stem cell sorting are at an early stage because of the low throughput manufacturing processes.

Neurology is another field that is greatly influenced by MFs applications. MFs neural interfaces have immense potential in clinical medicine and basic neuroscience research [59]. Moreover, MFs techniques and principles have been applied to isolated brain tissue culture and studies [59]. Recently, the MFs neural probes have led to the emergence of wireless flexible platforms to allow monitoring of natural movement and behavior while simultaneously minimizing chronic immune-reactive responses in the neural tissue [60, 61]. However, increased efforts are necessary to obtain improved and ready use of these devices in neurology.

2.3. MFs for Immunodiagnostics

MFs immunosensing is a nascent field that has an immense future in clinical and biological diagnostics. The integration of immunoassays (IAs) within MFs can be a smart answer to overcome the restrictions of the conventional IAs. The remarkable features of MFs can serve as an immune-reactor chamber, thereby decreasing the analysis time with minimal sample utilization. Most of the MFs based systems consist of micrometer size channels, which increase the surface-to-volume ratio, thus accelerating the mass transport and immuno-complex formation. Recently, some MFs based immuno-sensors were fabricated for electrochemical detection of mycotoxins [62, 63] and cardiac biomarkers [64, 65] **(Fig. 2 (D))**. However, these fabricated MFs platforms are not efficient as compared to the conventional sophisticated pathology tests as they lack portability and the higher cost for the

This article is protected by copyright. All rights reserved.

Accepted Article

fabrication. To overcome this problem, a better understanding of the IAs associated with portable MFs tools is necessary for the commercialization of the immunosensor based MFs platform. In this context, μPADs for IAs application have desired potential where the paper is used as a substrate, which is affordable, disposable, and user-friendly as compared to conventional substrates [66, 67]. With the introduction of electronics, the fabrication of μPAD with integrated electrodes is increasing [68]. However, there are challenges like miniaturization, integration, and cost that have hampered the commercialization of IAs [66]. We believe that the electrochemical immune μPADs will be commercially available in the near future once these problems have been addressed.

The current MFs technology for single or multiple detection of clinically important antigens is very inspiring as the detection limit has reached to attomole concentration. Further, a deep understanding of the antibody immobilization, fluid control, biological matrix interference, and signal detection modes is necessary for the miniaturization of IAs. Fluorescence, because of its higher signal sensitivity, has been the most common signal transduction platform for monitoring the immune-complex formation. However, other transduction techniques such as electrochemical can perhaps be utilized for simultaneous multiple analyte detection. Non-invasive determination of metabolites in body fluids containing important disease markers must also be considered for IAs based PoCDs.

3. Organ-on-a-Chip

MFDs can simulate the organs of the human body as it consists of cell culture chambers lined with living cells that mimic the physiological conditions in the human body. Organ-on-a-chip (OOC) systems combine MF technologies with tissue engineering to replicate tissue- and organ-level physiology. These devices mimic the organ functionality by recapitulating the physicochemical microenvironments, multicellular architectures, vascular perfusion of the body, and tissue-tissue interfaces [69].

Compared to the conventional *in vitro* models, OOC has the potential to mimic organs, substitute animal models, and lower the time and cost of experiments. These systems have been utilized to different chemical and physical stimuli, and efforts have been made to develop models of muscle, liver, bones, brain, heart, blood vessels, gut, kidney, etc. using the MFs technique [70]. Recently, researchers have micro-fabricated ‘Body-on-a-Chip’ (BOC) devices that consist of different types of cultured cells grown on multiple micro-chambers connected by a network of microchannels that allow the re-circulation and exchange in a

This article is protected by copyright. All rights reserved.

Accepted Article

physiologically-relevant method. The BOC has the potential to act as a relevant model for toxicity analysis and efficacy; and give a better insight into the tissue and organ level metabolic activities [71]. Compared to OOC, the BOC system can simulate the interactions and communications of different organs.

To study the tissue metabolism, cytotoxicity, and cell activity, various BOC systems containing multiple organ models (lung-liver model, gut-liver model, 13 organ model, etc.) have been reported till date [72-76]. In addition, these platforms have been used to investigate the migration of cancer cells in a multi-organ cell environment [77]. BOC systems have shown great potential in studying organ physiology, tissue development, and disease etiology. However, the major obstacle for these BOC systems is the medium circulation that helps in the transfer of nutrient and signaling factor exchange.

MFs based OOC technologies are promising and can be used for the study of tissue development, organ physiology, and disease etiology. Many OOC systems have been reported, that can precisely mimic the organ-level functions, but the major concern is the cell diversity. The PDMS used in the MFs OOC systems also hinders the function since it absorbs the chemical compounds generally used in drug-related research. Other materials like polyesters, natural hydrogels (matrigel, collagen) may be used to construct the devices but the high cost is currently an issue [78]. These limitations make it quite difficult to model all aspects of an organ-on- a- single chip. Overall, it appears that utilizing MFs to OOC systems is not a solution, but only an approach for creating a controlled environment .Hence, other technologies are required to enhance the OOC performance and simulations.

4. MFs in Food and Agriculture

In the food and agriculture sector, MF technology has paved new ways of different analyses that resulted in the improvement of crop production, food safety, as well as food processing. MF technology can be envisaged as the major economic driving force benefiting farmers as well as consumers with its new innovative solutions to the problems like crop protection and improvement, nutrient monitoring and enrichment [79, 80]. MFs technology has been used for precision agriculture by revolutionizing genotyping, phenotyping and nutrient monitoring of plants through integration of different steps like sample preparation, amplification and detection into one single unit using various methodologies are summarized in **Figure 3A**.

This article is protected by copyright. All rights reserved.

Accepted Article

4.1. MFs in Food Production

The use of MFs chips for high throughput genotyping analysis of plants is a well-known technique [81, 82]. It has revolutionized the field of NA analysis by providing a large number of plant species genotype data in less time and cost-effective manner, which has helped in designing genetically modified crops. In addition to genotyping MFs chips, the phenotypic analysis has recently shifted from traditional culturing of seeds in pots or petrie dishes or greenhouses to MFDs [83]. Several MFs chips have been used to study the dynamic change of plant traits with time under different nutrient stress conditions, fertilizer concentrations, pH, temperature, and other chemicals [84-87]. Recently, several MF systems named root-chips (RCs) have been reported to study root adaptations under different nutrient concentrations to determine the essential nutrients for increasing crop production [88]. *Arabidopsis thaliana* is the most studied plant in MF plant chips because of its small size with a single genome and can be easily grown. Massahla et al. studied the microbial interactions of *Bacillus subtilis* with the roots of *A. thaliana* in an MF chip with real-time imaging using a microscope [87]. Chai et al. studied the root response of *Oryza sativa* plant to different drought conditions induced by polyethylene glycol using an MF chip named as petaloid root growth microfluidics chip (PGRM). The plant seedlings were grown in the fabricated chip and the nutrient supply was controlled through the chip (**Fig. 3B)** [88]. The use of the MFs based system instead of open culture analysis may allow nondestructive observations at a very high spatial and temporal resolution.

In addition to the utilization of RCs, the plant nutrients have been monitored via MF systems that were designed for soil nutrient analysis. Several MFs chemical sensors for quantification of nitrates, ammonia, phosphates, and other ions are being used. Recently the integration of separating and quantifying units into one system have made the on-site and continuous monitoring of soil possible [79, 89]. Such type of MFDs consists of either a small separating chamber or a simple filtration membrane and an extraction liquid that extracts the target chemicals and then these are introduced to the MF chip for analysis [90]. Soil microbial ecology has been studied by specially designed “Soil-on-a-Chip” devices, which provides new insights into the soil enrichment by manipulating and simulating the physical and chemical conditions at the microscale in soil model habitats [91, 92]. Rapid and portable on-site detection of MFs multiplexed assays for several plant pathogens have helped in early and better disease control and prevention [93, 94].

The MFs chips for genotyping analysis have been commercialized, but for phenotypic analysis, the MFDs are under-utilized till date. A robust experimental framework should be

This article is protected by copyright. All rights reserved.

Accepted Article

established for high-throughput and precise plant phenotyping based on MF plant chip technology, along with the choice of embedded sensors. MFs assisted the synthesis of nano-encapsulated pesticides; antibiotics, as well as fertilizers, represent some of the areas that will create a future paradigm shift in the field of crop improvement and production.

4.2. MFs in Food Safety

MFDs have blazed new ways for the rapid and efficient detection of pathogens, allergens, toxins, pesticide residues, heavy metals, additives, and other chemical and physical food contaminants. The advent and integration of new detection technologies like SERS and surface wave acoustics are challenging the old age micro array-based gold standard detection methods that have made possible the detection of multiple analytes on a single chip [95, 96]. Mungroo et al. developed a multi-pathogen detection system based on SERS that can be utilized to distinguish between eight food-borne pathogens using silver nanoparticles for signal enhancement [95]. The shift of detection platforms from chip-based assays to μPADs has made it easier to operate, cost-effective, rapid and disposable detection devices rendering food safety analysis of reaching the common man possible. Further, the integration of SP with different portable sensing devices has provided globally accepted platforms for end-users to perform different complicated analysis of utility with ease and simplicity [3, 97]. Some SP based MFDs reported for food safety analysis have been summarized in **Table 1**.

Though the use of MFDs in food and water analysis is a burgeoning field, still, many issues remain to be answered. The major challenge with μPADs is the handling and pre-treatment of the samples prior to detection. Water and beverages sample, for example, does not need any pretreatment before analysis, but solid food samples require a particular pretreatment for target analyte sample collection. Despite the current limitations of sample pretreatment methods, selectivity, and sensitivity, the use of μPADs have led to improved performance of target detection in real food and water samples.

4.3. MFs in Food Processing Industry

In food processing industries, the major application of MFs is in the preparation of different nano-emulsions, nano-encapsulations and nano-suspensions. In milk and dairy industries where solids and liquids are mixed for several reasons, MFs can address the challenges of mixing due to having capabilities to modify the fluid viscosity, fluid density and rheological properties of the fluids [98]. MFs based chips are expected to have high solvent extraction abilities due to laminar flow, shorter diffusion distance and the relatively

This article is protected by copyright. All rights reserved.

Accepted Article

large interfacial area between organic and inorganic phases. This helps in the extraction of bioactive materials of interest from plants [99]. Moreover, MFs can be used to design encapsulation systems for oral administration of the functional and bioactive compounds for their controlled release and protection against the adverse conditions of the gastrointestinal tract. The most convenient method for food compound encapsulation is the formation of micro-beads by polymerizing the droplets. MF emulsification has been widely investigated on the lab scale; however, the commercialization continues to be a challenge due to some issues related to the scale-up for more significant throughput [100].

For the development of simple and highly ordered multiple food structures, MFDs has shown great potential. So far, these systems are limited to the lab-scale because working together with multiple systems results in numerous operational problems (instability, emulsification, etc), including the complexity of food. Thus, food scientists and technologists should be aware of the developments being made in other areas to expand the knowledge gained for designing devices for scale-up synthesis.

5. MFs for Environmental Monitoring

The field of environmental monitoring is another area where extensive research has been conducted on the use of MFs technology. MFDs have been designed for the determination of biological as well as chemical species of interest in air, water, and soil. The environmental sample analysis has its own challenges. These samples are very difficult to handle as they are large in number as well as *in situ* monitoring is very important and secondly, the sample pretreatment is a major challenge as the analyte is present in trace amount at the ppm-ppt level. Thus, either the sensitivity of the device should be too high, which is very difficult to achieve or a sample pre-concentration step is followed prior to detection. For sample pre-concentration, two types of approaches are developed during on-the-chip analysis of environmental samples: one is to pre-concentrate the sample first and then use the sensing chip for the analyte detection, and the other is to perform the pre-concentration step on an MF chip. On-chip pre-concentration methods are preferred to design an integrated system. In the next section, we discuss different MFDs designed for the analysis of contaminants in water, soil, and air.

5.1. Water Quality Assessment

The water quality monitoring devices can be designed according to the two types of requirements, one is for testing water for human usage and the other is to monitor the health

This article is protected by copyright. All rights reserved.

Accepted Article

of water bodies. The MF devices for testing of water usage should be sensitive, selective, rapid, reliable, portable, affordable, and easy to use. Recently introduced μPADs based technology, developed initially for PoCDs, has been extended for water assessment as these μPADs satisfy the entire criterion mentioned earlier. Now-a-days, with the advancement of technology, the color change can be measured using optical imaging, which has made possible the quantification of the analyte and the integration of SP to the μPADs. This has led to improved sensitivity as well as the portability of the devices [101]. Arduini et al. reported one such origami-based printed paper-based MFD for the detection of pesticides at the ppb level. Different screen printed office papers were used by folding and unfolding the filter paper-based structures for detecting several classes of pesticides by combining different enzyme-inhibition biosensors (**Fig.3C**) [102].

The monitoring of water bodies requires field-deployable systems (FDS) that are robust and function continuously, along with high sensitivity and selectivity. Several MFs based chips have already been used in the commercially available devices to reduce the volumes of analytes, which render them functional for months in adverse conditions. Regardless of the vast number of disposable systems designed, FDS are currently in the limelight. FDS has the advantages of being capable of being used *in situ* and continuously function for a long duration of time without any external assistance [103]. Nightingale et al. fabricated a pen-sized system utilizing a MFs chip operating on the standard Griess method for the continuous analysis of nitrates and nitrites in river water. The design of the device reduced the reagent consumption to 2.8 mL/day with a measurement frequency of per 10s**.** The use of MF sensor reduced the amount of reagent usage so that one liter of the reagents was sufficient for 11 months of continuous functioning [104]. Recently research has been directed towards the use of self-powering devices that are capable of generating power for their functioning themselves and when left in the environment, these devices can operate continuously [105]. Characteristics of some of the FDS reported in the literature for environmental monitoring are listed in **Table ST2**.

The main advantage of FDS is the capability of integrating into a small space all the processes involved in the analysis. However, self-functioning and wireless connectivity remain a problem. These devices are generally semi-autonomous and require an external power source to function. However, there are ample possibilities for designing self-powered FDS with improved wireless connectivity to monitor the remote areas incessantly.

This article is protected by copyright. All rights reserved.

Accepted Article

5.2. Air Quality Assessment

Most of the MFDs describe the handling of liquids, with only a small part dealing with gas-based applications. Recent studies have reported the use of MFs for measurements of interfacial tension to analyze the chemical mimics of atmospheric aerosol for the detection of bio-aerosol hazards [106, 107]. Jiang et al. developed a MFs system that could capture and enrich airborne pathogens (*Staphylococcus aureus*) and detect them using LAMP [108]. The analysis time required from loading the bio-aerosol to LAMP amplification was about 1.5 h with 24 cells per reaction as the detection limit for *S. aureus*. The specificity of the test was performed using five different air-borne bacteria. In another report Liu et al. used LAMP methodology for the detection of both airborne and waterborne pathogens [109]. The major problem, however, is the handling of gaseous samples as all the reported LOC devices contain a unit of extracting species of interest first in liquid form, and then it is analyzed on the chip. Thus, there is enough scope for the development of MFD for air monitoring by integrating different innovative technologies.

5.3. Soil Assessment

Though there are many MFDs to investigate the composition and various constituents of soil according to different perspectives, the MFDs have proved to be a boon in studying the soil ecology. Soil is the most challenging ecosystem on earth, comprising of varying physical, chemical, and biological structures. MFDs have addressed the challenges like enormous spatio-temporal heterogeneity of soil, difficulty in studying microbes and lack of optical access in real soil systems; which were difficult to be approached by any other technology [110, 111].

Handling of gaseous samples is very difficult and all the LOC devices reported contain a unit of extracting species of interest first in liquid form and then analysis on-chip. Still, there is a lot of work required to be done to realize the MFs based air monitoring devices. Further, for the fabrication of a perfect MFD, different steps such as sampling, pretreatment, separation, reaction, detection, and signal processing must be integrated into a single circuit that can function in adverse conditions.

6. Conclusions and Future Perspective

MFs is a rising star in all the fields of diagnostics and analysis as far as the realization of technology from the laboratory to the field is concerned. Since the field of MFs has matured in the last decade, many more applications and integration of complex functions on-

This article is protected by copyright. All rights reserved.

Accepted Article

chip have been demonstrated. Many commercial MFDs pertaining to different areas are manufactured by companies like μFluidix, Agilent Technologies, Caliper life-sciences, Biacore, PerkinElmer, Cepheid etc. Despite rapid advancements in this area, there are several limitations of technology like commercialization, increased signal-to-noise ratio due to miniaturization, a complex external system to work properly as well as human ethics and behavior related to untrained public diagnosing potential.

As compared to other biomedical devices, MFs and microsystems technologies are relatively under-researched for applications in agriculture and plant sciences and environmental monitoring that have enormous social and economic impacts. Examples of techniques that look promising for the future include smart contact lens sensors, real-time monitoring of physiological parameters from tear fluid for non-invasive diagnostics, or tattoo-based sensors. Another promising area of MF application relates to the space science arising due to the mass reduction benefit of the microscale. The LOC device that is capable of estimating blood sugar, kidney function and liver function will keep astronauts safer in space as they have to be diagnosed timely during space travel. Within the next two decades, it is believed that the "lab-on-a-chip revolution" is likely to occur, and the commercialization of the FDs based devices will lead to the expansion of MFDs from molecular diagnostics to other fields.

Acknowledgments

We thank Prof. Yogesh Singh, Vice-Chancellor, Delhi Technological University, for his interest in this work. SS is thankful to CSIR, India, for the award of RA. CMP acknowledges the Department of Science and Technology, New Delhi, India, for the DST-INSPIRE Faculty Award (DST/INSPIRE/04/2015/000932). B.D.M. thanks the Science & Engineering Research Board (Government of India) for the award of a Distinguished Fellowship (SB/DF/011/2019).

Conflict of interest

The authors declare no financial or commercial conflict of interest.

References

[1] Whitesides, G. M., *Nature* **2006**, *442*, 368.
[2] Gao, B., Li, X., Yang, Y., Chu, J., He, B., *Analyst* **2019**, *144*, 6497.
[3] Granica, M., Tymecki, Ł., *Talanta* **2019**, *197*, 319.
[4] Yager, P., Edwards, T., Fu, E., Helton, K., Nelson, K., Tam, M. R., Weigl, B. H., *Nature* **2006**, *442*, 412.
[5] Cui, P., Wang, S., *Journal of Pharmaceutical Analysis* **2019**, *9*, 238.

This article is protected by copyright. All rights reserved.

Accepted Article

[6] Hatamie, A., Angizi, S., Kumar, S., Pandey, C. M., Simchi, A., Willander, M., Malhotra, B. D., *Journal of The Electrochemical Society* **2020**, *167*, 037546.
[7] Salim, A., Lim, S., *Sensors* **2018**, *18*, 232.
[8] Vashist, S. K., Luppa, P. B., Yeo, L. Y., Ozcan, A., Luong, J. H. T., *Trends in Biotechnology* **2015**, *33*, 692.
[9] Qin, D., Xia, Y., Whitesides, G. M., *Nature protocols* **2010**, *5*, 491.
[10] Loo, J. F., Ho, A. H., Turner, A. P., Mak, W. C., *Trends in biotechnology* **2019**.
[11] Martinez, A. W., Phillips, S. T., Whitesides, G. M., Carrilho, E., ACS Publications **2009**.
[12] Pandey, C. M., Augustine, S., Kumar, S., Kumar, S., Nara, S., Srivastava, S., Malhotra, B. D., *Biotechnology journal* **2018**, *13*, 1700047.
[13] Kumar, S., Kumar, S., Ali, M. A., Anand, P., Agrawal, V. V., John, R., Maji, S., Malhotra, B. D., *Biotechnology journal* **2013**, *8*, 1267.
[14] Dixit, C. K., Kaushik, A. K., Kaushik, A., *Microfluidics for biologists*, Springer 2016.
[15] Squires, T. M., Quake, S. R., *Reviews of modern physics* **2005**, *77*, 977.
[16] Weibel, D. B., Whitesides, G. M., *Current opinion in chemical biology* **2006**, *10*, 584.
[17] Sackmann, E. K., Fulton, A. L., Beebe, D. J., *Nature* **2014**, *507*, 181.
[18] Unger, M. A., Chou, H.-P., Thorsen, T., Scherer, A., Quake, S. R., *Science* **2000**, *288*, 113.
[19] Chin, C. D., Linder, V., Sia, S. K., *Lab on a Chip* **2012**, *12*, 2118.
[20] Yetisen, A. K., Akram, M. S., Lowe, C. R., *Lab on a Chip* **2013**, *13*, 2210.
[21] Lafleur, J. P., Joensson, A., Senkbeil, S., Kutter, J. P., *Biosensors and Bioelectronics* **2016**, *76*, 213.
[22] Tomazelli Coltro, W. K., Cheng, C. M., Carrilho, E., de Jesus, D. P., *Electrophoresis* **2014**, *35*, 2309.
[23] Rivet, C., Lee, H., Hirsch, A., Hamilton, S., Lu, H., *Chemical Engineering Science* **2011**, *66*, 1490.
[24] Mauk, M. G., Liu, C., Sadik, M., Bau, H. H., Microfluidic Devices for Nucleic Acid (NA) Isolation, Isothermal NA Amplification, and Real-Time Detection, in: Rasooly, A., Herold, K. E. (Eds.), *Mobile Health Technologies: Methods and Protocols*, Springer New York, New York, NY 2015, pp. 15.
[25] Enzelberger, M. M., Hansen, C. L., Liu, J., Quake, S. R., Ma, C., Google Patents **2010**.
[26] Pandey, C. M., Dewan, S., Chawla, S., Yadav, B. K., Sumana, G., Malhotra, B. D., *Analytica Chimica Acta* **2016**, *937*, 29.
[27] Mauk, M. G., Song, J., Liu, C., Bau, H. H., *Biosensors (Basel)* **2018**, *8*, 17.
[28] Nguyen, H. V., Nguyen, V. D., Nguyen, H. Q., Chau, T. H. T., Lee, E. Y., Seo, T. S., *Biosensors and Bioelectronics* **2019**, *141*, 111466.
[29] Czilwik, G., Messinger, T., Strohmeier, O., Wadle, S., von Stetten, F., Paust, N., Roth, G., Zengerle, R., Saarinen, P., Niittymäki, J., McAllister, K., Sheils, O., O'Leary, J., Mark, D., *Lab on a Chip* **2015**, *15*, 3749.
[30] Oh, S. J., Park, B. H., Choi, G., Seo, J. H., Jung, J. H., Choi, J. S., Kim, D. H., Seo, T. S., *Lab on a Chip* **2016**, *16*, 1917.
[31] Loo, J. F.-C., But, G. W.-C., Kwok, H.-C., Lau, P.-M., Kong, S.-K., Ho, H.-P., Shaw, P.-C., *Food Chemistry* **2019**, *274*, 822.
[32] Choi, G., Prince, T., Miao, J., Cui, L., Guan, W., *Biosensors and Bioelectronics* **2018**, *115*, 83.
[33] Liu, Q., Zhang, X., Chen, L., Yao, Y., Ke, S., Zhao, W., Yang, Z., Sui, G., *Sensors and Actuators B: Chemical* **2018**, *270*, 371.
[34] Nguyen, H. V., Nguyen, V. D., Lee, E. Y., Seo, T. S., *Biosensors and Bioelectronics* **2019**, *136*, 132.

This article is protected by copyright. All rights reserved.

Accepted Article

[35] Nair, C. B., Manjula, J., Subramani, P. A., Nagendrappa, P. B., Manoj, M. N., Malpani, S., Pullela, P. K., Subbarao, P. V., Ramamoorthy, S., Ghosh, S. K., *PLOS ONE* **2016**, *11*, e0146961.
[36] Zhang, Y., Jiang, H.-R., *Analytica Chimica Acta* **2016**, *914*, 7.
[37] Mutlu, A. Y., Kılıç, V., Özdemir, G. K., Bayram, A., Horzum, N., Solmaz, M. E., *Analyst* **2017**, *142*, 2434.
[38] Qiu, X., Zhang, S., Mei, L., Wu, D., Guo, Q., Li, K., Ge, S., Ye, X., Xia, N., Mauk, M. G., *Biomicrofluidics* **2017**, *11*, 024103.
[39] Priye, A., Wong, S., Bi, Y., Carpio, M., Chang, J., Coen, M., Cope, D., Harris, J., Johnson, J., Keller, A., Lim, R., Lu, S., Millard, A., Pangelinan, A., Patel, N., Smith, L., Chan, K., Ugaz, V. M., *Analytical Chemistry* **2016**, *88*, 4651.
[40] Priye, A., Ugaz, V. M., Smartphone-Enabled Detection Strategies for Portable PCR–Based Diagnostics, in: Rasooly, A., Prickril, B. (Eds.), *Biosensors and Biodetection: Methods and Protocols Volume 1: Optical-Based Detectors*, Springer New York, New York, NY 2017, pp. 251.
[41] Ghrera, A. S., Pandey, C. M., Malhotra, B. D., *Sensors and Actuators B: Chemical* **2018**, *266*, 329.
[42] Qiu, X., Ge, S., Gao, P., Li, K., Yang, S., Zhang, S., Ye, X., Xia, N., Qian, S., *Microsystem Technologies* **2017**, *23*, 2951.
[43] Mulberry, G., White, K. A., Vaidya, M., Sugaya, K., Kim, B. N., *PLOS ONE* **2017**, *12*, e0179133.
[44] Rajendran, V. K., Bakthavathsalam, P., Bergquist, P. L., Sunna, A., *Biosensors and Bioelectronics* **2019**, *134*, 68.
[45] Mohan, H. K. S. V., Chee, W. K., Li, Y., Nayak, S., Poh, C. L., Thean, A. V. Y., *Materials & Design* **2019**, 108208.
[46] Tavakoli, H., Zhou, W., Ma, L., Perez, S., Ibarra, A., Xu, F., Zhan, S., Li, X., *TrAC Trends in Analytical Chemistry* **2019**, *117*, 13.
[47] Zhai, J., Yi, S., Jia, Y., Mak, P.-I., Martins, R. P., *TrAC Trends in Analytical Chemistry* **2019**, *117*, 231.
[48] Guo, S., Lin, W. N., Hu, Y., Sun, G., Phan, D.-T., Chen, C.-H., *Lab on a Chip* **2018**, *18*, 1914.
[49] Wu, M., Huang, P.-H., Zhang, R., Mao, Z., Chen, C., Kemeny, G., Li, P., Lee, A. V., Gyanchandani, R., Armstrong, A. J., Dao, M., Suresh, S., Huang, T. J., *Small* **2018**, *14*, 1801131.
[50] Chen, Y., Pei, S., Yan, L., Qiu, X., Li, R., Yu, S., Lodhi, A., Li, B., Manzoor, R., Zhang, P., Deng, Y., Ma, H., *Acta Astronautica* **2019**.
[51] Parekh, K., Sharifi, H., Khamenehfar, A., Beischlag, T. V., Payer, R. T. M., Li, P. C. H., The microfluidic capture of single breast cancer cells for multi-drug resistance assays, *Methods in Enzymology*, Academic Press 2019.
[52] Kuboki, Y., Fischer, C. G., Beleva Guthrie, V., Huang, W., Yu, J., Chianchiano, P., Hosoda, W., Zhang, H., Zheng, L., Shao, X., Thompson, E. D., Waters, K., Poling, J., He, J., Weiss, M. J., Wolfgang, C. L., Goggins, M. G., Hruban, R. H., Roberts, N. J., Karchin, R., Wood, L. D., *The Journal of Pathology* **2019**, *247*, 347.
[53] Yin, J., Deng, J., Du, C., Zhang, W., Jiang, X., *TrAC Trends in Analytical Chemistry* **2019**, *117*, 84.
[54] Yu, X., Wu, N., Chen, F., Wei, J., Zhao, Y., *TrAC Trends in Analytical Chemistry* **2019**, *117*, 27.
[55] Leung, C.-H., Wu, K.-J., Li, G., Wu, C., Ko, C.-N., Ma, D.-L., *TrAC Trends in Analytical Chemistry* **2019**, *117*, 78.

This article is protected by copyright. All rights reserved.

Accepted Article

[56] Menachery, A., Kumawat, N., Qasaimeh, M., *TrAC Trends in Analytical Chemistry* **2017**, *89*, 1.
[57] Xavier, M., Oreffo, R. O. C., Morgan, H., *Biotechnology Advances* **2016**, *34*, 908.
[58] Sreepadmanabh, M., Toley, B. J., *Biotechnology Advances* **2018**, *36*, 1094.
[59] Oddo, A., Peng, B., Tong, Z., Wei, Y., Tong, W. Y., Thissen, H., Voelcker, N. H., *Trends in Biotechnology* **2019**.
[60] Chae, U., Shin, H., Cho, I.-J., *IBRO Reports* **2019**, *6*, S264.
[61] Gutruf, P., Rogers, J. A., *Current Opinion in Neurobiology* **2018**, *50*, 42.
[62] Lu, L., Gunasekaran, S., *Talanta* **2019**, *194*, 709.
[63] Mondal, K., Ali, M. A., Srivastava, S., Malhotra, B. D., Sharma, A., *Sensors and Actuators B: Chemical* **2016**, *229*, 82.
[64] Singh, N., Rai, P., Ali, M. A., Kumar, R., Sharma, A., Malhotra, B. D., John, R., *Journal of Materials Chemistry B* **2019**, *7*, 3826.
[65] Singh, N., Ali, M. A., Rai, P., Sharma, A., Malhotra, B. D., John, R., *ACS Applied Materials & Interfaces* **2017**, *9*, 33576.
[66] Zhu, G., Yin, X., Jin, D., Zhang, B., Gu, Y., An, Y., *TrAC Trends in Analytical Chemistry* **2019**, *111*, 100.
[67] Martinez, A. W., Phillips, S. T., Butte, M. J., Whitesides, G. M., *Angewandte Chemie International Edition* **2007**, *46*, 1318.
[68] Alizadeh, N., Salimi, A., Hallaj, R., *Talanta* **2018**, *189*, 100.
[69] Bhatia, S. N., Ingber, D. E., *Nature biotechnology* **2014**, *32*, 760.
[70] Huh, D., Hamilton, G. A., Ingber, D. E., *Trends in cell biology* **2011**, *21*, 745.
[71] Perestrelo, A. R., Águas, A. C., Rainer, A., Forte, G., *Sensors* **2015**, *15*, 31142.
[72] Bovard, D., Sandoz, A., Luettich, K., Frentzel, S., Iskandar, A., Marescotti, D., Trivedi, K., Guedj, E., Dutertre, Q., Peitsch, M. C., Hoeng, J., *Lab on a Chip* **2018**, *18*, 3814.
[73] Chen, H. J., Miller, P., Shuler, M. L., *Lab on a Chip* **2018**, *18*, 2036.
[74] Oleaga, C., Riu, A., Rothemund, S., Lavado, A., McAleer, C. W., Long, C. J., Persaud, K., Narasimhan, N. S., Tran, M., Roles, J., Carmona-Moran, C. A., Sasserath, T., Elbrecht, D. H., Kumanchik, L., Bridges, L. R., Martin, C., Schnepper, M. T., Ekman, G., Jackson, M., Wang, Y. I., Note, R., Langer, J., Teissier, S., Hickman, J. J., *Biomaterials* **2018**, *182*, 176.
[75] Satoh, T., Sugiura, S., Shin, K., Onuki-Nagasaki, R., Ishida, S., Kikuchi, K., Kakiki, M., Kanamori, T., *Lab on a Chip* **2018**, *18*, 115.
[76] Miller, P. G., Shuler, M. L., *Biotechnology and Bioengineering* **2016**, *113*, 2213.
[77] Ma, Y., Pan, J.-Z., Zhao, S.-P., Lou, Q., Zhu, Y., Fang, Q., *Lab on a Chip* **2016**, *16*, 4658.
[78] Tian, C., Tu, Q., Liu, W., Wang, J., *TrAC Trends in Analytical Chemistry* **2019**, *117*, 146.
[79] Antonacci, A., Arduini, F., Moscone, D., Palleschi, G., Scognamiglio, V., *TrAC Trends in Analytical Chemistry* **2018**, *98*, 95.
[80] Felix, F., Baccaro, A., Angnes, L., *Sensors* **2018**, *18*, 4124.
[81] Zec, H. C., Zheng, T., Liu, L., Hsieh, K., Rane, T. D., Pederson, T., Wang, T.-H., *Microsystems & Nanoengineering* **2018**, *4*, 17097.
[82] Ren, P., Liu, J., Zhao, H., Fan, X.-P., Xu, Y.-C., Li, C.-X., *Forensic Science International: Genetics* **2019**, *41*, 145.
[83] França, D., Messa, L. L., Souza, C. F., Faez, R., Nano and Microencapsulated Nutrients for Enhanced Efficiency Fertilizer, *Polymers for Agri-Food Applications*, Springer 2019, pp. 29.
[84] Jiang, H., Wang, X., Aluru, M. R., Dong, L., *Sensors and Actuators A: Physical* **2019**, *298*, 111572.

This article is protected by copyright. All rights reserved.

Accepted Article

[85] Wang, X., Jiang, H., Wang, Y., Dong, L., *2017 IEEE 12th International Conference on Nano/Micro Engineered and Molecular Systems (NEMS)*, IEEE **2017**, pp. 398.
[86] Sytar, O., Zivcak, M., Olsovska, K., Brestic, M., Perspectives in High-Throughput Phenotyping of Qualitative Traits at the Whole-Plant Level, *Eco-friendly Agro-biological Techniques for Enhancing Crop Productivity*, Springer 2018, pp. 213.
[87] Massalha, H., Korenblum, E., Malitsky, S., Shapiro, O. H., Aharoni, A., *Proceedings of the National Academy of Sciences* **2017**, *114*, 4549.
[88] Chai, H., Chen, F., Zhang, S. J., Li, Y. D., Lu, Z., Kang, Y., Yu, L., *Lab on a Chip* **2019**.
[89] Ali, M. A., Jiao, Y., Tabassum, S., Wang, Y., Jiang, H., Dong, L., *2017 19th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS)*, IEEE **2017**, pp. 238.
[90] Smolka, M., Puchberger-Enengl, D., Bipoun, M., Klasa, A., Kiczkajlo, M., Śmiechowski, W., Sowiński, P., Krutzler, C., Keplinger, F., Vellekoop, M., *Precision agriculture* **2017**, *18*, 152.
[91] Aleklett, K., Kiers, E. T., Ohlsson, P., Shimizu, T. S., Caldas, V. E., Hammer, E. C., *The ISME journal* **2018**, *12*, 312.
[92] Stanley, C. E., Grossmann, G., i Solvas, X. C., deMello, A. J., *Lab on a Chip* **2016**, *16*, 622.
[93] Khater, M., de la Escosura-Muñiz, A., Merkoçi, A., *Biosensors and Bioelectronics* **2017**, *93*, 72.
[94] Abd-Elsalam, K. A., Prasad, R., *Nanobiotechnology applications in plant protection*, Springer 2018.
[95] Chen, L., Mungroo, N., Daikuara, L., Neethirajan, S., *Journal of nanobiotechnology* **2015**, *13*, 45.
[96] Jahn, I., Žukovskaja, O., Zheng, X.-S., Weber, K., Bocklitz, T., Cialla-May, D., Popp, J., *Analyst* **2017**, *142*, 1022.
[97] Almeida, M. I. G., Jayawardane, B. M., Kolev, S. D., McKelvie, I. D., *Talanta* **2018**, *177*, 176.
[98] Comunian, T. A., Ravanfar, R., Alcaine, S. D., Abbaspourrad, A., *Food research international* **2018**, *106*, 945.
[99] Cheng, X., Qi, Z., Burdyny, T., Kong, T., Sinton, D., *Bioresource technology* **2018**, *250*, 481.
[100] Ushikubo, F., Oliveira, D., Michelon, M., Cunha, R., *Food Engineering Reviews* **2015**, *7*, 393.
[101] Chung, S., Breshears, L. E., Perea, S., Morrison, C. M., Betancourt, W. Q., Reynolds, K. A., Yoon, J.-Y., *ACS Omega* **2019**, *4*, 11180.
[102] Arduini, F., Cinti, S., Caratelli, V., Amendola, L., Palleschi, G., Moscone, D., *Biosensors and Bioelectronics* **2019**, *126*, 346.
[103] Clinton-Bailey, G. S., Grand, M. M., Beaton, A. D., Nightingale, A. M., Owsianka, D. R., Slavik, G. J., Connelly, D. P., Cardwell, C. L., Mowlem, M. C., *Environmental science & technology* **2017**, *51*, 9989.
[104] Nightingale, A. M., Hassan, S.-u., Warren, B., Makris, K., Evans, G. W., Papadopoulou, E., Coleman, S., Niu, X., *Environmental science & technology* **2019**.
[105] Yang, W., Wei, X., Fraiwan, A., Coogan, C. G., Lee, H., Choi, S., *Sensors and Actuators B: Chemical* **2016**, *226*, 191.
[106] Boyer, H. C., Dutcher, C. S., *The Journal of Physical Chemistry A* **2017**, *121*, 4733.
[107] Metcalf, A. R., Boyer, H. C., Dutcher, C. S., *Environmental science & technology* **2016**, *50*, 1251.
[108] Jiang, X., Jing, W., Sun, X., Liu, Q., Yang, C., Liu, S., Qin, K., Sui, G., *ACS Sensors* **2016**, *1*, 958.

This article is protected by copyright. All rights reserved.

Accepted Article

[109] Liu, Q., Zhang, X., Yao, Y., Jing, W., Liu, S., Sui, G., *Sensors and Actuators B: Chemical* **2018**, *258*, 1138.
[110] Cai, P., Sun, X., Wu, Y., Gao, C., Mortimer, M., Holden, P. A., Redmile-Gordon, M., Huang, Q., *Soil Ecology Letters* **2019**, 1.
[111] König, S., Köhnke, M. C., Firle, A.-L., Banitz, T., Centler, F., Frank, K., Thullner, M., *Frontiers in Ecology and Evolution* **2019**, *7*, 290.
[112] Vidal, E., Lorenzetti, A. S., Lista, A. G., Domini, C. E., *Microchemical Journal* **2018**, *143*, 467.
[113] San Park, T., Li, W., McCracken, K. E., Yoon, J.-Y., *Lab on a Chip* **2013**, *13*, 4832.
[114] Guzman, J. M. C. C., Tayo, L. L., Liu, C.-C., Wang, Y.-N., Fu, L.-M., *Sensors and Actuators B: Chemical* **2018**, *255*, 3623.
[115] Fu, L.-M., Liu, C.-C., Yang, C.-E., Wang, Y.-N., Ko, C.-H., *Food chemistry* **2019**, *286*, 316.
[116] Shahvar, A., Saraji, M., Gordan, H., Shamsaei, D., *Talanta* **2019**, *197*, 578.
[117] Wang, S., Zheng, L., Cai, G., Liu, N., Liao, M., Li, Y., Zhang, X., Lin, J., *Biosensors and Bioelectronics* **2019**, *140*, 111333.
[118] Machado, J. M., Soares, R. R., Chu, V., Conde, J. P., *Biosensors and Bioelectronics* **2018**, *99*, 40.
[119] Motalebizadeh, A., Bagheri, H., Asiaei, S., Fekrat, N., Afkhami, A., *RSC advances* **2018**, *8*, 27091.
[120] Zhang, X.-X., Song, Y.-Z., Fang, F., Wu, Z.-Y., *Analytical and bioanalytical chemistry* **2018**, *410*, 2665.
[121] Zheng, L., Cai, G., Wang, S., Liao, M., Li, Y., Lin, J., *Biosensors and Bioelectronics* **2019**, *124*, 143.

This article is protected by copyright. All rights reserved.

Accepted Article

Figure Captions

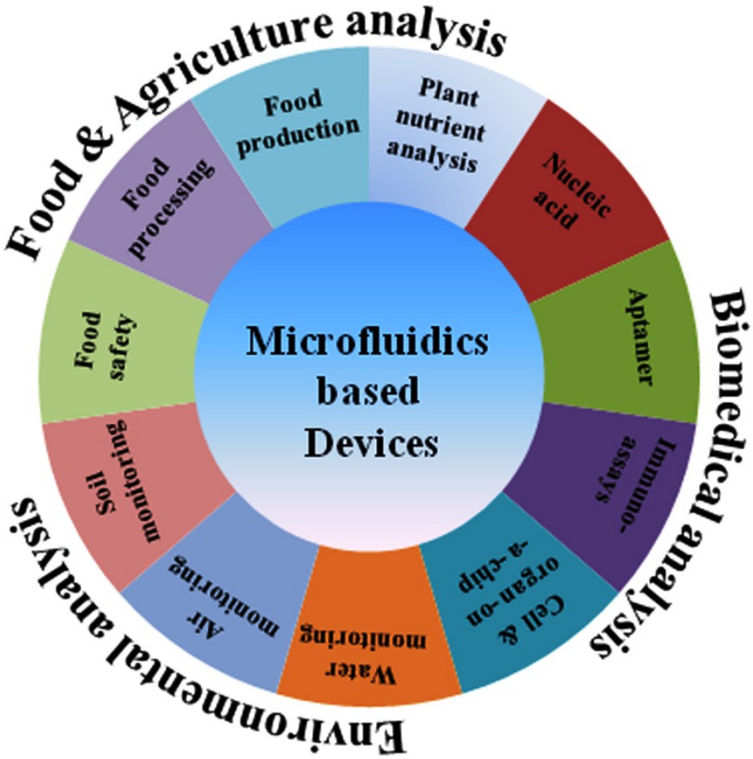


Figure 1: Scheme showing emerging applications of MFDs.

This article is protected by copyright. All rights reserved.

Accepted Article

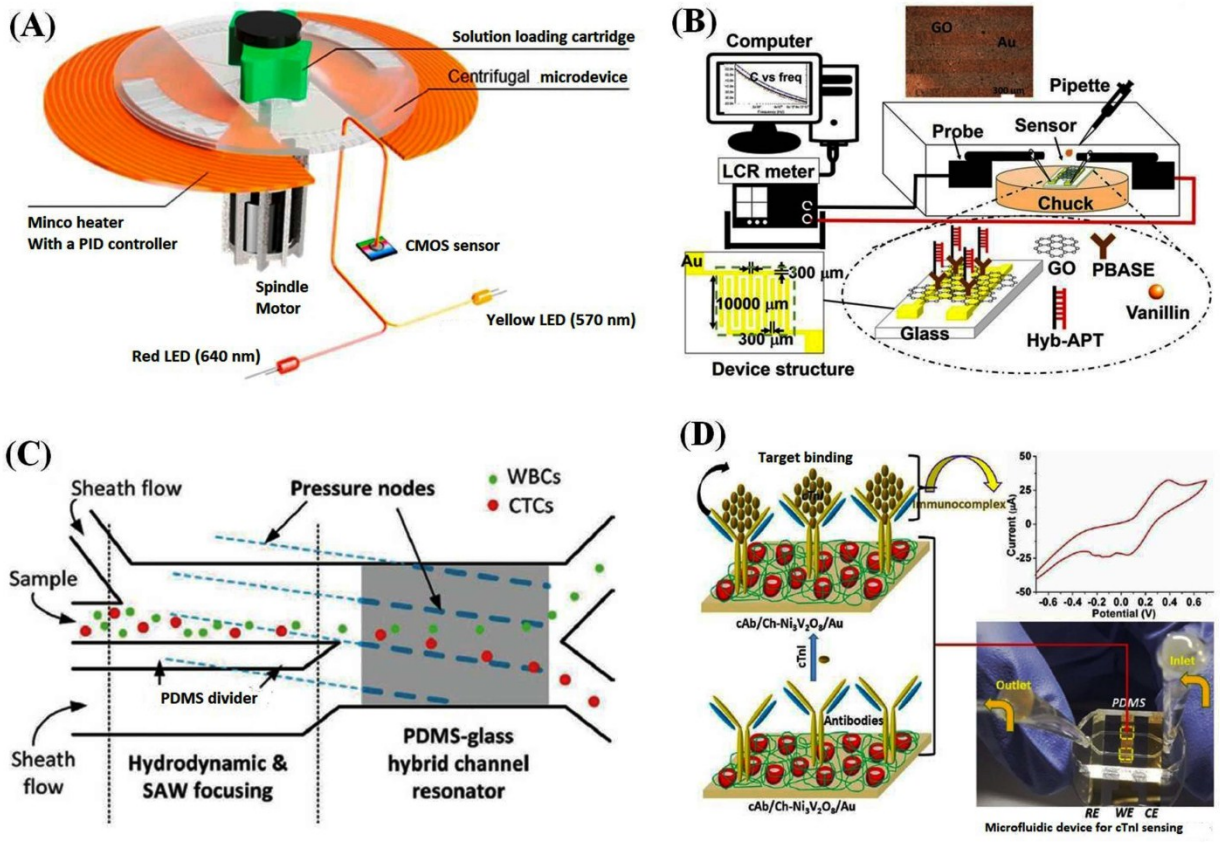


Figure 2: Schematic illustration of (A) integrated genetic analyzer platform with a motor, a couple of Minco heaters, and a UV–vis detector with two light sources at 640 and 570 nm. Reproduced with permission [34]. Copyright 2019, Elsevier, (B) enlarged view of the GO aptasensor, IDE structure showing sensing channel (enclosed by dotted lines) and an optical microscopic image of GO on glass with Au electrodes. Reproduced with permission [45]. Copyright 2019, Elsevier, (C) throughput acoustic circulating tumor cells (CTC) separation MFD. Cells were injected from the central inlet and focused along the PDMS divider. In the hybrid PDMS–glass channel resonating area, CTC (red dots) and WBC (green dots) are separated because of the difference in the lateral shift. Reproduced with permission [49]. Copyright 2019, John Wiley and Sons, and (D) Graphic illustration of the fabrication of antibody-based MFD for quantification of cTnI. Reproduced with permission [64]. Copyright 2019, Royal Society of Chemistry.

This article is protected by copyright. All rights reserved.

Accepted Article

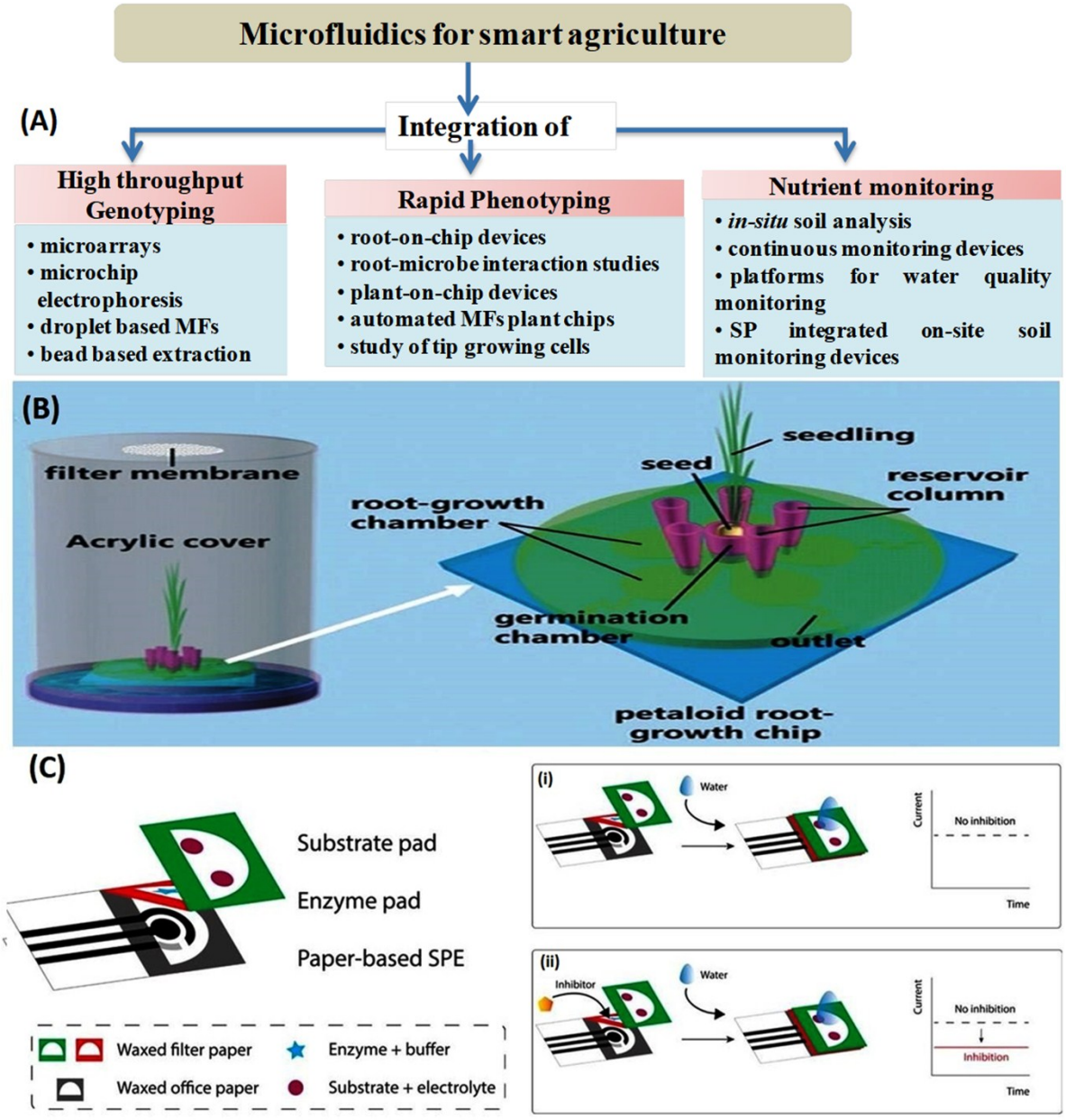


Figure 3: (A) Schematic showing an overview of MFDs used for smart agriculture, (B) Scheme showing culture system consisting of petaloid root growth microfluidic chip, a water tank and a transparent acrylic cover. Reproduced with permission [82]. Copyright 2019, Royal Society of Chemistry, and (C) Details of the enzymatic activity measurement using electrodes printed on office paper combined with a multi-pad system (i) and the mechanism of enzyme inhibition and signal production (ii). Reproduced with permission [97]. Copyright 2019, Elsevier.

This article is protected by copyright. All rights reserved.

Accepted Article

Table 1: Comparison of SP integrated MFDs for food analysis.

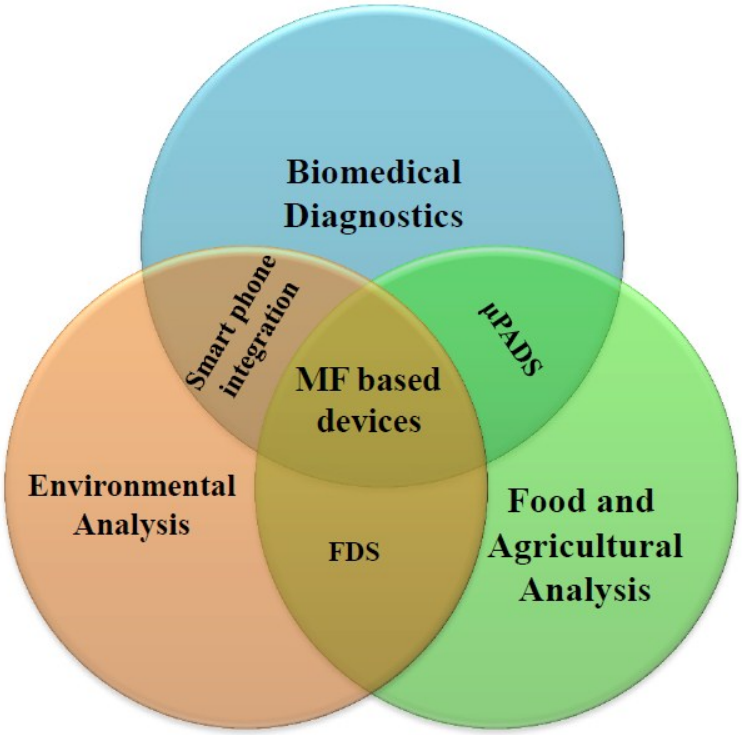
Detection target	Detection method	MFD	LOD	Ref.
Fluoride and nitrite	Image capture	Paper-based colorimetric system	0.13 mg L^{-1} for fluoride and 0.03 mg L^{-1} for nitrite	[112]
Salmonella typhi	Mie scattering	Paper-based immunosensing system	10^2 CFU/mL	[113]
Formaldehyde	Image analysis	Paper-based colorimetric system	0.2 ppm	[114]
Sulphur dioxide	CMOS	PET/paper chip	1.45 ppm/a.u	[115]
Sulfite	Image analysis	µPAD	0.04 µg /L	[116]
Salmonella typhimurium	Image capture analysis	PDMS MFC	58 CFU/mL	[117]
Ochratoxin A, Aflatoxin B1 and Deoxynivalenol	Image capture	PDMS MFC	< 40, 0.1–0.2 and < 10 ng/mL	[118]
As^{3+} and Hg^{2+}	Image processing	PDMS MFC with gold nanoparticles	-	[119]
Nitrite	Image processing	Paper-based MFC	73 ng/mL	[120]
E. coli	Imaging	PDMS based MFC	50 CFU/mL	[121]

This article is protected by copyright. All rights reserved.

Accepted Article

Graphical Abstract

Microfluidics has been used for various applications in healthcare, along with other biological and medical areas. This review focuses on the emerging demands of microfluidics in different fields, i.e., molecular diagnostics, environmental analysis and food and agriculture research, etc. It also aims to reveal new opportunities in these areas and future prospects of commercial microfluidics based devices.



This article is protected by copyright. All rights reserved.

RSC Advances

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *RSC Adv.*, 2023, **13**, 33675

Energy transfer mechanisms and color-tunable luminescence of $Tm^{3+}/Tb^{3+}/Eu^{3+}$ co-doped $Sr_4Nb_2O_9$ phosphors for high-quality white light-emitting diodes

Ravina,[a] Kanishk Poria,[b] Mukesh K. Sahu,[c] A. Kumar,[a] Anu,[d] Sajjan Dahiya,[e] Nisha Deopa *[a] and A. S. Rao[d]

This paper investigates the synthesis and luminescence characteristics of $Tm^{3+}/Tb^{3+}/Eu^{3+}$ co-doped $Sr_4Nb_2O_9$ (SNB) phosphors as potential candidates for white light-emitting diodes (WLEDs). The study explores the energy transfer mechanisms and color-tunable characteristics of these phosphors. The SNB phosphors were prepared using a solid-state reaction method, and their structural and morphological properties were characterized using X-ray diffraction (XRD), scanning electron microscopy (SEM), energy-dispersive X-ray spectroscopy (EDS), and Fourier-transform infrared (FT-IR) spectroscopy. The diffuse reflectance, photoluminescence (PL) and time resolved photoluminescence (TRPL) properties were investigated, revealing efficient energy transfer processes from Tm^{3+} to Tb^{3+} and Eu^{3+} ions. The energy transfer mechanisms were determined through critical distance calculations and analysis of multipolar interactions. The co-doped phosphors exhibited tunable emission colors ranging from blue to white light, with controllable correlated color temperatures (CCTs) and high color rendering indices (CRIs). The CIE chromaticity coordinates were optimized to approach neutral white light. The PL intensity is maintained at 81.19% at 150 °C of that of room temperature which showcases the remarkable thermal stability of the as-prepared phosphors. The results highlight the potential of $Tm^{3+}/Tb^{3+}/Eu^{3+}$ co-doped SNB phosphors for generating high-quality, color-tunable white light for advanced lighting applications.

Received 14th August 2023
Accepted 8th November 2023

DOI: 10.1039/d3ra05519a

rsc.li/rsc-advances

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

1 Introduction

With technological advancements, especially the widespread utilization of rare earth ion-doped phosphors, white light-emitting diodes (WLEDs) have progressively emerged as the optimal choice for next-generation lighting solutions. This preference arises from their inherent advantages, including cost-effectiveness, high efficiency, and environmental friendliness.[1–3] Currently, white light can be achieved through two approaches. The first involves integrating a blue InGaN chip with YAG:Ce^{3+} yellow phosphors. However, this approach lacks the red segment of the spectrum, leading to a diminished color rendering index (CRI) and an elevated correlated color temperature (CCT). This deviation from the natural white light region poses a challenge to achieving high-quality white light illumination. Another approach involves using a near-ultraviolet chip combined with three-color phosphors. This method is capable of meeting the requirements for a high color rendering index (CRI) and achieving a suitable correlated color temperature (CCT).[4] By employing this approach, it becomes possible to produce white light that closely resembles natural lighting conditions. Therefore, the development of color-tunable phosphors on a single-phase substrate is a highly significant research direction within the field of white light-emitting diodes (WLEDs). This approach aims to create phosphors that can be adjusted to produce a wide range of colors, enabling versatile and customizable lighting solutions.

Rare-earth ions exhibit a remarkably diverse electronic structure, rendering them capable of emitting photoluminescence across a wide range of wavelengths, including visible and infrared light. Particularly in the visible light region, their emission ability and vivid colors stand out, making rare-earth ions highly desirable for various applications, such as lighting and display technologies.[5] The fusion of three primary colors—red, green, and blue—is well-known to yield white light.

[a]*Department of Physics, Chaudhary Ranbir Singh University, Rohtak Bypass Road, Jind-126102, Haryana, India. E-mail: nispectro999@gmail.com; Tel: +918860551723*
[b]*Department of Physics, Panjab University, Chandigarh-160014, India*
[c]*School of Computing, Graphic Era Hill University, Haldwani, Uttarakhand-263136, India*
[d]*Department of Applied Physics, Delhi Technological University, Bawana Road, New Delhi-110 042, India*
[e]*Department of Physics, Maharshi Dayanand University, Rohtak-124001, Haryana, India*

© 2023 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

In this context, different phosphors utilize Tm^{3+}, Eu^{3+}, and Tb^{3+} ions as blue, red, and green emitters, respectively. These ions are chosen for their abundant electron transitions, contributing to the emission properties of the phosphors. For instance, Tm^{3+} doping in inorganic compounds often yields excellent blue-emitting performance due to the blue transition from $^1D_2 \rightarrow {}^3F_4$.[6,7] Eu^{3+} ions emit in the orange-red region of the spectrum through the $^5D_0 \rightarrow {}^7F_J$ ($J = 0$–4) transitions,[8,9] while Tb^{3+} ions are commonly used as blue-green emitters with the $^5D_4 \rightarrow {}^7F_J$ ($J = 6$–3) transitions.[10] Co-doping Tm^{3+}, Eu^{3+}, and Tb^{3+} ions into a single host material allows for the potential generation of a single-phased phosphor emitting white light.[11] This approach permits the generation of white light by combining emissions from these different ions. Furthermore, the selection of an appropriate host material is vital for synthesizing phosphors with high performance. The host material plays a crucial role by providing suitable crystal fields for the activator ions. The crystal structure and properties of the host material significantly influence the luminescent properties and efficiency of the phosphor. Hence, careful consideration and optimization of the host material are essential to achieve phosphors with desirable performance characteristics.

The extensive study of niobate group materials is driven by their remarkable electro-optic, dielectric, ferroelectric, pyroelectric, and piezoelectric properties.[12,13] Niobate serves as a highly favorable host matrix for doping rare-earth ions due to its exceptional luminescent properties, excellent chemical stability, and good thermal stability. Concerning characteristics related to light emission, the optical absorption observed in niobates is attributed to charge transfer occurring within the NbO_6 complex.[14] This absorption behavior significantly influences their luminescent properties. While certain pure niobates may not exhibit luminescence, substantial research has been conducted on rare-earth-doped niobates due to their potential as light-emitting materials. The incorporation of rare-earth ions into niobate structures allows for the creation of luminescent centers, expanding the range of applications in areas such as lighting, displays, and optoelectronics.

2 Experimental procedure and instrumentations

Through the solid-state reaction method, we synthesized samples of strontium niobate phosphor $Sr_4Nb_2O_9$ (SNB) doped with Tm^{3+}, Tb^{3+}, and Eu^{3+}. The precursor materials $SrCO_3$ (98%), Nb_2O_5 (99.9%), Tm_2O_3 (99.9%), Tb_4O_7 (99.9%), Eu_2O_3 (99.99%) were of analytical grade. Following their precise stoichiometric ratio, we weighed and thoroughly mixed the precursors using an agate mortar until achieving uniformity. These homogeneous mixtures were then poured in an alumina crucible and sintered at 1350 °C for 7 hours, followed by gradual cooling to ambient temperature. We labeled the resulting phosphors as follows: SNB:xTm^{3+} ($x = 0.01$, 0.03, 0.05, 0.07, 0.10 mol), SNB:0.03Tm^{3+}, yTb^{3+} ($y = 0.01$, 0.02, 0.03, 0.04, 0.05, 0.06, 0.08, 0.10 mol) denoted as TT1-TT10, SNB:0.03Tm^{3+}, 0.06Tb^{3+}, zEu^{3+} ($z = 0.01$, 0.03, 0.05 and 0.07 mol) labeled as TTE1-TTE7. Finally, we finely powdered the prepared samples for subsequent characterization.

Structural analysis of the samples was conducted using X-ray diffraction (XRD). A Bruker D8 Advance diffractometer equipped with a nickel filter and a Cu Kα radiation source (wavelength $\lambda = 1.5406$ Å) was utilized within the 2θ range of 20° to 80°. For examination of sample morphology and elemental composition, scanning electron microscopy (SEM) and energy-dispersive X-ray spectroscopy (EDS) were performed using a Zeiss Gemini 300 instrument. Additionally, Fourier-transform infrared (FT-IR) spectra of the as-synthesized samples were recorded using a Nicolet IS50 FT-IR apparatus. UV spectrum analysis was conducted using a Jasco V-770 spectrophotometer with light sources of deuterium and halogen and a resolution of 1.0 nm. For validation of PL excitation and emission spectra, a Jasco 8300 FP spectrofluorophotometer employing 300 W xenon lamp acts as excitation source and a resolution of 1.0 nm was used. PL decay measurements were performed using the Hitachi F 7000 spectrofluorometer in conjunction with a 450 watts microsecond xenon flash lamp. The time-resolved photoluminescence (TDPL) was performed using the FLAME-S-XR1-ES Ocean Optics spectrometer along with a sample holder along with a heating assembly.

3 Results and discussion

3.1. Study of XRD pattern

Fig. 1 presents the XRD patterns of SNB phosphors doped with 3 mol% Tm^{3+}, 3 mol% Tm^{3+} and 6 mol% Tb^{3+}. The diffraction patterns demonstrate consistency with standard data (Card No. 048-0558), confirming the successful incorporation of Tm^{3+}, Tb^{3+} and Eu^{3+}ions into the host lattice. The XRD patterns confirm that the as-prepared phosphor possesses a crystalline cubic structure within the *Fm*3*m* space group, showing no structural alterations.[15] The impurities at 28° and 50° in the XRD is because of the difference between both ionic radius and charge numbers between the Ln^{3+} ions and Sr^{2+} ions. The average crystallite sizes (D) were calculated using Debye–Scherer equation (eqn (1)) from the literature,[16]

$$D = \frac{K\lambda}{\beta \cos \theta} \quad (1)$$

resulting in values of 41.64 nm for SNB:3 mol% Tm^{3+}, 40.61 nm for SNB:3 mol% Tm^{3+} and 6 mol% Tb^{3+}, and 41.07 nm for SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+}, and 5 mol% Eu^{3+}.

Fig. 2 demonstrates the XRD patterns of the SNB:3Tm^{3+}/6Tb^{3+}/zEu^{3+} ($z = 1$, 3, 5 and 7 mol%) samples which are consistent with the standard data. The average crystallite size were evaluated and found to be 42.26, 41.88, 41.07 and 42.12 nm, respectively.

3.2. SEM and EDS studies

SEM imaging, SEM-EDS mapping, and EDS spectrum analysis were employed to investigate the morphology and elemental composition of the SNB sample doped with 3 mol% Tm^{3+} and 1 mol% Tb^{3+}, as illustrated in Fig. 3. In Fig. 3(a), the SEM image

 © 2023 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

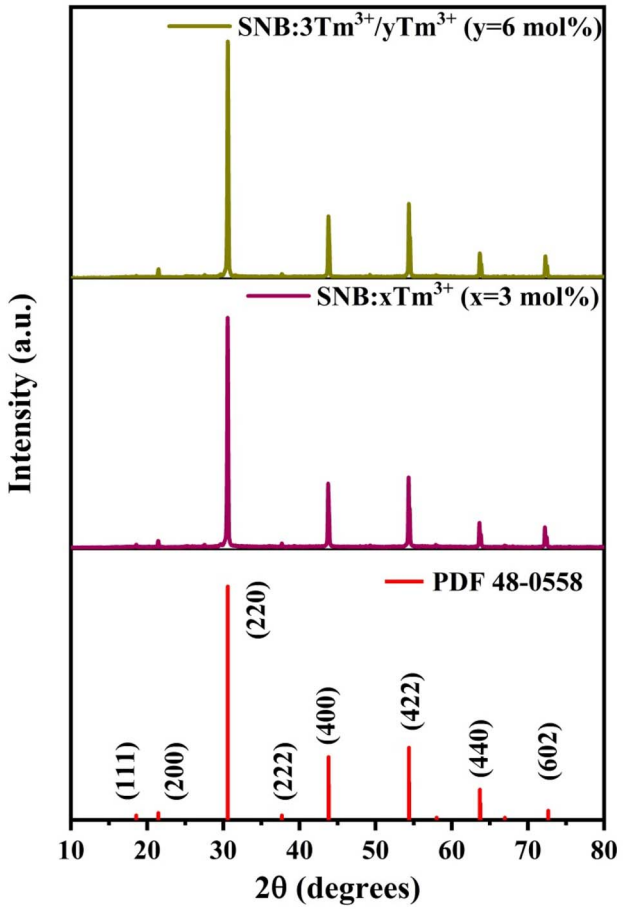


Fig. 1 XRD patterns of SNB:xTm^{3+} (x = 3 mol%), SNB:3Tm^{3+}/yTb^{3+} (y = 6 mol%) samples and compared with standard data.

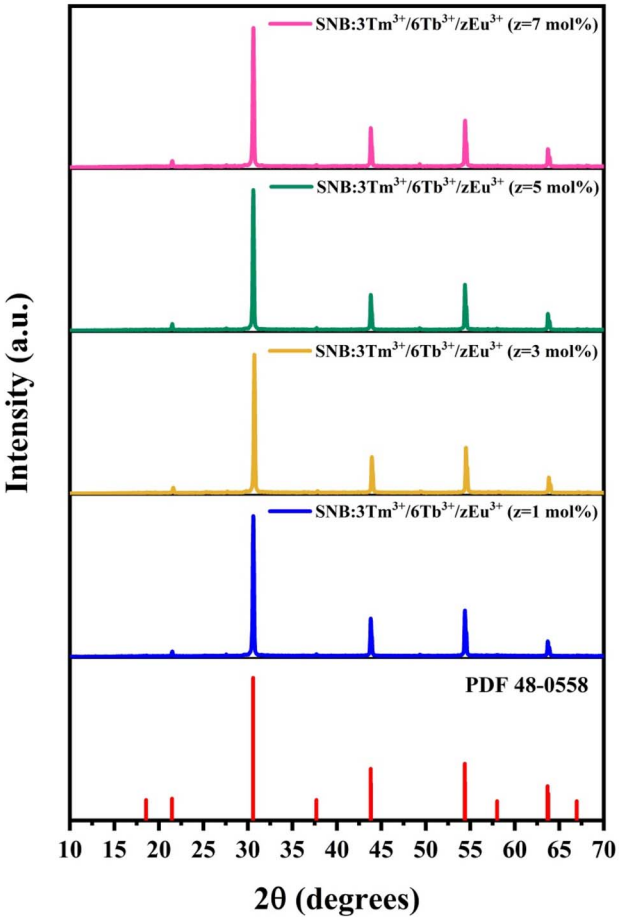


Fig. 2 XRD patterns of SNB:3Tm^{3+}/6Tb^{3+}/zEu^{3+} (z = 1, 3, 5 and 7 mol%) samples and compared with standard data.

unveils the presence of regular and smooth micron-sized particles within the SNB sample. Elemental mapping results in Fig. 3(b–f) exhibit a uniform distribution of Sr, Nb, O, Tm, and Tb elements across the grains. The EDS spectrum depicted in Fig. 3(g) displays distinct peaks corresponding to Nb, O, and Sr, along with smaller peaks representing Tm and Tb. No indications of impurity elemental peaks are evident, indicating the successful integration of Tm^{3+} and Tb^{3+} ions into the SNB host matrix. The inset within Fig. 3(g) illustrates the elemental compositions.

3.3. FTIR studies

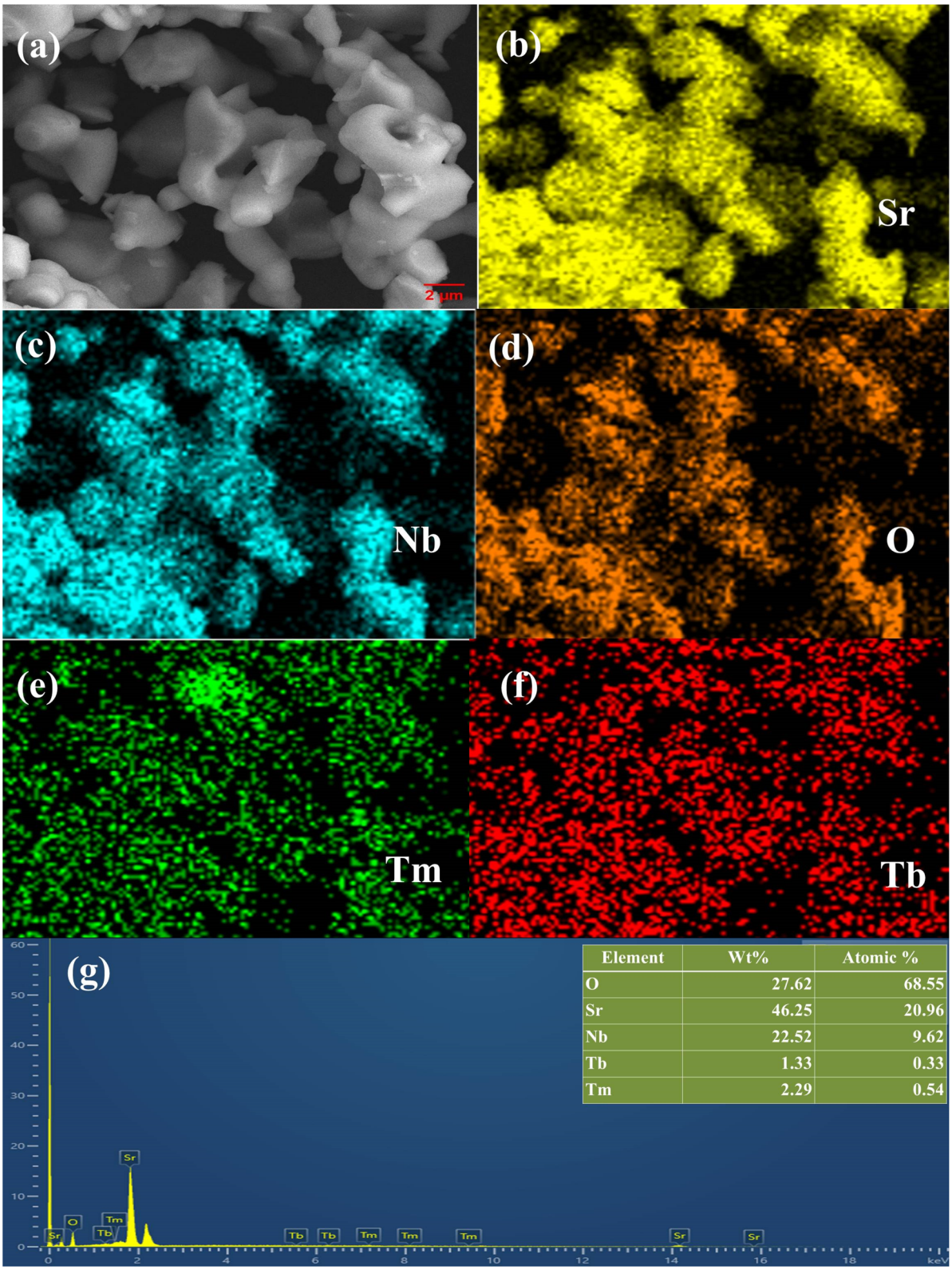
The structural components and functional groups present in the as-synthesized SNB:xTm^{3+} (x = 1, 3, 5, 7, and 10 mol%) phosphor were analyzed by recording FT-IR spectra in the wavenumber range of 4000–400 cm^{-1}. Fig. 4 illustrates the FT-IR spectra of the synthesized phosphor. FT-IR spectroscopy was employed in the initial stage of material analysis to ensure the quality of the product. The shift in peak positions indicated the presence of contaminants. The spectra exhibited analogous positions of shoulder, as shown in Fig. 4. The FT-IR spectra revealed the presence of nine peaks at approximately 418, 435, 525, 651, 861, 1442, 2885, 2948, and 2989 cm^{-1}. These peaks were assigned based on previous reported papers. The peak observed near 418 and 525 cm^{-1} corresponds to the Sr–O stretching modes.[17] The peaks around 435 and 861 cm^{-1} are attributed to the coupling between Nb–O stretching mode.[18] The peak at approximately 651 cm^{-1} is due to Tm–O stretching vibrations.[19] The peaks observed in the range of 1442 and 2989 cm^{-1} are associated with water molecules.[20] The disturbances in the form of peaks between 1000 to 3000 cm^{-1} may indicate the presence of residual organic compounds.[21] Furthermore, the absence of significant evidence of carbonates or carboxyl groups in the FT-IR spectra confirms the synthesis of phosphor materials from their precursor carbonates.

3.4. Diffuse reflectance study

Diffuse reflectance spectra were collected for the SNB samples doped with 3 mol% Tm^{3+} and SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+} over the wavelength range of 200–1800 nm, spanning the UV-visible and near-IR regions at room temperature, as depicted in Fig. 5.

In the phosphors doped solely with Tm^{3+}, the absorption bands located at approximately 684, 788, 1216, and 1644 nm

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.



Element	Wt%	Atomic %
O	27.62	68.55
Sr	46.25	20.96
Nb	22.52	9.62
Tb	1.33	0.33
Tm	2.29	0.54

Fig. 3 (a) SEM images. (b–f) SEM-EDS mapping. (g) EDS spectrum of SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+} phosphor [inset shows the elements weight and atomic percent].

correspond to transitions from the ground level $^3H_6$ to $^3F_3$, $^3H_4$, $^3H_5$ and $^3F_4$ levels, respectively.[22] In phosphors co-doped with both Tm^{3+} and Tb^{3+}, a captivating interplay of energy level transitions is observable. Tm^{3+} ions exhibit transitions at 684, 788, 1216, and 1644 nm, while Tb^{3+} ions exhibits the $^7F_6 \rightarrow {}^5D_3$ energy level transition at 386 nm.[23]

© 2023 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

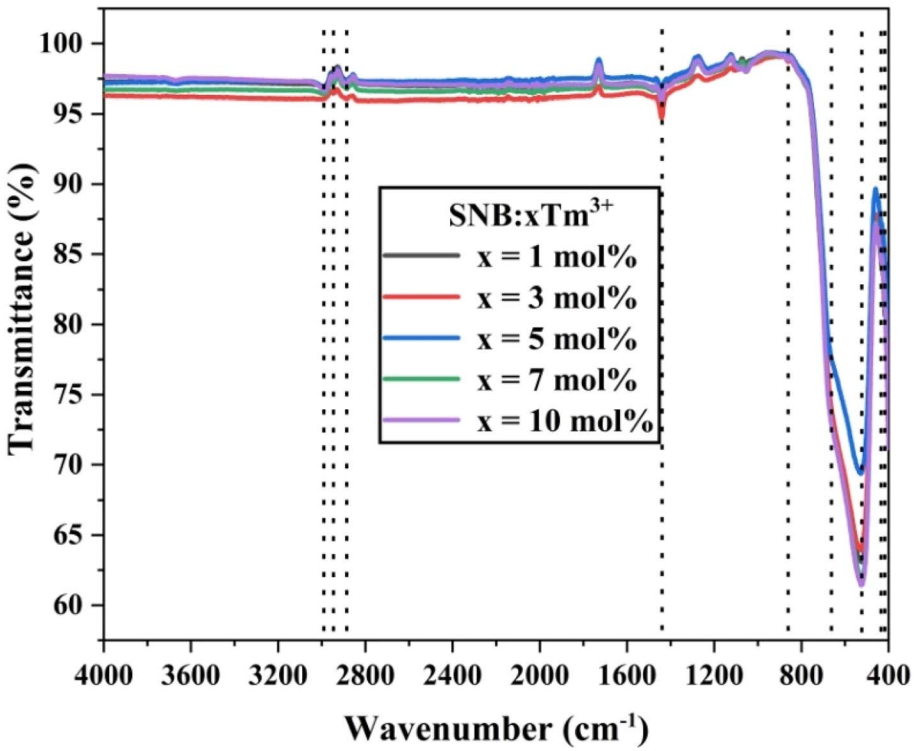


Fig. 4 FT-IR spectra of SNB:xTm^{3+} samples.

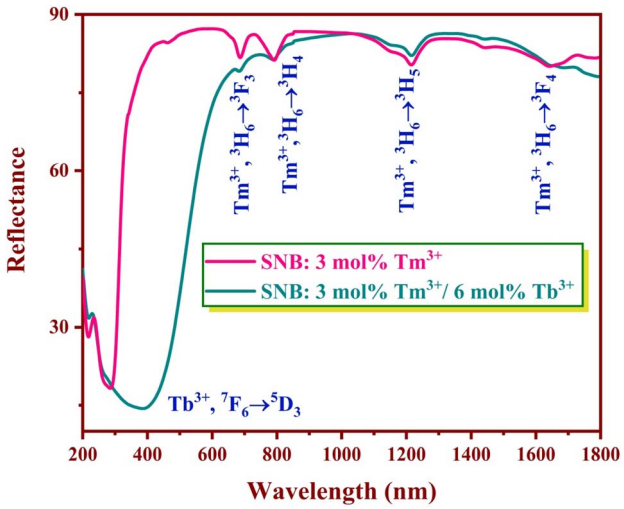


Fig. 5 Diffuse reflectance spectra of SNB:3 mol% Tm^{3+} and SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+} phosphors.

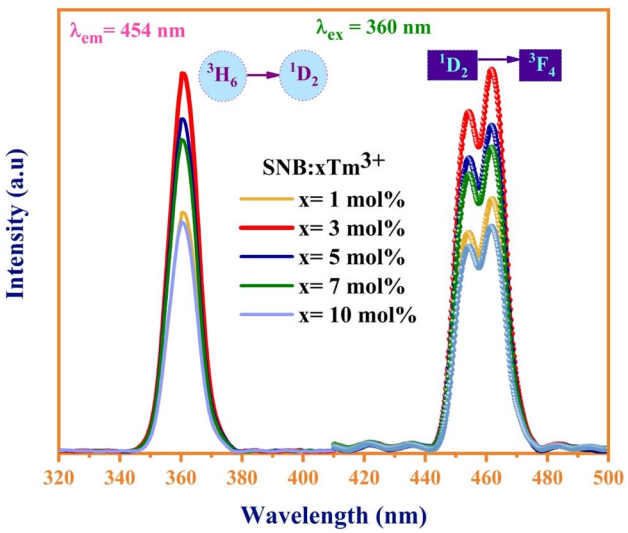


Fig. 6 PLE and PL spectra of SNB:xTm^{3+} phosphors.

3.5. Photoluminescence properties

Fig. 6 displays the photoluminescence excitation and emission spectra of the synthesized SNB:xTm^{3+} (x = 1, 3, 5, 7, 10 mol%) phosphors. The dominant excitation peak at 360 nm can be attributed to the distinctive $^3H_6 \rightarrow {}^1D_2$ transitions of Tm^{3+} ions, effectively stimulated by near-ultraviolet (NUV) light. In the emission spectra, a broad peak ranging from 440 to 480 nm under the 360 nm excitation for SNB:xTm^{3+} (x = 1, 3, 5, 7, 10) phosphors is associated with the $^1D_2 \rightarrow {}^3F_4$ transition.[24]

Interestingly, the peak intensity of the samples initially increases and then diminishes as the Tm^{3+} doping concentration rises. This intensity attains its peak at x = 3 mol%, which is attributed to concentration quenching, as depicted in Fig. 6. This phenomenon arises from the cross relaxation mechanism due to variations in the spacing between luminescence centers. At lower concentrations of Tm^{3+} ions, the spacing between them is sufficiently small to effectively neglect fluorescence quenching.[25] However, with an increasing concentration of Tm^{3+} ions,

© 2023 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

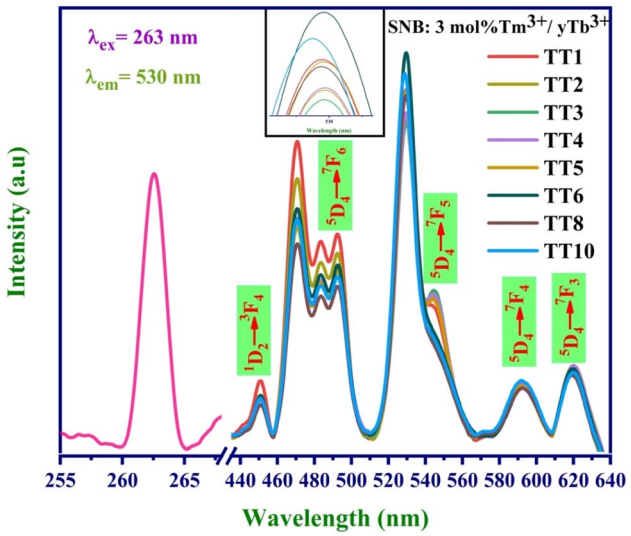


Fig. 7 PLE and PL spectra of SNB:3 mol% Tm^{3+}/yTb^{3+} phosphors.

the gap between the ions diminishes, falling below the critical transfer distance. This results in a higher utilization of Tm^{3+} ions at the $^1D_2$ level, consequently leading to a decrease in luminous efficiency.

3.6. Energy transfer from Tm^{3+} to Tb^{3+}

To explore the energy transfer mechanism from Tm^{3+} to Tb^{3+} ions, we prepared SNB phosphors doped with 3 mol% Tm^{3+} and varying concentrations of Tb^{3+} (denoted as y mol%, where $y = 1$, 2, 3, 4, 5, 6, 8, and 10 mol%). Fig. 7 illustrates the emission spectra under 263 nm UV light excitation, and a spectral comparison reveals that the emission peak positions remain nearly consistent across the 8 samples, appearing at approximately 450, 492, 530, 591, and 620 nm.[26]

Analysis of the emission spectra indicates that as the concentration of Tb^{3+} ions increases, the luminescent intensity of the distinct blue emission peak at 450 nm decreases. Meanwhile, the emission peaks attributed to Eu^{3+} ions at 530 nm initially experience an increase, reaching a maximum for Tm-3 mol% and Tb-6 mol%, followed by a gradual decrease. This observation suggests the occurrence of a concentration quenching phenomenon, signifying the energy transfer from Tm^{3+} to Tb^{3+} ions.

The efficiency of energy transfer from Tm^{3+} to Tb^{3+} ions can be quantified using the formula:[27]

$$\eta = 1 - \frac{I_T}{I_T + I_D} \quad (2)$$

In this equation, I_T and I_D denote the photoluminescence intensities of the emission peaks at 450 and 530 nm, respectively.

The calculated efficiency values for the TT1, TT2, TT3, TT4, TT5, TT6, TT8, and TT10 phosphors are 86.1%, 88.5%, 89.7%, 90.2%, 90%, 91.3%, 90.3%, and 90.4%, respectively. These results underscore a significant improvement in energy transfer efficiency with increasing concentration, reaching its peak value in the TT6 phosphor sample. The higher concentration of Tb^{3+} ions surrounding the Tm^{3+} ions promotes a greater likelihood of cross relaxation between the transitions of Tm^{3+} ($^1D_2 \rightarrow {}^3F_4$) and Tb^{3+} ($^5D_4 \rightarrow {}^7F_5$).[28] This phenomenon elucidates the concentration quenching and the reduction in emission intensity at 450 nm. The quantum yield for the TT6 phosphor was evaluated using the equations mentioned in literature and comes out to be 41% which is higher than the other reported works.[29,30]

Fig. 8 graphically portrays the energy transfer efficiency between Tm^{3+} and Tb^{3+} ions, alongside the emission intensities corresponding to the $^1D_2 \rightarrow {}^3F_4$ transition of Tm^{3+} and the $^5D_4$

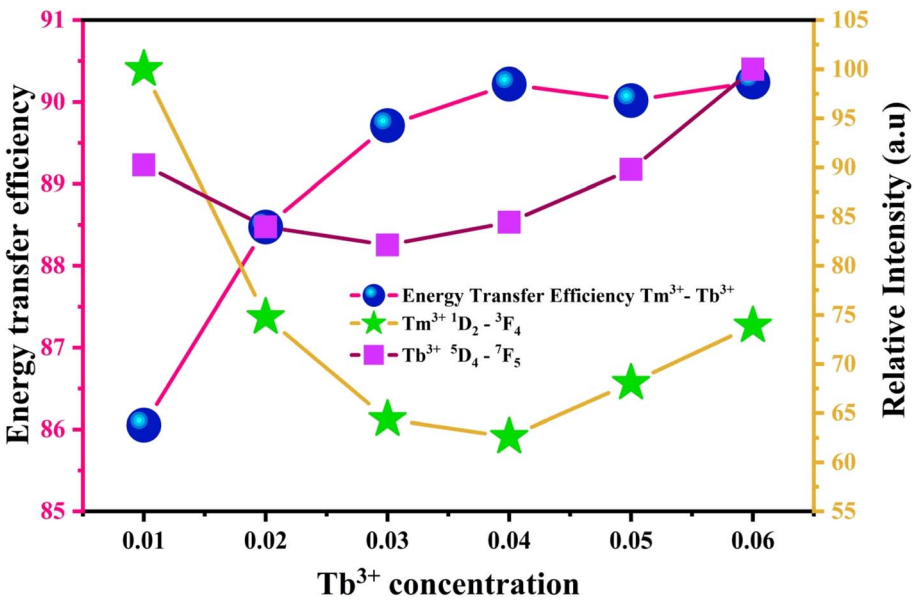


Fig. 8 Energy transfer efficiency of $Tm^{3+} \rightarrow Tb^{3+}$ and the luminous intensity curve of the monitoring wavelength were 450 nm (Tm^{3+}) and 530 nm (Tb^{3+}).

© 2023 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

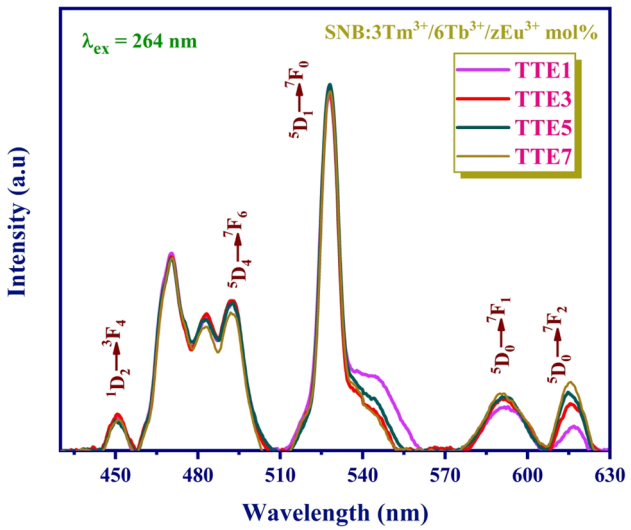


Fig. 9 PL spectra of SNB:3 mol% Tm^{3+}/6 mol% Tb^{3+}/*z* mol% Eu^{3+} phosphors under 264 nm excitation wavelength.

→ $^7F_5$ transition of Tb^{3+}. As the concentration of Tb^{3+} ions increases, the emission intensity of Tm^{3+} decline, while conversely, the emission intensity of Tb^{3+} and the energy transfer efficiency increases.

The emission spectra of SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+}, and *z* mol% Eu^{3+} phosphors illuminated by 264 nm UV light are depicted in Fig. 9. These spectra reveal distinct emission peaks associated with Tm^{3+}, Tb^{3+}, and Eu^{3+} ions at 450, 493, 528, 593, and 618 nm.[31,32] Notably, the positions of these characteristic peaks remain largely consistent across varying Eu^{3+} concentrations. Specifically, the emission intensity at 528 nm initially increases until the TTE5 phosphor, after which it diminishes with higher Eu^{3+} ion concentrations.

Fig. 10 provides further insight, illustrating changes in emission intensity. The emission intensity from Tm^{3+} ions at

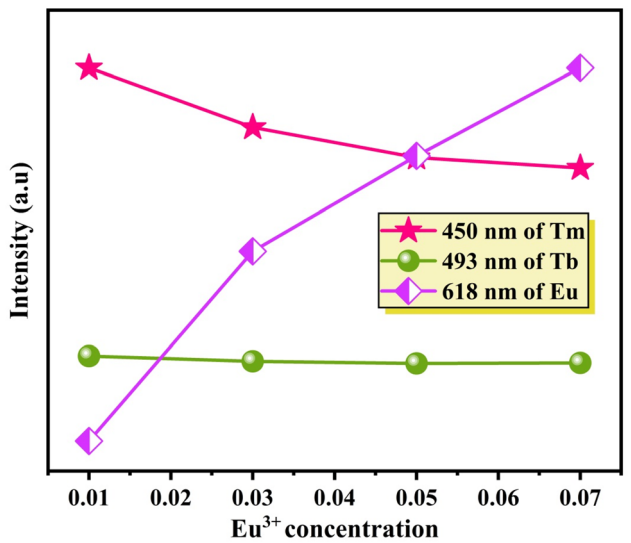


Fig. 10 Variation of the emission intensity of Tm^{3+}/Tb^{3+}/Eu^{3+} ions with the concentration of Eu^{3+} ions.

450 nm and Tb^{3+} ions at 493 nm declines, while the emission intensity attributed to Eu^{3+} ions at 618 nm rises as the Eu^{3+} concentration increases. This observation strongly implies a dual energy transfer mechanism, occurring from both Tm^{3+} and Tb^{3+} ions to Eu^{3+} ions.

Energy transfer mechanism of Tm^{3+}, Tb^{3+} and Eu^{3+} ions is illustrated in the schematic diagram presented in Fig. 11. The ions at ground state get excited to the higher energy levels of Tm^{3+}, Tb^{3+} and Eu^{3+} by the 264 nm UV excitation. After that ions gets de-excited to lower energy states $^5D_{1,0}$ of Eu^{3+} and $^5D_4$ of Tb^{3+} through non-radiative emission aided by phonons and relaxed to the ground state. In the case of Tm^{3+} ions, the blue photons emitted at 545 nm associated with $^1D_2 \rightarrow {}^3F_4$ transition. In the case of Tb^{3+} ions, the cyan and green photons emitted at 492 nm and 530 nm associated with $^5D_4 \rightarrow {}^7F_6$ transition and $^5D_4 \rightarrow {}^7F_5$ transition, respectively and for Eu^{3+} ions transitions at 593 nm ($^5D_0 \rightarrow {}^7F_1$) and 618 nm ($^5D_0 \rightarrow {}^7F_2$) emit orange and red photons, respectively. As the energy gap between the excited state $^1G_4$ of Tm^{3+} ions and $^5D_4$ of Tb^{3+} ions is smaller so they can bridge with the phonons of host lattice which results in non-radiative relaxation in Tm^{3+}. Some part of the energy is migrated to Tb^{3+} which results in transition from $^7F_6$ to $^5D_4$.[33] The non-radiative resonant energy transfer from Tm^{3+} ($^1G_4$) to Eu^{3+} ions ($^5D_2$) is because of lesser energy difference between these two states. This result in the Eu^{3+} ions transition from $^7F_1$ level to $^5D_2$ state and then return to $^5D_0$ state through non-radiatively which accompanied further radiative emission from this excited state to $^7F_1$ and $^7F_2$ state.

The energy transfer interactions between sensitizers and activators can be broadly categorized into two primary types: exchange interaction and electric multipolar interaction. The characterization of these interactions hinges on a critical parameter known as the critical distance (R_c) between the sensitizers and activators. If the value of R_c remains below 5 Å, exchange interaction is likely; however, if R_c surpasses this threshold, electric multipolar interaction becomes more probable.[34]

In order to determine the nature of the interaction between Tm^{3+} (sensitizer) and Tb^{3+} (activator), the critical distance R_c can be quantified using an equation outlined in existing literature.[35]

$$R_c = 2\left(\frac{3V}{4\pi N X_c}\right)^{\frac{1}{3}} \quad (3)$$

The computed R_c value is found to be 18.04 Å, which significantly exceeds the 5 Å threshold. This outcome strongly suggests that the energy transfer mechanism is primarily driven by electric multipolar interactions.

Utilizing Dexter's energy transfer formula for multipolar interaction and applying Reisfeld's approximation, the relationship can be presented as follows:[36,37]

$$\frac{I_{SO}}{I_S} \propto C_{Tm+Tb}^{\,n/3} \quad (4)$$

Here, I_{SO} and I_S denotes the luminescence intensities of Tm^{3+} ions without and with Tb^{3+} ions, respectively. C_{Tm+Tb} is the

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

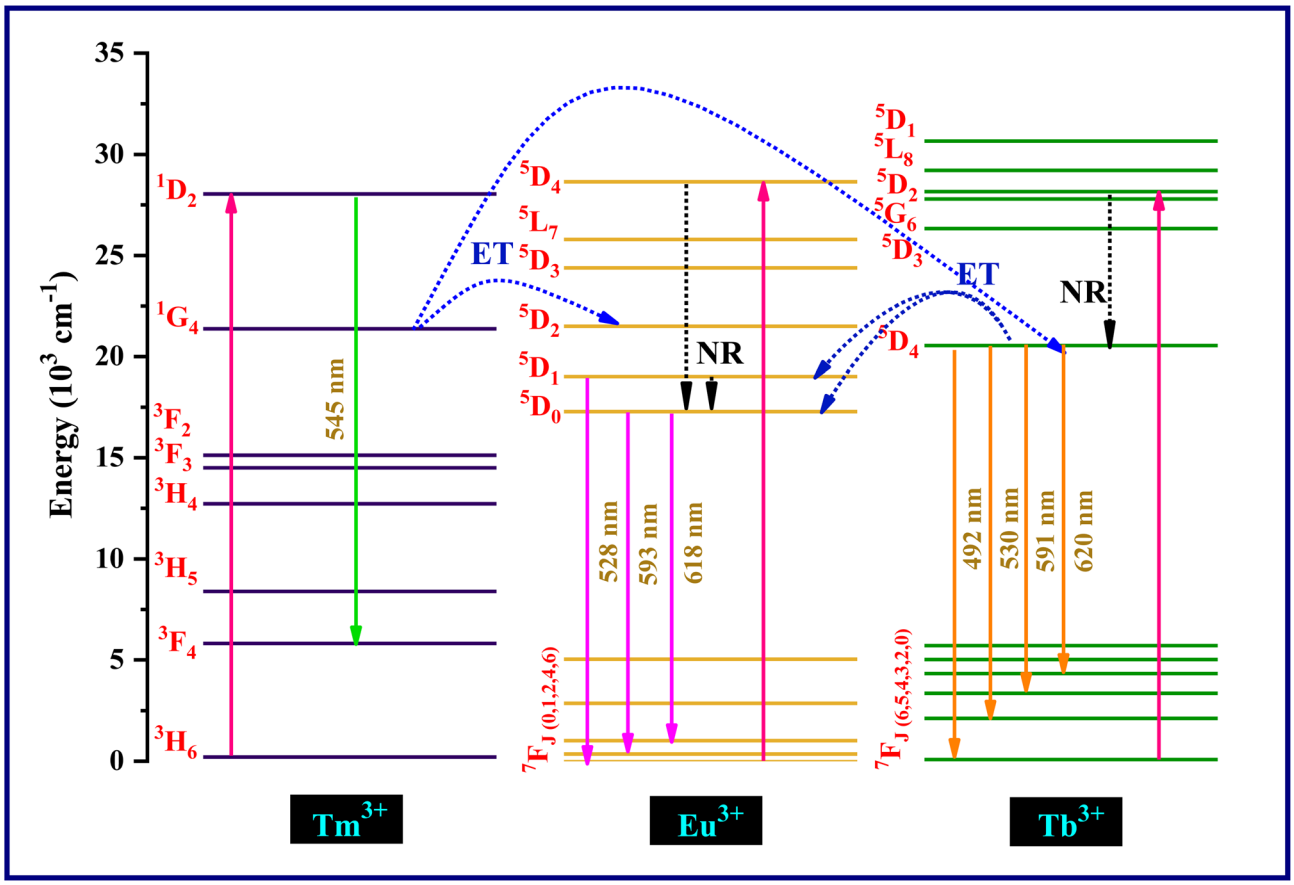


Fig. 11 Energy level diagram of Tm^{3+}/Tb^{3+}/Eu^{3+} ions at 264 nm.

combined concentration of Tb^{3+} and Tm^{3+} ions. The ratio $I_{SO}/I_S \propto C_{Tm+Tb}{}^{n/3}$ with $n = 6$, 8 and 10 corresponds to different exchange interactions: dipole–dipole, dipole–quadrupole, and quadrupole–quadrupole interactions, respectively.[38] The relationship between I_{SO}/I_S and $C_{Tm+Tb}{}^{n/3}$ is illustrated in Fig. 12. It is evident that the most optimal linear fitting is observed when

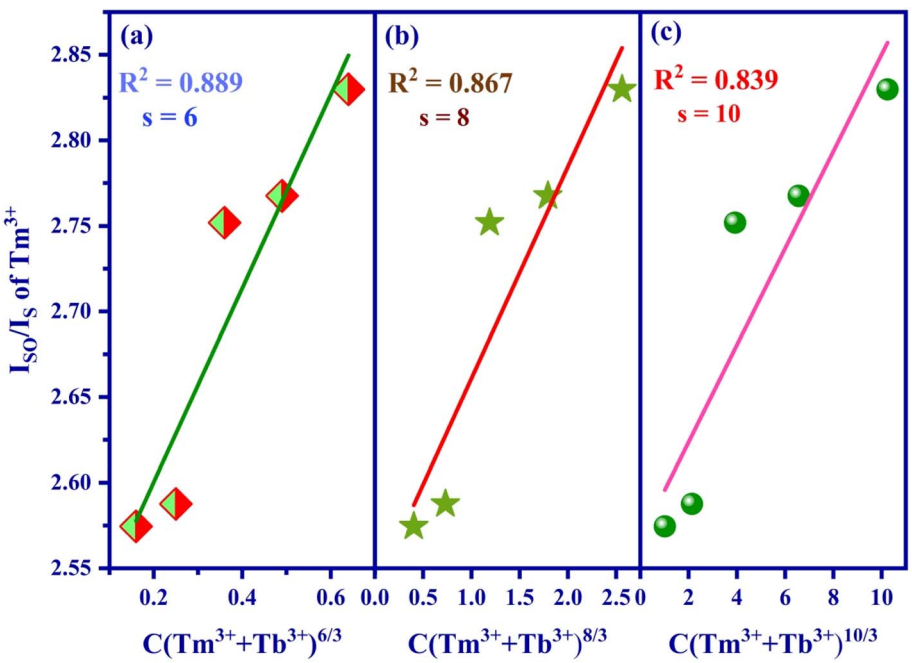


Fig. 12 Fitting relationship of I_{so}/I_s of Tm^{3+} on (a) $C\,(Tm^{3+} + Tb^{3+})^{6/3}$, (b) $C\,(Tm^{3+} + Tb^{3+})^{8/3}$ and (c) $C\,(Tm^{3+} + Tb^{3+})^{10/3}$.

© 2023 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

$n = 6$. Consequently, the dominant mechanism for energy transfer from Tm^{3+} to Tb^{3+} is ascribed to dipole–dipole interactions.

To provide additional evidence for $Tm^{3+} \rightarrow Tb^{3+}$ energy transfer in SNB phosphor. Theoretically, it is known that when 3 mol% Tm^{3+} 6 mol% $Tb^{3+} \rightarrow z$ Eu^{3+} energy transfers, the time attenuation of Tb^{3+} ions will be accelerated energy transfer process 3 mol% Tm^{3+} 6 mol% $Tb^{3+} \rightarrow z$ Eu^{3+}. If 3 mol% Tm^{3+} 6 mol% Tb^{3+} and Eu^{3+} ions work independently without occurrence of 3 mol% Tm^{3+} 6 mol% $Tb^{3+} \rightarrow z$ Eu^{3+} energy transfers, the fluorescence lifetime of both remain equivalent to that observed in the single-doped sample. The experimental analysis of fluorescence lifetime of Eu^{3+} ions in SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+}, z Eu^{3+} phosphor was tested and the double exponential decay function was used to fit well:[39,40]

$$I(t) = I_0 + A_1 \exp\left(-\frac{t}{\tau_1}\right) + A_2 \exp\left(-\frac{t}{\tau_2}\right) \quad (5)$$

In eqn (5), I_t represents the luminous intensity at specific time t, I_0 is the background or detector zero offset. A_1 and A_2 are constants while τ_1 and τ_2 are the rapid and gradual life of exponential components, respectively. The computation of the

Table 1 Colorimetric parameters for SNB:$x$$Tm^{3+}$, SNB:3$Tm^{3+}$/$y$$Tb^{3+}$ and SNB:3Tm^{3+}/6Tb^{3+}/$z$$Eu^{3+}$ samples with excitation of 350 nm

	Color co-ordinates			
Sample	X	Y	CRI	CCT (K)
T1	0.288	0.172	—	6500
T3	0.271	0.157	—	3032
T5	0.279	0.163	—	3616
T7	0.280	0.165	—	4337
T10	0.293	0.176	—	7858
TT1	0.282	0.289	88.82	9677
TT2	0.293	0.303	89.41	8163
TT3	0.303	0.313	89.51	7256
TT4	0.302	0.314	89.52	7309
TT5	0.304	0.315	89.53	7165
TT6	0.298	0.309	89.48	7622
TT8	0.306	0.316	89.55	6984
TT10	0.302	0.312	89.45	7291
TTE1	0.299	0.307	89.34	7623
TTE3	0.298	0.303	88.87	7815
TTE5	0.301	0.307	89.11	7467
TTE7	0.299	0.303	88.64	7662

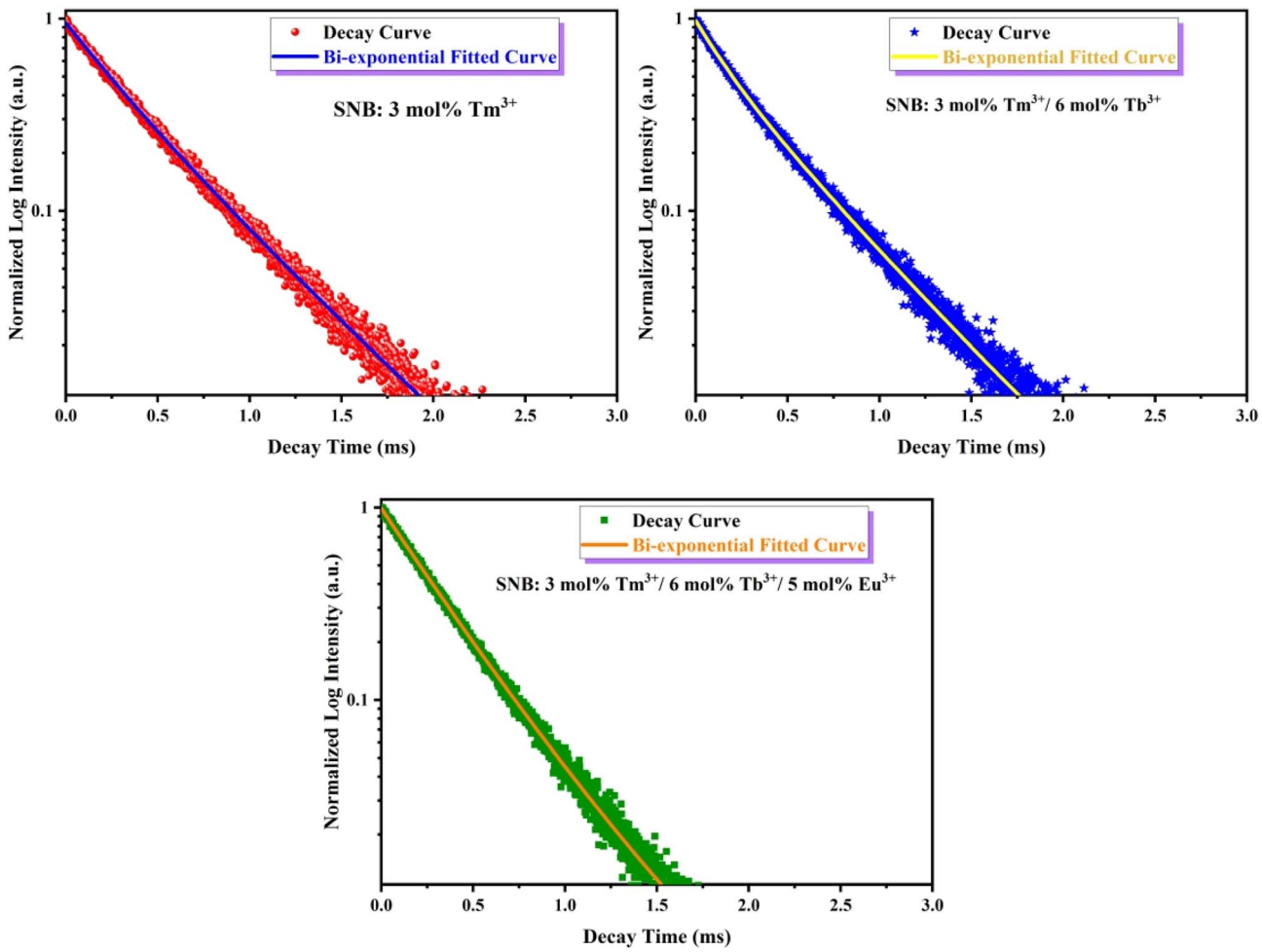


Fig. 13 PL decay curve for SNB:3 mol% Tm^{3+}, SNB:3 mol% Tm^{3+}/6 mol% Tb^{3+} and SNB:3 mol% Tm^{3+}/6 mol% Tb^{3+}/5 mol% Eu^{3+} phosphors.

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

average fluorescence lifetime (τ_{av}) can be performed using eqn (6):[41]

$$\tau_{avg} = \frac{A_1\tau_1^2 + A_2\tau_2^2}{A_1\tau_1 + A_2\tau_2} \quad (6)$$

The fluorescence lifetime decay curve of SNB:3 mol% Tm^{3+}, SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+} and SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+}, z Eu^{3+} phosphors is shown in Fig. 13 and the calculated τ_{av} value are 2.11, 1.90 and 1.64 ms, respectively. From the result it is evident that the decrement in the average life time of co doping sample SNB:3 mol% Tm^{3+}, 6 mol% Tb^3 is due to energy transfer between Tm^{3+} to Tb^{3+}. Further, the average lifetime decreased further when it is doped with Eu^{3+}, thus powerfully confirming the existence of 3 mol% Tm^{3+} 6 mol% Tb^{3+} → z Eu^{3+}.

The CIE chromaticity coordinates of SNB:$x$$Tm^{3+}$, SNB:3 mol% Tm^{3+}, $y$$Tb^{3+}$ and SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+}, $z$$Eu^{3+}$ phosphors are evaluated using formulation from the reported work in literature[42,43]

$$CCT = -449n^3 + 3525n^2 - 6823.3n + 5520.33 \quad (7)$$

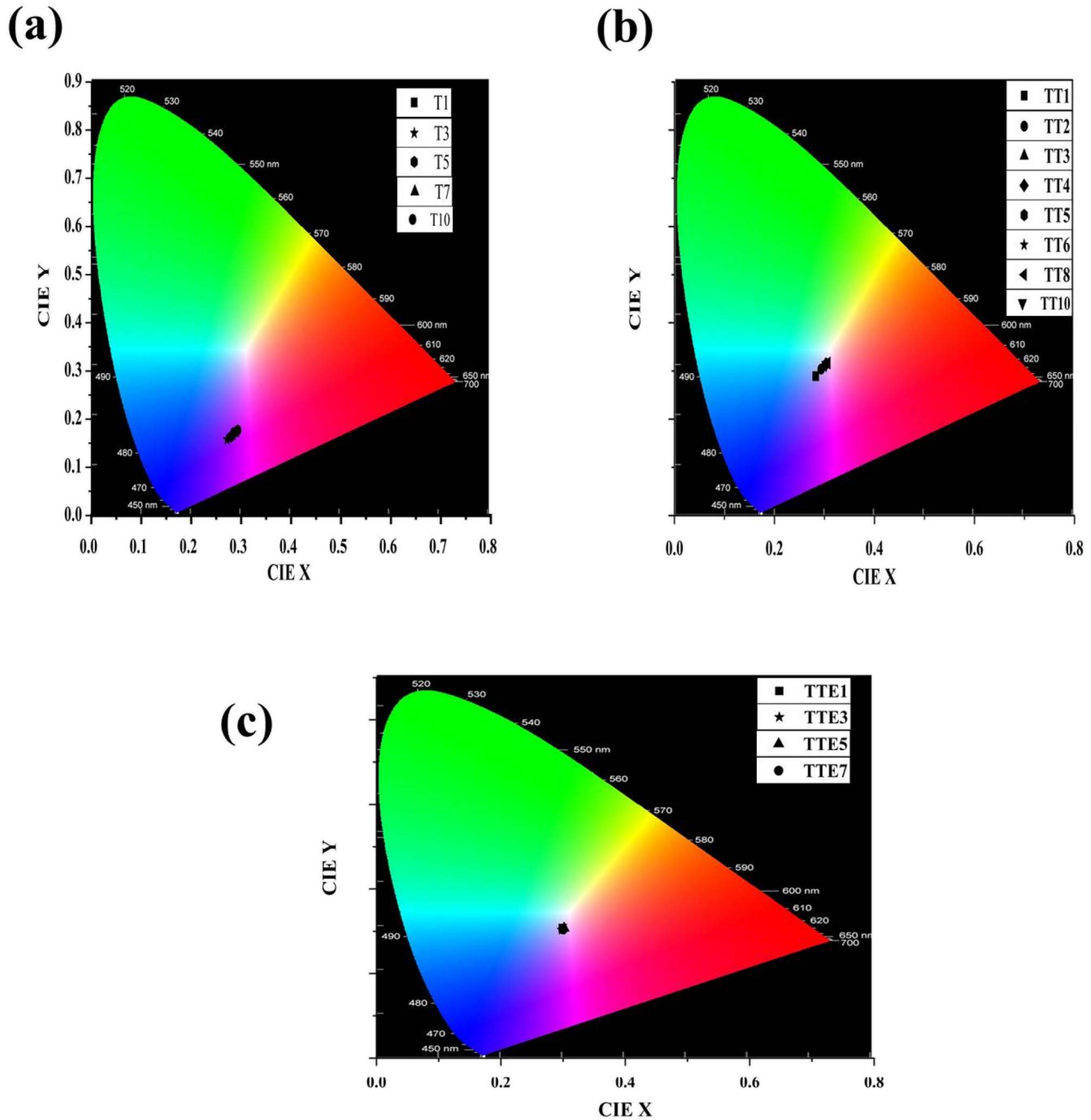


Fig. 14 CIE coordinates of (a) SNB:$x$$Tm^{3+}$, (b) SNB:3$Tm^{3+}$/$y$$Tb^{3+}$ and (c) SNB:3Tm^{3+}/6Tb^{3+}/$z$$Eu^3$ samples.

© 2023 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

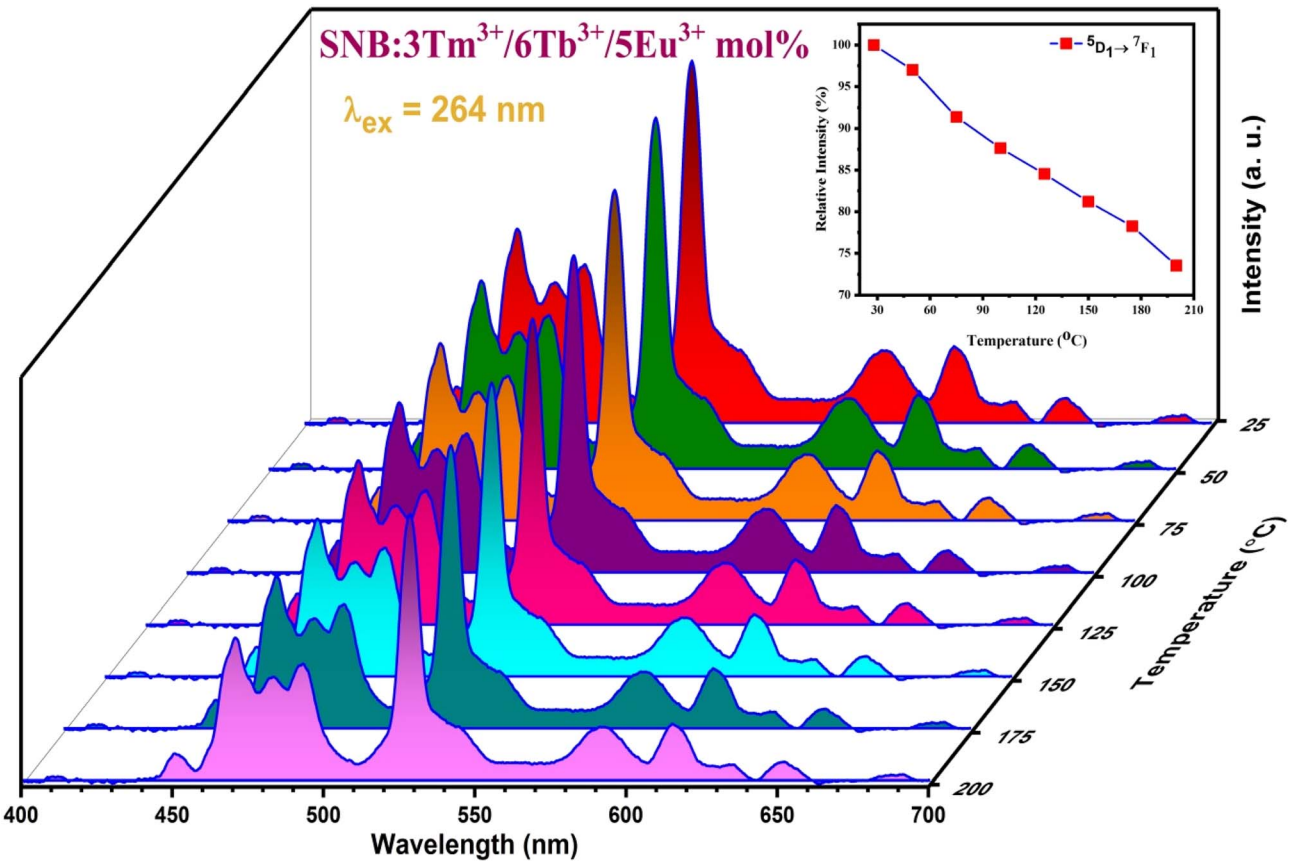


Fig. 15 TDPL spectra of SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+}, 5 mol% Eu^{3+} phosphor [inset exhibits the relative emission intensity variation for $^5D_1 \rightarrow {}^7F_1$ transition in the temperature range spanning from 28 to 200 °C].

and indicated in Table 1 and Fig. 14. These phosphors offer a remarkable avenue for color tunability, enabling the creation of customized illumination. The warm light can be achieved by adjusting the CCT and CRI values with NUV excitation of SNB:0.03Tm^{3+}, 0.06Tb^{3+}, zEu^{3+} phosphors. Notably, the CIE coordinates of TTE5 phosphor (0.3016, 0.3073) is close to standard value (0.31006, 0.3136). As highlighted in Table 1, increasing the concentration of Eu^{3+} concentration in the codoped phosphors has the effect of modifying the color temperature and enhancing the color rendering index by introducing the red emission into the spectra of white light. TTE1 phosphor represents the CCT of 7623 K and CRI 89.34 while the TTE3 phosphor emits the 7815 K CCT and 88.87 CRI white light. The result reveals that the TTE5 phosphor with higher Eu^{3+} concentration emits the cool light with CCT value of 7467 K and CRI value of 89.12. Hence, the white light with alterable CCT (7467–7815 K) and high CRI (>88) can be achieved for SNB:0.03Tm^{3+}, 0.06Tb^{3+}, zEu^{3+} phosphors with the control of Eu^{3+} concentration. Obviously, the TTE5 phosphor yield neutral white light with CCT of 7467 K and CRI of 89.12 with CIE coordinates (0.3016, 0.3073). These attributes position these phosphors as a strong contender for single-phased phosphors, holding significant promise for advancing neutral white light-emitting diodes (w-LEDs) with tunable color characteristics.

The thermal stability of as-prepared phosphors is crucial parameter in the application of W-LEDs device as operating temperature may increase significantly after the constant working of device at high temperature. The temperature dependent photoluminescence of SNB:3 mol% Tm^{3+}, 6 mol% Tb^{3+}, 5 mol% Eu^{3+} phosphor under 263 nm excitation was conducted in the range spanning from room temperature (28 °C) to 200 °C is shown in Fig. 15. The results indicate that the emission intensities of the as-prepared sample decreases with the rise in temperature because of thermal quenching but position remains unaltered. This decrement is attributed to the non-radiative transitions, therefore rate of intensity reduction changes over different emission peaks. This characteristic maintains the color stability of the as-prepared phosphors which indicates their reliable performance. Thermal stability was determined through the ratio of emission intensity at particular temperature to that at room temperature. The emission trend of the transition $^5D_1 \rightarrow {}^7F_1$ is depicted in the inset of Fig. 15. The strength of intensity at 150 °C and 200 °C relative to that at room temperature comes out to be 81.19% and 73.51%, respectively which is higher than reported work in the literature.[28] The results indicate the excellent thermal stability of the as-prepared phosphor.

3.7. Conclusions

We have synthesized a series of single-phase SNB phosphors co-doped with Tm^{3+}, Tb^{3+}, and Eu^{3+} ions. XRD analysis of the SNB phosphor confirms the formation of cubic structure within the

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Fm3m space group and its match with standard data from the JCPDS file. The quantum efficiency of TT6 phosphor found to be 91.3%. The TDPL spectra reveal that the PL intensity persists to 81.19% at 150 °C in compare to room temperature. Under an excitation wavelength of 360 nm, Tm^{3+} single-doped SNB samples exhibit blue luminescence, characterized by an initial increase followed by a subsequent decrease in luminescence intensity. The maximum intensity is reached at a 3 mol% doping concentration. Additionally, the photoluminescence emission spectra of SNB phosphors co-doped with Tm^{3+} and Tb^{3+} ions, excited at 263 nm, reveal a declining intensity of Tm^{3+} emissions alongside an increasing intensity of Tb^{3+} emissions. This shift in the behavior of Tm^{3+} and Tb^{3+} ion intensities signifies the occurrence of energy transfer from Tm^{3+} to Tb^{3+} ions. Upon doping Eu^{3+} ions into the optimized co-doped sample (3 mol% Tm^{3+} and 6 mol% Tb^{3+}), both Tb^{3+} and Tm^{3+} ion intensities decrease as the concentration of Eu^{3+} ions increases. This variation in intensity is attributed to the energy transfer process from the optimized co-doped sample to Eu^{3+} ions. By carefully adjusting the doping concentrations of Tm^{3+}, Tb^{3+}, and Eu^{3+} ions to approximately 3 mol%, 6 mol%, and 5 mol%, respectively, the emission color can be fine-tuned to achieve white light emission.

Author contribution

Ravina: conceptualization, methodology, writing – original draft. Kanishk Poria: conceptualization, methodology, writing – original draft. Mukesh K Sahu: data curation, writing – review & editing. A. Kumar: formal analysis. Anu: formal analysis. Sajjan Dahiya: resources, formal analysis. Nisha Deopa: supervision, methodology, software, validation, writing – review & editing formal analysis. A. S. Rao: writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

One of the authors, 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) to her.

References

1 J. Zheng, X. Wu, Q. Ren, W. Bai, Y. Ren, M. Wang and O. u. Hai, Investigation of luminescence properties and energy transfer in Sm^{3+} and Eu^{3+} co-doped $Sr_3Y(BO_3)_3$ red phosphors, *Opt. Laser Technol.*, 2020, **122**, 105857.

2 W. Zhang, H. Shen, X. Hu, Y. Wang, J. Li, Z. J. Zhu, Z. You and C. Tu, Solid-state synthesis, structure and spectroscopic analysis of Dy:$CaYAl_3O_7$ phosphors, *J. Alloys Compd.*, 2019, **781**, 255–260.

3 Q. Ren, Y. Zhao, X. Wu and O. Hai, Tb^{3+}/Eu^{3+}/Tm^{3+} co-doped single-phase phosphors emitting close to standard white light, *Opt. Laser Technol.*, 2021, **136**, 106774.

4 H. Patnam, S. K. Hussain, L. K. Bharat and J. S. Yu, Synthesis and luminescence properties of Eu^{3+} and Dy^{3+} ions single and co-doped $Ba_2LaV_3O_{11}$ phosphors for white-light applications, *Dyes Pigm.*, 2019, **162**, 583–589.

5 D. Rajesh, K. Brahmachary, Y. C. Ratnakaram, N. Kiran, A. P. Baker and G.-G. Wang, Energy transfer based emission analysis of Dy $^{3+}$/Eu $^{3+}$ co-doped ZANP glasses for white LED applications, *J. Alloys Compd.*, 2015, **646**, 1096–1103.

6 J. Li, H. Yan and F. Yan, A novel high color purity blue-emitting phosphor: $CaBi_2B_2O_7$: Tm^{3+}, *Mater. Sci. Eng., B*, 2016, **209**, 56–59.

7 T. S. Atabaev, H. H. Thi Vu, H. K. Kim and Y. H. Hwang, The optical properties of Eu^{3+} and Tm^{3+} codoped Y_2O_3 submicron particles, *J. Alloys Compd.*, 2012, **525**, 8–13.

8 G. Zhu, Z. P. Ci, Y. R. Shi, M. D. Que, Q. Wang and Y. H Wang, Synthesis, Crystal Structure and Luminescence Characteristics of a Novel Red Phosphor $Ca_{19}Mg_2(PO_4)_{14}$:Eu^{3+} for Light Emitting Diodes and Field Emission Displays, *J. Mater. Chem. C*, 2013, **1**, 5960–5969.

9 A. G. Bispo Jr, S. A. M. Lima and A. M. Pires, Energy transfer between terbium and europium ions in barium orthosilicate phosphors obtained from sol-gel route, *J. Lumin.*, 2018, **199**, 372–378.

10 P. C. Ricci, M. Salis, R. Corpino, C. M. Carbonaro, E. Fortin and A. Anedda, A Kinetics Model for Tb^{3+} Recombinations in Low Doped Tb:$Lu_{1.8}Y_{0.2}SiO_5$ Crystals, *J. Appl. Phys.*, 2010, **108**, 043512.

11 W. Zhao, X. Feng and B. Fan, Novel color tunable phosphors $NaYGeO_4$: Tm^{3+},Tb^{3+},Eu^{3+} for ultraviolet excited white LEDs with good thermal stability, *J. Mater. Sci.: Mater. Electron.*, 2020, **31**, 14434–14442.

12 V. V. Atuchin, B. I. Kidyarov and N. V. Pervukhina, Shortest chemical bond length as a criterion for searching new noncentrosymmetric niobate and tantalite crystals with high optical nonlinearity, *J. Cryst. Growth*, 2005, **275**, 1941–1946.

13 V. V. Atuchin, J. C. Grivel, A. S. Korotkov and Z. Zhang, Electronic parameters of $Sr_2Nb_2O_7$ and chemical bonding, *J. Solid State Chem.*, 2008, **181**, 1285–1291.

14 A. K. Vishwakarma, K. Jha, M. Jayasimhadri, A. S. Rao, K. Jang, B. Sivaiah and D. Haranath, Red light emitting $BaNb_2O_6$:Eu^{3+} phosphor for solid state lighting applications, *J. Alloys Compd.*, 2015, **622**, 97–101.

15 R. Lohan, A. Kumar, M. K. Sahu, A. Mor, V. Kumar, N. Deopa and A. S. Rao, Structural, thermal, and luminescence kinetics of $Sr_4Nb_2O_9$ phosphor doped with Dy^{3+} ions for cool w-LED applications, *J. Mater. Sci.: Mater. Electron.*, 2023, **34**, 694.

16 L. Alexander and H. P. Klug, Determination of crystallite size with the x-ray spectrometer, *J. Appl. Phys.*, 1950, **21**, 137–142.

17 M. Sharma, D. Kumar, J. P. Gewali and A. Thakur, Yttrium doped strontium nano hexaferrites: Influence of yttrium on the structural and magnetic properties of sol-gel synthesized nanocrystalline strontium hexaferrite, *Int. J. Mod. Phys.: Conf. Ser.*, 2022, **2267**, 012024.

© 2023 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 16 November 2023. Downloaded on 12/7/2023 6:09:41 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

18 K. Poria, R. Lohan, S. Bhatia, A. Kumar, R. Singh, N. Deopa, R. Punia, J. S. Shahi and A. S. Rao, Lumino-structural properties of Dy^{3+} activated $Na_3Ba_2LaNb_{10}O_{30}$ phosphors with enhanced internal quantum yield for w-LEDs, *RSC Adv.*, 2023, **13**(17), 11557–11568.

19 F. A. Kerim, H. F. Aly and A. Agramy, Infrared absorption spectra of some lanthanide acetylacetonate complexes, *Proc. Indian Acad. Sci.*, 1977, **85**, 559–566.

20 M. Afqir, A. Tachafine, D. Fasquelle, M. Elaatmani, J. C. Carru, A. Zegzouti, M. Daoud, S. Sayouri, T. D. Lamcharfi and M. Zouhairi, Structural, electric and dielectric properties of Eu-doped $SrBi_2Nb_2O_9$ ceramics obtained by co-precipitation route, Process, *Appl. Ceram.*, 2012, **12**, 72–77.

21 I. P. Sahu, D. P. Bisen, N. Brahme and R. K. Tamrakar, Generation of white light from dysprosium-doped strontium aluminate phosphor by a solid-state reaction method, *J. Mater. Sci.: Mater. Electron.*, 2016, **45**, 2222–2232.

22 W. T. Carnall, P. R. Fields and K. Rajnak, Electronic Energy Levels in the Trivalent Lanthanide Aquo Ions. I. Pr^{3+}, Nd^{3+}, Pm^{3+}, Sm^{3+}, Dy^{3+}, Ho^{3+}, Er^{3+}, and Tm^{3+}, *J. Chem. Phys.*, 1968, **49**, 4424.

23 W. T. Carnall, P. R. Fields and K. Rajnak, Electronic Energy Levels of the Trivalent Lanthanide Aquo Ions. III. Tb^{3+}, *J. Chem. Phys.*, 1968, **49**, 4447.

24 M. Wang, Y. Zhang, Z. Hu and J. Zhang, A single-phased $Ca_4(PO_4)_2O$: Tm^{3+}, Tb^{3+}, Eu^{2+} white light-emitting phosphor with tunable emission color, *J. Lumin.*, 2020, **224**, 117301.

25 W. Xiulan, D. Liang, Y. Zheng, M. Pei, Q. Ren and O. Hai, $Tm^{3+} \rightarrow Tb^{3+}$ energy transfer induced color-tunable in double-doped $LiLaSiO_4$ phosphors, *J. Lumin.*, 2021, **235**, 118027.

26 N. Deopa and A. S. Rao, Spectroscopic studies of single near ultraviolet pumped Tb^{3+} doped Lithium Lead Alumino Borate glasses for green lasers and tricolour w-LEDs, *J. Lumin.*, 2018, **194**, 56–63.

27 R. Pillania, A. Prasad, P. Rohilla, A. Shandilya and A. S. Rao, Photoluminescence and energy transfer studies on Tm^{3+}/Dy^{3+}/Eu^{3+}doped borosilicate glasses for color tunability and warm white light generation, *J. Non-Cryst. Solids*, 2023, **606**, 122192.

28 X. Wu, W. Bai, O. Hai, Q. Ren, J. Zheng and Y. Ren, Tunable color of Tb^{3+}/Eu^{3+}/Tm^{3+}-coactivated $K_3La(PO_4)_2$ via energy transfer: A single-phase white-emitting phosphor, *Opt. Laser Technol.*, 2019, **115**, 176–185.

29 X. Wu, L. Du, Y. Zheng, M. Pei, Q. Ren and O. Hai, $Tm^{3+} \rightarrow Tb^{3+}$ energy transfer induced color-tunable in double-doped $LiLaSiO_4$ phosphors, *J. Lumin.*, 2021, **235**, 118027.

30 X. Liu, C. Chen, S. Li, Y. Dai, H. Guo, X. Tang, Y. Xie and L. Yan, Host-sensitized and tunable luminescence of $GdNbO_4$: Ln^{3+} (Ln^{3+}= Eu^{3+}/Tb^{3+}/Tm^{3+}) nanocrystalline phosphors with abundant color, *Inorg. Chem.*, 2016, **55**, 10383–10396.

31 Y. Chen, X. Liu, G. Chen, T. Yang and C. Yuan, Luminescent characteristics of Tm^{3+}/Tb^{3+}/Eu^{3+} tri-doped borophosphate glasses for LED applications, *J. Mater. Sci.: Mater. Electron.*, 2017, **28**, 5592–5596.

32 Z. Wang, H. Tang, J. Xie, X. Zhang, X. Liu, X. Mi, Q. Liu, L. Lu and Z. Bai, Tunable color and energy transfer of $Y_2Mg_3Ge_3O_{12}$: Tb^{3+}, Tm^{3+}, Eu^{3+} phosphors with excellent thermal stability, *J. Solid State Chem.*, 2022, **314**, 123371.

33 H. Thakur, A. K. Gathania, I. Kumar and R. K. Singh, Tunable white light photoluminescence of a single phase Tm^{3+}/Tb^{3+}/Eu^{3+} codoped $GdPO_4$ phosphor, *Mater. Today Commun.*, 2023, **36**, 106458.

34 G. Blasse and B. C. Grabmaier, *Energy Transfer in Luminescent Materials*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1994, pp. 91–107.

35 G. Blasse, Energy transfer in oxidic phosphors, *Philips Res. Rep.*, 1969, **24**, 131–144.

36 K. Li, M. Shang, H. Lian and J. Lin, Recent development in phosphors with different emitting colors via energy transfer, *J. Mater. Chem. C*, 2016, **24**, 5507–5530.

37 J. Xiong, Q. Meng and W. Sun, Luminescent properties and energy transfer mechanism from Tb^{3+} to Eu^{3+} in $CaMoO_4$:Tb^{3+},Eu^{3+} phosphors, *J. Rare Earths*, 2016, **34**, 251–258.

38 D. L. Dexter, A theory of sensitized luminescence in solids, *J. Chem. Phys.*, 1953, **21**, 836–850.

39 C. H. Huang, T.-W. Kuo and T.-M. Chen, Novel Red-Emitting Phosphor $Ca_9Y(PO_4)_7$:Ce^{3+},Mn^{2+} with Energy Transfer for Fluorescent Lamp Application, *ACS Appl. Mater. Interfaces*, 2010, **2**, 1395–1399.

40 N. Deopa, M. K. Sahu, P. R. Rani, R. Punia and A. S. Rao, Realization of warm white light and energy transfer studies of Dy^{3+}/Eu^{3+} co-doped Li_2O-PbO-Al_2O_3-B_2O_3 glasses for lighting applications, *J. Lumin.*, 2020, **222**, 117166.

41 Y. Zhu, Q. Meng, W. Sun and S. Lü, $NaLa(MoO_4)_2$: Sm^{3+}, Tb^{3+} phosphor: optical temperature sensing material with a wide change range of luminescence color, *J. Lumin.*, 2020, **218**, 116854.

42 A. Kumar, A. Mor, M. K. Sahu, R. Pillania, S. Dahiya, N. Deopa, A. Malik, R. Punia and A. S. Rao, Spectral characteristics of Tb^{3+} doped ZnF_2–K_2O–Al_2O_3–B_2O_3 glasses for epoxy free tricolor w-LEDs and visible green laser applications, *J. Lumin.*, 2022, **244**, 118676.

43 C. S. McCamy, Correlated color temperature as an explicit function of chromaticity coordinates, *Color Res. Appl.*, 1992, **17**, 142–144.

The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2755-0761.htm

Evaluating public service value within collaborative e-governance: a study in the Indian context

Jitendra Kumar Pandey
Delhi School of Management, Delhi Technological University, Delhi, India

Received 19 August 2023
Revised 23 September 2023
3 October 2023
Accepted 4 October 2023

Abstract

Purpose – This study aims to meticulously evaluate the public service value-generation process facilitated by collaborative e-governance services within the framework of the National e-governance Plan (NeGP).

Design/methodology/approach – The study formulates a comprehensive research model through a combination of literature review, insights from domain experts and hands-on experience gained from the e-governance project. A conceptual research model was meticulously structured, validated, and interpreted by using a reflective measurement theory. The analytical tool SmartPLS3 was used to assess the proposed model rigorously.

Findings – The analysis of collected data reveals a statistically significant positive correlation between the implementation of collaborative e-governance strategies and the creation of public service value. This relationship is further reinforced by a strong alignment between the perceived aspects of collaborative e-governance, such as responsiveness, transparency and service delivery and their substantial contribution to the enhancement of public service value.

Originality/value – This study contributes to the scholarly discourse by introducing an innovative methodology for assessing public service value through analyzing empirical data from citizen-centric collaborative e-governance projects. It is noteworthy that no prior studies have examined the nuanced concept of public service value in the context of collaborative e-governance.

Keywords Collaborative e-governance, E-governance impact on public value, Indian e-governance initiatives, National e-governance plan (NeGP), Public service value, Reflective measurement theory

Paper type Research paper

1. Introduction

Government organizations often work in silos, with independent processes, control structures and decision-making environments (McDermott, 2010; Suri, 2014). This can lead to inefficiencies, ineffective decision-making and a lack of trust and cooperation. Citizens, on the other hand, expect their problems to be addressed seamlessly, without having to navigate the long channels of a traditional government department (O'Leary & Vij, 2012). To meet these expectations, government organizations need to build cross-organizational strategic collaboration. Strategic collaboration can lead to several benefits for creating public service value, including Increased efficiency and effectiveness (Ansell & Gash, 2008), Improved decision-making (Suri, 2014) and Greater citizen satisfaction (Kumar, Singh, & Shankar, 2016). To be successful, strategic collaboration must be carefully planned and implemented with the integration of newer technologies.

© Jitendra Kumar Pandey. Published in *Digital Transformation and Society*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

The author duly acknowledges the support of Prof. (Dr.) PK Suri of DTU, Delhi, and Sh. PK Mohapatra, General Manager (IT), NHAI, Delhi for their insightful questionnaire reviews and generously sharing expertise and providing valuable research feedback. The author also acknowledges the anonymous supporters whose undisclosed assistance was pivotal in developing and completing this paper.

Funding: This research work was conducted without external financial support.

Digital Transformation and Society
Emerald Publishing Limited
e-ISSN: 2755-077X
p-ISSN: 2755-0761
DOI 10.1108/DTS-08-2023-0065

The integration of information and communication technology (ICT) into governance can significantly enhance government institutions' capabilities. Specifically, the use of collaborative governance mechanisms can extend the involvement of citizens in government decision-making, as noted by Lofstedt (2012) and Pandey and Suri (2020). A key aspect of collaborative governance is the adoption of good governance practices for the creation of public service value, which encompasses elements of transparency, citizen engagement, improved service delivery and streamlined administrative procedures (Deng, Karunasena, & Xu, 2018; Pandey & Suri, 2020). The effectiveness of intergovernmental collaboration in addressing complex challenges, such as resource allocation, technological interoperability and compliance with legal and regulatory models, is highlighted in the work of O'Leary and Vij (2012), Pardo, Nam, and Burke (2012) and Kumar *et al.* (2016). As a result, the combination of ICT and collaborative governance emerges as a transformative pathway for strengthening government functionality and responsiveness.

The National e-Governance Plan (NeGP) was introduced in India in 2006 to provide affordable, efficient, transparent and reliable government services. The plan of 2006 is referred to as NeGP 1.0 and it could partially remain successful as it had some limitations, such as a lack of comprehensive need analysis, integration of processes and collaboration across applications and databases. NeGP 2.0 or e-Kranti was launched in 2014 to address these limitations and to enhance public service value by improving governance efficiency and service delivery. Notable citizen-centric e-governance projects such as BHOOMI, CARD, e-SEVA, FRIENDS, GYANDOOT, LOKVANI, etc., have had a significant impact on citizens' lives. However, strategic collaboration within and across government agencies emerged as a key challenge to the full success of e-governance projects in India (Suri, 2016). Overcoming this challenge for bettering public service value is essential for the success of e-governance projects.

This research examines four collaborative e-governance initiatives in India that are aligned with the objectives of NeGP and were accessible to the researcher. Of these projects, three were affiliated with the Indian central government and one was associated with the state government of Delhi (the capital city of India). The projects were chosen based on the criteria that they have been in operation for at least five years, are citizen-centric and require citizens to visit government offices. The research also explores the potential for corruption in these initiatives drawing on the work of Sheikh and Mandelkern (2014) and the CVC (GOI) Annual Report (2020-2021). The following are the research objectives:

(1) To clarify and analyze the interdependencies between public service value components and collaborative e-governance.

(2) To propose an empirically validated research model for assessing public service value derived from collaborative e-governance.

The paper has a clear structure and is well-organized. The introduction provides an overview of the research, including the objectives and significance. The literature review section provides a comprehensive overview of the research topic, including the theoretical underpinnings, the development of the conceptual research model and the formulation of research hypotheses. The research methodology section describes a case study approach of selected projects for evaluation, a detailed description of the questionnaire development and sampling methods. The data analysis and results section presents the analysis of measurement and structural models along with the tested hypotheses results. The final section of the paper highlights the discussions, implications, limitations and scope for future research.

2. Literature review

E-governance has the potential to improve organizational efficiency and meet citizens' expectations, but its full potential is still not being realized. A comprehensive literature

review was conducted to explore the research topic, including the concepts of e-governance, strategic collaboration, public service value and good governance principles (Bannister & Connolly, 2014; Gupta & Suri, 2017; Soni, Dey, Anand, Malhotra, & Banwet, 2017).

2.1 Theoretical background

Several scholars have explored the relationship between collaborative e-governance and public service value, emphasizing its transformative potential. For example, Bannister and Connolly (2014) found that ICT can make government more open, transparent, participatory and responsive, which are all important public values. Austin (2010), Flak, Solli-Saether, and Straub (2015), and Gupta and Suri (2017) also found that collaborative e-governance can improve planning, decision-making and control processes, ultimately contributing to the creation of public service value. These perspectives align with the broader research trend in examining how collaborative approaches in e-governance can yield substantial benefits for public service delivery and citizen engagement.

Furthermore, investigations by Osman *et al.* (2014) and Gupta and Suri (2017) have highlighted that collaborative e-governance not only streamlines service delivery but also ensures accessible and simplified information dissemination, particularly benefiting marginalized individuals. These insights emphasize the practical implications of collaborative e-governance in addressing social inclusivity and equity, thereby contributing to the overall enhancement of public service value. Chun, Luna-Reyes, and Sandova (2012) and Gupta and Suri (2017) have additionally emphasized how collaborative e-governance aligns with domain-specific goals, including safety, security, privacy and responsiveness. Such alignment serves to fortify the theoretical underpinning that collaborative e-governance strategies can effectively address specific service-related challenges, contributing to the creation of public service value.

Moreover, the works of Reddick (2011), Suri (2014) and Picazo-Vela, Gutierrez-Martınez, Duhamel, Luna, and Luna-Reyes (2015) emphasize that collaborative e-governance fosters transparency and citizen participation, further enhancing the public service value proposition. This aligns with the theoretical model that effective collaboration in e-governance can lead to more accountable, citizen-centric and participatory governance, ultimately resulting in value creation for the public.

2.1.1 E-governance. The integration of ICTs into government has transformed the way that information is disseminated and services are delivered. This has led to improvements in transparency, accountability and responsiveness, which have enhanced the value of public services (Tripathi, Gupta, & Bhattacharya, 2012). E-governance offers several advantages, such as high-quality services, enhanced accessibility, improved information security, cost reduction and increased citizen participation (Bannister & Connolly, 2014; Soni, Dey *et al.*, 2017). However, the predominant technical focus of e-governance can sometimes overshadow citizens' needs (Chun *et al.*, 2012). Additionally, there are some challenges to e-governance in developing countries, such as inadequate ICT infrastructure, disparities between design and implementation, institutional support, content localization and the digital divide (Luna-Reyes & Gill-Gracia, 2011; Gómez & Heeks, 2016; Soni, Anand, Dey, Dash, & Banwet, 2017). These challenges require aligning expectations with limited resources and advocating for strategic collaboration (Pandey & Suri, 2020).

2.1.2 Strategic collaboration. Notably, governmental entities aspire to optimize online service delivery via single-window systems. However, achieving this goal requires seamless coordination and integration across horizontal and vertical dimensions (Pardo *et al.*, 2012). While strategic collaboration can take various forms (Ansell & Gash, 2008), its significance in the e-governance domain transcends conceptual boundaries. It serves as a mechanism for enhancing citizen engagement, generating public service value and effectively achieving governance objectives (Austin, 2010; Tripathi *et al.*, 2012; United Nations Department of

Economic and Social Affairs, 2016). Yet, pursuing strategic collaboration faces obstacles such as organizational culture differences, conflicting goals, financial constraints and interoperability challenges (Tripathi *et al.*, 2012; Pham & Tanner, 2014). Shifting toward a collaborative governance paradigm becomes essential for successful initiative execution (Bertot, Jaeger, & Grimes, 2012). Thus, fostering a collaborative mindset within government agencies and providing essential competencies becomes crucial (McDermott, 2010), thereby enhancing the process of collaborative e-governance.

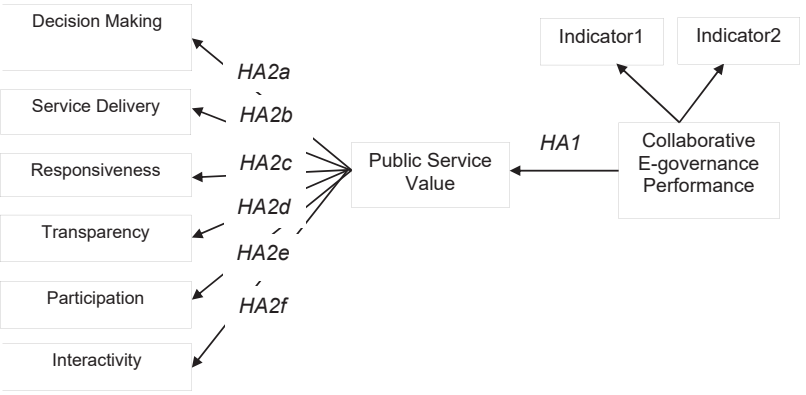
2.1.3 Collaborative e-governance performance. Collaborative e-governance is a combination of strategic collaboration and electronic governance that is powered by economic incentives, citizen engagement, technological advancements and the pursuit of public service value (Chun *et al.*, 2012; Gupta & Suri, 2017; Pandey & Suri, 2020). It has the potential to enhance public service value by strengthening processes like planning, executing and evaluating (Gupta & Suri, 2017). The assessment of collaborative e-governance performance is based on the degree of advancement of the system, which is characterized by positional advantage and relationship advantage (Dyer & Singh, 1998). Positional advantage refers to the ability to provide enhanced functionality and efficient service delivery, while relationship advantage refers to the ability to actively involve citizens in the decision-making process. The measurement of outcomes in collaborative e-governance is rooted in the cultivation of public service values (Kelly, Mulgan, & Muers, 2002; Kernaghan, 2003; Moore, 2014).

2.1.4 Public service value. Public value and public service value are interrelated concepts that have distinct meanings and implications in public administration and governance. Public value refers to the broader societal benefits and outcomes pursued by public organizations for citizens and the community (Moore, 1995; Kelly *et al.*, 2002; O'Cass & Ngo, 2011; Luna-Reyes, Picazo-Vela, Luna, & Gil-Garcia, 2016), while public service value refers to the principles and beliefs that guide public policy and the delivery of public services. Public service value is a more specific and tangible facet of public value, focusing on the quality, effectiveness and efficiency of government-provided services to the public (Golubeva, 2007; Karunasena & Deng, 2012). It is often evaluated in terms of customer satisfaction, accessibility, affordability and contribution to public well-being (Chun *et al.*, 2012; Cordella & Bonina, 2012; Gupta & Suri, 2017). A spectrum of public service values, including ethical, democratic, professional and people-oriented values, has been systematically categorized by Kernaghan (2003) and Bannister and Connolly (2014) taken as the guiding principle for this research.

Within the scope of this research, public service value denotes the specific value derived from the quality and delivery of public services to citizens by government organizations (Bryson, Crosby, & Bloomberg, 2014; Luna-Reyes *et al.*, 2016). This study adopts constructs of public service values that encapsulate the value intended for citizens, shaped by government efforts in collaborative e-governance (Cordella & Bonina, 2012; Bannister & Connolly, 2014; Gupta & Suri, 2017). The Public Service Value for the study includes effective planning, decision-making and control (Gupta & Suri, 2017), improved service delivery with accessible and understandable information and benefiting marginalized and disabled populations (Chun *et al.*, 2012; Flak *et al.*, 2015). Additionally, it fosters transparency by tracking and displaying information, involving citizens in policy-making and facilitating seamless government-citizen interactions (Ramaswamy & Ozcan, 2018; Green & Sergeeva, 2019; Pereira, Lohmann, & Houghton, 2021).

2.2 Hypotheses development

The study adopted variables from a literature review, expert insights and project experience. These variables were relevant to collaborative e-governance and public service value. A conceptual research model and hypotheses were constructed based on the variables. Figure 1



Source(s): Figure by authors

Figure 1. Research conceptual model

shows the interrelations between public service value and collaborative e-governance performance.

The theoretical underpinning guided by the "reflective measurement theory" elucidated by Hair, Ringle, and Sarstedt (2011), was instrumental in developing a conceptual research model and formulating hypotheses to assess the public service value engendered by collaborative e-governance. This theory posits that latent factors "cause" or "reflect" the indicative variable, resulting in a hierarchical arrangement. The model comprises six first-order factors (micro variables), a second-order factor (macro variable) and a third-order factor (outcome variable). The reflective-reflective model underwent validation in stages and is depicted in Figure 4.

Within the realm of collaborative e-governance performance, Indicator1 and Indicator2 are the two key factors that contribute to the overall success of collaborative e-governance initiatives, aligned with the process of public service value creation. To be more specific, Indicator1 measures the extent to which e-governance services are aligned with the needs of citizens. Meanwhile, Indicator2 measures the extent to which e-governance services are comprehensive and effective in addressing the concerns of citizens.

The proposed model posits that Public Service Value (VCEG), constitutes Collaborative E-governance Performance (CEGP) comprising six fundamental elements essential for the generation of public service value: Decision Making (VCDM), Service Delivery (VCSD), Responsiveness (VCRP), Transparency (VCTP), Participation (VCPT) and Interactivity (VCIT). To validate this proposed model, two sets of hypotheses were constructed-one for the main dimension and the other for its sub-dimensions.

Enhanced functionality, efficient service delivery and citizen participation in decision-making have a positive effect on public service value. It can, therefore, be argued that the existence of strong public service value substantially shapes and contributes to the realization of collaborative e-governance performance.

HA1. Public Service Value is a significant constituent of Collaborative E-governance Performance

Additionally, the interrelationships among the six micro variables within Public Service Value have also been subject to hypothesis formulation.

Decision Making: The augmentation of public service value within a governmental entity encompasses the expeditious resolution of citizens' concerns, the enactment of well-informed decisions, the vigilant oversight of procedural activities and the execution of efficacious

regulatory measures aimed at achieving the desired outcomes (Suri, 2014; Green & Sergeeva, 2019; Pereira *et al.*, 2021). The augmentation of public service value is a complex undertaking, but it is essential for the effective functioning of a democracy.

HA2a. Decision-making is a significant element of Public Service Value.

Service Delivery: The optimization of citizen service delivery for public service value creation is enhanced through the establishment of accessible and up-to-date information on processes and procedures via the designated online portal. This includes multi-language content ensuring its availability through common service centres. Moreover, the facilitation of accessibility for marginalized and disabled citizens should be ensured (Cordella & Bonina, 2012; Osman *et al.*, 2014; Gupta & Suri, 2017; Deng *et al.*, 2018).

HA2b. Service Delivery is a significant element of Public Service Value.

Responsiveness: Issuing receipts and token numbers to citizens when they submit applications and promptly handling inquiries, grievances recommendations, are proactive ways to provide public service values (Satapathy, 2014; Gupta & Suri, 2017). These proactive practices help to prevent problems from arising in the first place and improve the overall quality of public services.

HA2c. Responsiveness is a significant element of Public Service Value.

Transparency: The enrichment of transparency in e-governance through web-based functionalities that present comprehensive information about government initiatives, schemes, strategies, procedures the status of submitted applications of the citizens can enhance public service value (Suri, 2014; Deng *et al.*, 2018; Schoenmaker & Schramade, 2019). Such transparent activities also allow citizens to have a clear understanding of how public services are delivered and how their paid taxes are being used.

HA2d. Transparency is a significant element of Public Service Value.

Participation: The active involvement of citizens plays a critical role in enhancing the overall value of public service delivery, particularly through the model of collaborative e-governance. This entails the integration of citizens' constructive contributions, including their suggestions and remarks, into the processes of policy conceptualization and decision-making (Reddick, 2011; Chun *et al.*, 2012; Bryson *et al.*, 2014; Green & Sergeeva, 2019).

HA2e. Participation is a significant element of Public Service Value.

Interactivity: The augmentation of public service value through collaborative e-governance is facilitated by the seamless engagement of stakeholders, both within and outside government entities. This requires government organizations to accord high priority to stakeholder engagement. Interactivity, a pivotal dimension in this context, entails the utilization of a range of communication channels, including but not limited to social media platforms, call centres and organized events (Karunasena & Deng, 2012; Suri, 2014, 2016; Ramaswamy & Ozcan, 2018).

HA2f. Interactivity is a significant element of Public Service Value.

2.3 Concise project descriptions for evaluation

The study defined selection criteria for e-governance projects, which involved on-site visits to government offices, collaborative methodologies, a track record of at least five years and the exclusion of early-stage projects. It encompassed projects at state, autonomous and national levels within the NeGP initiative, including those with documented corruption concerns. Table 1 presents succinct information on the projects selected for assessment.

Project name/Operating organization	Coverage	Focus	Implementation status
Passport SEWA project/ ministry of external affairs (MEA)	All India	To enhance public service values by providing passport-related services to the citizens in a collaborative manner with private agencies and several other government organizations. The government agencies look at on fundamental functions such as verification, authorization, and passport issuance, leaving technology-oriented services to the purview of the private company. This initiative holds the potential to substantially enhance the efficiency and efficacy of passport services within India	Operational since 2010
Leased hold to freehold conversion of properties/Delhi development authority (DDA)	All India	Enabling citizens to access services through a unified platform that embodies efficiency, transparency, reliability, and user-friendliness. This initiative engages multiple stakeholders to ensure interactive and timely service delivery, all within a streamlined timeframe of 45 days	Operational since 2014
Booking of community hall/open spaces and parks/DDA	Delhi	To encourage and foster interactions between citizens from diverse social and cultural backgrounds, prioritizing their collective well-being, the approach for reservations operates on a 'First Come First Serve (FCFS)' basis. This reservation system is inclusive, welcoming individuals of all religious affiliations, and facilitating the organization of various activities	Operational since 2012
Driving License/Government of National Capital Territories (NCT) of Delhi	Delhi	A collaborative e-governance projects engaged in the issuance of driving licenses across all types of motor vehicles, including personal, passenger, and commercial transport. This approach aims to improve efficiency by reducing human involvement and enhancing security, transparency, and overall performance	Operational since 2007

Source(s): Table by authors

Table 1.
Projects selected for assessment

3. Research methodology

In order to acquire data for this study, structured questionnaires were administered to individuals utilizing e-governance services. The research framework underwent validation and testing employing Partial Least Squares Structural Equation Modeling (PLS-SEM). PLS-SEM is primarily used in exploratory analysis, and it does not require the normal distribution of data and is suitable for a small size sample (Hair *et al.*, 2011) as is the case of this study.

The selection of the PLS-SEM approach in this research aligns with the recommendations set forth by Sarstedt, Ringle, and Hair (2017). This method is deemed suitable when (1) the research objective is to explain and predict target constructs and/or to detect important driver

constructs, (2) the structural model has reflectively measured constructs and (3) the researcher is working with small sample size. For data collection and model evaluation Smart PLS version 3.0 software has been employed.

3.1 Questionnaire development

The questionnaire was designed using a five-point Likert-type scale, allowing respondents to express their agreement on a scale of intensity from 1 (no extent) to 5 (very large extent). A closed-ended questionnaire was distributed between April 1, 2020 and December 31, 2020, during the Covid-19 pandemic. The initial version underwent pre-testing with a panel of 10 participants, including academic experts and e-governance practitioners, to refine the wording, order and structure of questions. Feedback refined the questionnaire, addressing potential biases, including non-response bias and Common Method Bias (CMB). The questionnaire was neutral to prevent personal opinions from influencing responses and double-barreled questions were avoided. Options were designed to align with the "interest hypothesis," which posits that respondents more interested in the subject matter would respond more promptly (Donald, 1960). The questionnaires, along with the constructs' are provided in Supplementary B (Supplementary file).

3.2 Sampling method and data collection

Survey administration began with the distribution of questionnaires to the initially identified beneficiaries in person. The beneficiaries were requested to distribute the surveys to individuals who had utilized the services of the selected projects under study.

3.2.1 Target respondents and sample size. The survey aimed for 300 participants but concluded with 250 responses, out of which 210 were suitable for analysis. This sample size of 210 was deemed adequate for the conceptual model analysis using PLS-SEM (Sarstedt *et al.*, 2017; Ringle, Sarstedt, Mitchell, & Gudergan, 2020). Data collection used both offline and online methods. In the offline approach, respondents were given physical questionnaires. For online, Google Forms were used and distributed via email, Facebook and WhatsApp.

3.2.2 Participant demographics and service usage. The study included respondents aged 18 to 55, with a small portion above 56 years (4.28%). Males constituted 77.14% of the sample, while females were 22.85%. Most had a graduation or higher degree (91.40%). The professional distribution was diverse, with government employees (50.4%), private employees (33.33%) and students (12.38%) being prominent. Regarding e-governance service use, 78.56% engaged with services like 'Passport Sewa' and 'Driving License', while 21.42% used 'Online Booking of DDA's Park and Open Spaces' and 'Lease-hold to Free-hold Conversion of Properties'. Participants' demographics and e-governance services used are given in Supplementary A.

4. Data analysis and results

Structural equation modeling (SEM) comprising measurement model and structural model was used to evaluate the reliability and validity of the constructs and the predictive capabilities of the target constructs. The results of the evaluation are presented below.

4.1 Measurement model analysis

The study assesses construct reliability using Cronbach's alpha ($\alpha \geq 0.70$) and Composite Reliability ($CR \geq 0.70$). Convergent validity is examined through Average Variance Extracted (AVE >0.50) and factor loadings (≥0.70). Discriminant validity is established using the Fornell-Larcker criterion, cross-loadings and HTMT ratio. According to the Fornell and Larcker (1981) criterion, AVE (>0.50) should exceed squared correlations between latent

variables, representing the square root of the construct's AVE. The measurement model analysis shows that the constructs' reliability and validity (Table 2) meet the recommended thresholds of all the constructs.

HTMT (Table 3) assesses the correlation between constructs, with a threshold of 0.90 at 95% confidence (Henseler, Ringle, & Sarstedt, 2015). A value over 0.90 indicates weak discriminant validity. In this study, all values are below 0.90, except "Service Delivery", which is 0.91, still within acceptable limits. Cross-loadings indicate construct items loading better within their constructs (Table 4).

4.2 Structural model analysis

The structural model's predictive capacity is assessed through collinearity, path coefficients, coefficients of determination (R^2), effect size (f2), blindfolding and predictive relevance (Q^2).

Constructs	Items	Factor loadings	α	CR	AVE	Fornnel Larcker criteria (SQRT AVE)
Decision making	DMC1	0.812	0.896	0.924	0.708	0.841
	DMC2	0.876				
	DMC3	0.888				
	DMC4	0.782				
	DMC5	0.843				
Interactivity	ITN1	0.789	0.818	0.880	0.647	0.804
	ITN2	0.805				
	ITN3	0.827				
	ITN4	0.795				
Participation	PTN1	0.915	0.912	0.945	0.851	0.922
	PTN2	0.943				
	PTN3	0.909				
Responsivity	RSP1	0.762	0.790	0.864	0.614	0.784
	RSP2	0.851				
	RSP3	0.772				
	RSP4	0.746				
Service delivery	SDL1	0.796	0.785	0.861	0.608	0.780
	SDL2	0.813				
	SDL3	0.733				
	SDL4	0.775				
Transparency	TRP1	0.849	0.806	0.873	0.633	0.795
	TRP2	0.762				
	TRP3	0.782				
	TRP4	0.785				

Source(s): Table by authors

Table 2. Constructs reliability and validity

	VCDM	VCIN	VCPT	VCRP	VCSD	VCTP
VCDM						
VCIN	0.49					
VCPT	0.29	0.74				
VCRP	0.89	0.74	0.40			
VCSD	0.89	0.70	0.53	0.90		
VCTP	0.83	0.76	0.60	0.92	0.91	

Source(s): Table by authors

Table 3. HTMT for the constructs (based on 0.90 criterion)

	VCDM	VCIN	VCPT	VCRP	VCSD	VCTP
DMC1	0.812	0.313	0.041	0.553	0.603	0.564
DMC2	0.876	0.356	0.122	0.631	0.648	0.638
DMC3	0.888	0.325	0.152	0.659	0.706	0.622
DMC4	0.782	0.396	0.273	0.654	0.673	0.586
DMC5	0.843	0.396	0.144	0.666	0.690	0.614
ITN1	0.379	0.789	0.526	0.493	0.437	0.475
ITN2	0.347	0.805	0.486	0.489	0.495	0.486
ITN3	0.335	0.827	0.539	0.492	0.379	0.484
ITN4	0.307	0.795	0.519	0.462	0.354	0.557
PTN1	0.155	0.594	0.915	0.363	0.222	0.388
PTN2	0.182	0.607	0.943	0.326	0.262	0.389
PTN3	0.152	0.578	0.909	0.294	0.172	0.353
RSP1	0.698	0.354	0.062	0.762	0.602	0.591
RSP2	0.640	0.524	0.427	0.851	0.595	0.674
RSP3	0.562	0.393	0.162	0.772	0.493	0.488
RSP4	0.462	0.603	0.433	0.746	0.542	0.580
SDL1	0.718	0.361	0.082	0.598	0.796	0.592
SDL2	0.690	0.414	0.159	0.606	0.813	0.603
SDL3	0.463	0.448	0.381	0.466	0.733	0.581
SDL4	0.583	0.402	0.139	0.550	0.775	0.583
TRP1	0.634	0.518	0.287	0.619	0.659	0.849
TRP2	0.416	0.497	0.427	0.495	0.525	0.762
TRP3	0.730	0.443	0.176	0.680	0.666	0.782
TRP4	0.482	0.525	0.437	0.575	0.541	0.785

Table 4.
Correlations between the constructs and their indicators

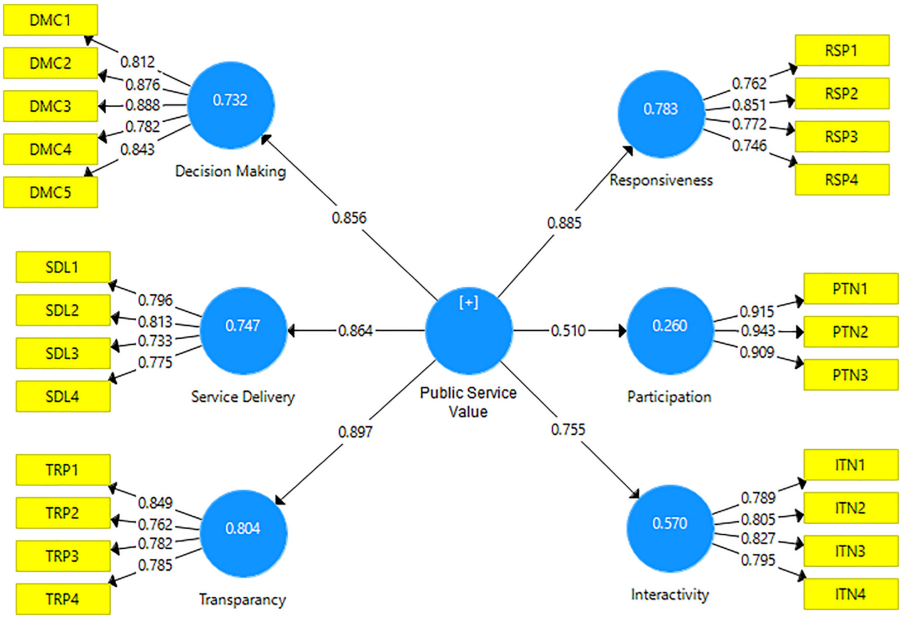
Source(s): Table by authors

Collinearity, evaluated by the Variance Inflation Factor (VIF), should be < 5 (Hair *et al.*, 2011). Path coefficients denote relationships among constructs, with values near +1 indicating strong positive connections and those near zero suggesting weak links. Figure 2 illustrates path coefficient interconnections, demonstrating robust relationships among all constructs. The significance of path coefficients depends on bootstrapped standard error. Empirical t-values and p-values are computed for structural path coefficients, with values exceeding critical thresholds (1.65 for 10% significance, 1.96 for 5% significance) indicating statistical significance. Figure 3 portrays bootstrapping outcomes with measures exceeding the critical value of 1.96 at 5% significance.

The R^2 value, on a scale of 0 to 1, measures predictive accuracy with 0.67, 0.33 and 0.19 denoting substantial, moderate and weak predictability respectively (Chin, 1998). This study's R^2 values mostly fall within the acceptable range, implying substantial predictive accuracy, except for "Participation". Effect size (f^2), interprets exogenous variable effects as small (0.02), medium (0.15), or large (0.35), with below 0.02 indicating no effect (Cohen, 1988). Predictive relevance is appraised via Stone-Geisser's (1974) Q^2 value, indicating model predictive power. A Q^2 value > 0 implies predictive relevance, while <0 suggests otherwise. All constructs exhibit positive predictive relevance. Refer to Table 5 for a summary of relevant measures in the structural model.

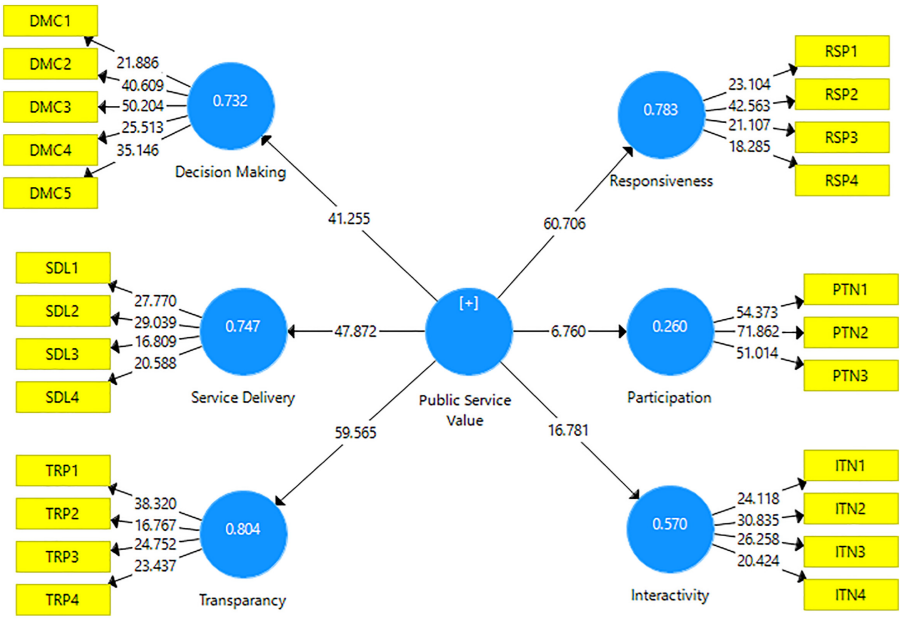
4.3 Hypothesis testing

The reflective-reflective model validation process is conducted in stages, starting with the low-order construct (LOC) and then moving on to the high-order construct (HOC) (Sarstedt *et al.*, 2017). In this study, VCEFG is initially considered the HOC, but it is treated as the LOC because there is another HOC, CEGP, as the outcome variable. The proposed conceptual



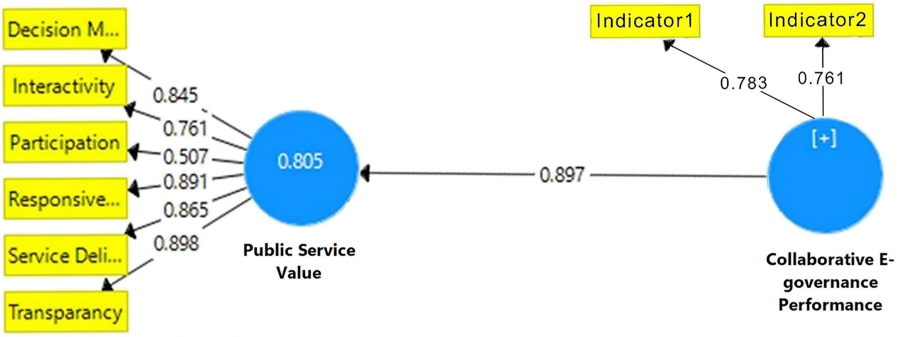
Source(s): Figure by authors

Figure 2.
Path coefficients of the research model (LOC)



Source(s): Figure by authors

Figure 3.
Bootstrapping of the research model (LOC)



Source(s): Figure by authors

Figure 4.
Results of the research model (HOC)

Structural relationship	VIF	R^2	*t*-value*	Path coefficients	*P*-values**
CEGP → VCEG	1.00	0.805	51.847	0.897	0.000
VCEG → VCDM	1.00	0.732	41.255	0.856	0.000
VCEG → VCIN	1.00	0.570	16.781	0.755	0.000
VCEG → VCPT	1.00	0.260	06.710	0.510	0.000
VCEG → VCRP	1.00	0.783	60.706	0.885	0.000
VCEG → VCSD	1.00	0.747	47.872	0.864	0.000
VCEG → VCTP	1.00	0.804	59.565	0.897	0.000

Note(s): *Critical *t*-value for two-tailed test >1.96 (significant level = 5%)
**Critical *P*-value for two-tailed test <0.001 (significant level = 1%)
Source(s): Table by authors

Table 5.
Summary of structural model findings

model for CEGP demonstrates a strong relationship between all paths in its structural model, with path coefficients falling within the recommended threshold of 0-1. The validated research model is shown in Figure 4.

The results of the hypothesis testing are presented in Table 6. The study confirmed the significance of the proposed hypotheses and recommended their acceptance. This validates the conceptual research model and provides valuable insights into the relationship between the different variables. The findings also contribute to the understanding of collaborative e-governance performance and its impact on public service value.

#Hypotheses	Result
HA1. Public service value is a significant constituent of collaborative e-governance performance	Supported
HA2a. Decision making is a significant element of public service value	Supported
HA2b. Service delivery is a significant element of public service value	Supported
HA2c. Responsiveness is a significant element of public service value	Supported
HA2d. Transparency is a significant element of public service value	Supported
HA2e. Participation is a significant element of public service value	Supported
HA2f. Interactivity is a significant element of public service value	Supported

Source(s): Table by authors

Table 6.
Hypotheses and results

4.4 Goodness-of-fit

Goodness-of-fit (GOF) indices evaluate model fitness, computed as ($\sqrt{(AVE \times R^2)}$) (Tenenhaus, Vinzi, Chatelin, & Lauro, 2005). Model fit indices' acceptable range is 0 to 1, with a suggested threshold of 0.36 (Akter, D'ambra, & Ray, 2011). The obtained GOF value, 0.592, surpasses the cut-off, confirming comprehensive model fitness. Standardized Root Mean Square Residual (SRMR) by Henseler *et al.* (2015) was also used to measure GOF in PLS-SEM, where SRMR values of 0.77 (saturated) and 0.088 (estimated) denote a good fit.

5. Discussions

This study investigated the impact of all constructs adopted for the study on enhancing public service value within collaborative e-governance initiatives. The conceptual model for the study posits that VCEG is a constituent construct of CEGP and is composed of six foundational elements essential for the generation of public service value. These elements are VCDM, VCSD, VCRP, VCTP, VCPT and VCIT. The statistical measures for constructs reliability (α) are found to be > .80 which is above the recommended threshold of ≥ .70. Similarly, the validity of all of the constructs in terms of convergent validity (factor loadings ≥.70 and AVE >.50) and divergent validity (Fornell-Larcker Criterion (SQRT AVE) and HTMT ≤.90 (doesn't include 1)) are also found to be significant as per the recommended threshold. Please refer to Table 2 for all statistical values for the measurement model.

The validity of the model was assessed by formulating two sets of hypotheses: one for the primary dimension (VCEG) and another for its constituent sub-dimensions (VCDM, VCSD, VCRP, VCTP, VCPT, VCIT). All hypotheses were supported by the data, as evidenced by the significant structural relationship in terms of path coefficient (β-values), *t*-statistics (≥1.96 at significant level = 5%) and p-values (<0.001 at significant level = 1%). The structural relationship of the primary dimension, CEGP → VCEG ($\beta = 0.897, t = 51.847, p = 0.000$) was also found to be significant. Please refer to Table 5 for statistical values of the structural relationship of all foundational elements for public service value creation.

Key among the findings is the undeniable impact of transparency [VCEG → VCTP ($\beta = 0.897, t = 59.565, p = 0.000$)], responsiveness [VCEG → VCRP ($\beta = 0.885, t = 60.706, p = 0.000$)] and decision-oriented governance [VCEG → VCDM ($\beta = 0.856, t = 41.255, p = 0.000$)] on the overall perception of public service value. Equally noteworthy is the pivotal role of citizen participation and interactivity in the creation of public service value. However, the findings show that although the structural relationship between public service value creation and citizen participation [VCEG → VCPT ($\beta = 0.510, t = 06.710, p = 0.000$)] is significant, its path coefficients. and t-statistic is the lowest among all the correlations in the model. This suggests that citizen participation in government organizations requires greater attention, as it is a key determinant of collaborative e-governance performance and public service value creation.

This research study adds significant value to academic and theoretical discourse through its rigorous examination, statistical validation of variables, validation of conceptual research framework and tested hypotheses. The study findings provide crucial insights through transparency, responsiveness, decision-making, participation and interactivity, this research provides valuable insights that extend beyond mere theoretical postulation. The elucidation of the intricate relationships between these variables lends empirical credence to the theoretical constructs that underlie collaborative e-governance initiatives.

6. Research implications, limitations and future scope

The outcome of this research has significant implications for a variety of stakeholders, including practitioners and policymakers, societies and beneficiaries, as well as researchers and the academic community.

(1) Practitioners and policymakers: The findings can help practitioners and policymakers to improve the design and implementation of collaborative e-governance initiatives. For example, the findings suggest that transparency and responsiveness are essential for creating public service value. Practitioners and policymakers can use this information to design initiatives that are more transparent and responsive to the needs of citizens. Notably, the study's focus on dimensions like citizens' participation and interactivity highlights critical areas for improvement. This emphasizes the need for policymakers to establish strategic mechanisms that encourage citizen engagement, fostering effective communication among government stakeholders.

(2) Societies and beneficiaries: The findings can help societies and beneficiaries to better understand the potential benefits of collaborative e-governance for public service value. For example, the findings suggest that collaborative e-governance can lead to improved service delivery, increased citizen participation and stronger government decisions. This information can help societies and beneficiaries advocate for the development and implementation of collaborative e-governance initiatives for the generation of public service values.

(3) Researchers and academia: The findings of this research can help researchers and academics to better understand the factors that contribute to the success of collaborative e-governance initiatives in generating public service value. The information provided in this paper, which adopts relevant constructs based on a comprehensive literature review and the results of a rigorously tested model, can be used to develop new theories and models of collaborative e-governance. It can also be used to guide future research on this topic.

6.1 Limitations of the study

The study has two limitations. First, it focuses on a limited set of variables, which may not capture the full intricacies of collaborative e-governance. Second, it only considers government-to-citizen (G2C) projects, which narrows the research scope. Future research should explore other types of collaborative e-governance projects, such as government-to-business (G2B) and government-to-government (G2G) projects. This would help us to understand this transformative paradigm better.

6.2 Future research scope

This study can be expanded in three ways: 1) to include G2B and G2G interfaces-this would allow for a more comprehensive understanding of the complex dynamics of collaborative e-governance among stakeholders, 2) to adopt a content-based collaborative approach-this would focus on the exchange of information, knowledge, resources, policies and expertise and 3) to explore other developing countries-for enhancing the applicability of the findings by investigating similar contest. This would help to understand the intricate dynamics of collaborative e-governance and its universality across different contexts.

References

Akter, S., D'ambra, J., & Ray, P. (2011). An evaluation of PLS based complex model: the roles of power analysis, predictive relevant and GOF index. *AMCIS 2011 Proceedings*. Available from: https://aisel.aisnet.org/amcis2011_submissions/151 (accessed 22 August 2022).

Ansell, C., & Gash, A. (2008). Collaborative governance in theory and practice. *Journal of Public Administration Research and Theory*, *18*, 543–571.

Austin, J. E. (2010). From organization to organization: On creating value. *Journal of Business Ethics*, *94*(1), 13–15.

Bannister, F., & Connolly, R. (2014). ICT, public values and transformative government: A model and programme for research. *Government Information Quarterly*, *31*(1), 119–128.

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2012). Promoting transparency and accountability through ICTs, social media, and collaborative e-government. *Transforming Government: People, Process and Policy*, *6*(1), 78–91.

Bryson, J. M., Crosby, B. C., & Bloomberg, L. (2014). Public service value governance: Moving beyond traditional public administration and the new public management. *Public Administration Review*, *74*(4), 445–456.

Central Vigilance Commission (CVC), & Government of India (2020-2021). Annual reports, Available from: https://cvc.gov.in/sites/default/files/Final2021-.pdf (accessed 20 August 2022).

Chin, W. W. (1998). Commentary: Issues and opinion on structural equation modeling. *MIS Quarterly*, 7–16.

Chun, S.A., Luna-Reyes, L. F., & Sandova, R. (2012). Collaborative e-government. *Transforming Government: People, Process and Policy*, *6*(1), 5–12.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cordella, A., & Bonina, M. C. (2012). A public service value perspective for ICT-enabled public sector reforms: A theoretical reflection. *Government Information Quarterly*, *29*, 512–520.

Deng, H., Karunasena, K., & Xu, W. (2018). Evaluating the performance of e-government in developing countries: A public service value perspective. *Internet Research*, *28*(1), 169–190.

Donald, M. N. (1960). Implications of nonresponse for the interpretation of mail questionnaire data. *Public Opinion Quarterly*, *24*(1), 99–114.

Dyer, J. H., & Singh, H. (1998). The relational view: Cooperative strategy and sources of inter-organizational competitive advantage. *Academy of Management Review*, *23*(4), 660–679.

Flak, L., Solli-Saether, H., & Straub, D. (2015). Towards a theoretical model for co-realization of IT value in government. In *Proceedings of 48th Hawaii International Conference on System Sciences* (pp. 2486–2494). IEEE.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.

Geisser, S. (1974). A predictive approach to the random effects model. *Biometrika*, *61*(1), 101–107.

Gómez, L. F., & Heeks, R. (2016). Measuring the barriers to big data for development: Design-reality gap analysis in Colombia's public sector. Development Informatics Working Paper, *62*.

Golubeva, A. A. (2007). Evaluation of regional government portal on the basis of Public service value concept. Case study from Russian Federation. In *ACM International Conference Proceeding Series*, Available from: https://doi.org/10.1145/1328057.1328139 (accessed 5 January 2021).

Green, S. D., & Sergeeva, N. (2019). Value creation in projects: Towards a narrative perspective. *International Journal of Project Management*, *37*(5), 636–651.

Gupta, P. J., & Suri, P. (2017). Measuring public service value of e-governance projects in India: Citizens' perspective. *Transforming Government: People, Process and Policy*, *11*(2), 236–261.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed, a silver bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modelling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135.

Karunasena, K., & Deng, H. (2012). Critical factors for evaluating the public service value of e-government in Sri Lanka. *Government Information Quarterly*, *29*(1), 76–84.

Kelly, G., Mulgan, G., & Muers, S. (2002). *Creating public value*. London: Cabinet Office.

Kernaghan, K. (2003). Integrating values into public service: The values statement as centerpiece. *Public Administration Review*, *63*(6), 711–719.

Kumar, R., Singh, R. K., & Shankar, R. (2016). Study on collaboration and information sharing practices for SCM in Indian SMEs. *International Journal of Business Information Systems*, *22*(4), 455–475.

Lofstedt, U. (2012). E-Governance assessment of current research and some proposals for future directions. *International Journal of Public Information Systems*, *1*, 39–59.

Luna-Reyes, L. F., & Gil-Garcia, J. R. (2011). Using institutional theory and dynamic simulation to understand complex e-Government phenomena. *Government Information Quarterly*, *28*(3), 329–345.

Luna-Reyes, L. F., Picazo-Vela, S., Luna, D. E., & Gil-Garcia, J. R. (2016). Creating public service value through digital government: Lessons on inter-organizational collaboration and information technologies. *System Sciences (HICSS), 2016 49th Hawaii International Conference* (pp. 2840–2849). IEEE.

McDermott, P. (2010). Building open government. *Government Information Quarterly*, *27*, 401–413.

Moore, M. (1995). *Creating public value-strategic management in government*. Cambridge, MA: Harvard University Press.

Moore, M.H. (2014). Public service value accounting: Establishing the philosophical basis. *Public Administration Review*, *74*(4), 465–477.

O'Cass, A., & Ngo, L.V. (2011). Examining the firm's value creation process: A managerial perspective of the firm's value offering strategy and performance. *British Journal of Management*, *22*(4), 646–671.

Osman, I. H., Anouze, A. L., Irani, Z., Ayoubi, B. A., Lee, H., Balc, A., ... Weerakkody, V. (2014). COBRA model to evaluate e-government services: A citizen-centric perspective. *Government Information Quarterly*, *31*(2), 243–256.

O'Leary, R., & Vij, N. (2012). Collaborative public management: Where have we been and where are we going? *The American Review of Public Administration*, *42*(5), 507–522.

Pandey, J. K., & Suri, P. K. (2020). Collaboration competency and e-governance performance. *International Journal of Electronic Governance*, *12*(3), 246–275.

Pandey, J. K., & Suri, P. K. (2020). Drivers of strategic collaboration for e-governance in India: A qualitative study. *Journal for Global Business Advancement*, *13*(5), 605–642.

Pardo, T. A., Nam, T., & Burke, G. B. (2012). E-government interoperability: Interaction of policy, management, and technology dimensions. *Social Science Computer Review*, *30*(1), 7–23.

Pereira, B., Lohmann, G., & Houghton, L. (2021). The role of collaboration in innovation and value creation in the aviation industry. *Journal of Creating Value*, *7*(1), 44–59.

Pham, H. T., & Tanner, K. (2014). Collaboration between academics and librarians: A literature review and model for analysis. *Library Review*, *63*(1/2), 15–45.

Picazo-Vela, S., Gutiérrez-Martínez, I., Duhamel, F., Luna, D. E., & Luna-Reyes, L. F. (2015). Inter-organizational collaboration and value creation in digital government projects. *Proceedings of the 16th Annual International Conference on Digital Government Research*, Fornell ACM (pp. 133–139).

Ramaswamy, V., & Ozcan, K. (2018). What is co-creation? An interactional creation model and its implications for value creation. *Journal of Business Research*, *84*, 196–205.

Reddick, C. G. (2011). Citizen interaction and e-government. *Transforming Government: People, Process and Policy*, *5*(2), 167–184.

Ringle, C. M., Sarstedt, M., Mitchell, R., & Gudergan, S. P. (2020). Partial least squares structural equation modelling in HRM research. *The International Journal of Human Resource Management*, *31*(12), 1617–1643.

Sarstedt, M., Ringle, C.M., & Hair, J.F. (2017). Partial least squares structural equation Modeling. In Homburg, C., Klarmann, M., & Vomberg, A. (Eds.), *Handbook of market research*. Heidelberg: Springer.

Satapathy, S. (2014). An analysis for service quality enhancement in electricity utility sector of India by SEM. *Benchmarking: An International Journal*, *21*(6), 964–986.

Schoenmaker, D., & Schramade, W. (2019). Investing for long-term value creation. *Journal of Sustainable Finance & Investment*, *9*(4), 356–377.

Sheikh, S., & Mandelkern, B. (2014). The Delhi development authority: accumulation without development. In *A Report of the Cities Of Delhi Project*, Centre for Policy Research: New Delhi.

Soni, V., Dey, P. K., Anand, R., Malhotra, C., & Banwet, D. K. (2017). Digitizing grey portions of e-governance. *Transforming Government: People, Process and Policy*, *11*(3), 419–455.

Soni, V., Anand, R., Dey, P. K., Dash, A. P., & Banwet, D. K. (2017). Quantifying e-governance efficacy towards Indian–EU strategic dialogue. *Transforming Government: People, Process and Policy*, *11*(4), 535–571.

Suri, P. K. (2014). Flexibility of processes and e-governance performance. *Transforming Government: People, Process and Policy*, *8*(2), 230–250.

Suri, P. K. (2016). A Strategic analysis of actor competencies and e-governance performance. *Flexible Work Organizations* (pp. 241–259). New Delhi: Springer.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis*, *48*(1), 159–205.

Tripathi, R., Gupta, M. P., & Bhattacharya, J. (2012). Interoperability adoption among government and corporate portals in India: A study. *Journal of Enterprise Information Management*, *25*(2), 98–122.

United Nations Department of Economic and Social Affairs (2016). *E-government survey 2016*. New York: E-government in Support of Sustainable Development.

Supplementary

The supplementary material for this article can be found online.

About the author

Dr Jitendra Kumar Pandey serves as the Deputy Director at the Delhi Development Authority (DDA), a prominent central autonomous government organization under the Ministry of Housing and Urban Affairs, Government of India. He is a highly qualified professional, holding a Master's in Business Administration (MBA) with dual specializations in Strategy and Technology Management from the prestigious Indian Institute of Technology/Delhi. Additionally, he has earned a doctoral degree (PhD) in the field of e-governance from the esteemed Delhi Technological University/Delhi. Dr Pandey's areas of expertise encompass diverse domains such as strategy, soft technology, collaboration, e-government and public policy, among others. He has actively contributed to the field through his scholarly work, with notable publications and presentations at both national and international conferences, showcasing his dedication to advancing knowledge and innovation in the realm of e-governance and related disciplines. Jitendra Kumar Pandey can be contacted at: jitenpandey@gmail.com

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**

Colloid and Polymer Science
https://doi.org/10.1007/s00396-023-05198-8

RESEARCH

Fabrication and evaluation of stimuli-sensitive xanthan gum-based hydrogel as a potential carrier for a hydrophobic drug ibuprofen

Priyanka Meena[1] · Poonam Singh[1] · Sudhir G. Warkar[1]

Received: 26 September 2023 / Revised: 28 October 2023 / Accepted: 8 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

This research work aimed at the synthesis and characterization of novel xanthan gum/polyacrylamide/polyvinyl pyrrolidone (XG/PAM/PVP)-based hydrogels for the delivery of the hydrophobic drug ibuprofen. The influence of crosslinker, biopolymer, and initiator amount on the hydrogel's swelling ratio, porosity, and gel fraction was also investigated. Several techniques like Powder X-ray Diffraction (PXRD), Attenuated Total Reflection-Fourier Transform Infrared spectroscopy (ATR-FTIR), Scanning Electron Microscopy (SEM), and Thermogravimetry Analysis (TGA) were employed to characterize the synthesized hydrogels. Sol–gel analysis revealed an increase in the gel fraction with a rise in the concentration of MBA, XG, and KPS. Porosity measurements indicated a greater porosity with higher XG and KPS amounts, while porosity decreased as the MBA amount increased. Furthermore, the presence of PVP in the hydrogel was observed to greatly enhance various properties, including gel fraction, swelling ratio, porosity, drug loading, and drug release percent. Notably, under alkaline conditions (pH 7.4), the swelling ratio and release of ibuprofen were significantly increased, as compared to acidic (pH 1.2). The ibuprofen release from the XG/PAM/PVP hydrogel is governed by a Fickian diffusion mechanism, and the Korsmeyer-Peppas model serves as the most appropriate model for elucidating the drug release kinetic in both pH 1.2 and 7.4. The pH-responsive nature exhibited by the XG/PAM/PVP hydrogel highlights its potential as a drug delivery system.

Keywords Hydrogel · Biopolymer · Porosity · Drug release · Kinetic modeling

Introduction

Ibuprofen is one of the extensively used nonsteroidal anti-inflammatory drug (NSAID) having poor solubility, high permeability, and limited bioavailability [1]. It is chemically called as 2-(4-(2-methylpropyl)phenyl) propionic acid and is commonly prescribed for its analgesic, antipyretic, and anti-inflammatory properties, particularly in the treatment of rheumatoid arthritis and osteoarthritis [2]. However, owing to its shorter half-life and rapid elimination from the body, multiple dosages are required to maintain an extended pharmacological effect [3]. This frequent dosing regimen is associated with side effects, i.e., gastrointestinal problems, nausea, ulcers etc. [4]. To address these challenges, the development of a novel biopolymer-based hydrogel is required to minimize adverse effects and achieve extended delivery, thereby enhancing therapeutic efficacy [5].

Xanthan gum (XG) is an anionic biopolymer derived from the microorganism *Xanthomonas campestris* [6–8]. The primary structure of this naturally occurring polysaccharide comprises the side chain of glucuronic acid and pyruvic acid [9]. XG-based hydrogels have emerged as a promising material for drug delivery systems due to their unique properties such as hydrophilicity, tunable release rate, and biocompatibility [10, 11].

Hydrogels are three-dimensional networks mainly composed of hydrophilic polymers that possess the capability to hold large amounts of fluid while maintaining their structural integrity [12]. Hydrogels exhibit high biocompatibility and biodegradability, ensuring compatibility with biological systems and minimizing adverse reactions [13, 14]. They also protect drugs from degradation, enzymatic activity, and harsh physiological conditions, leading to enhanced drug stability and bioavailability.

✉ Poonam Singh
poonam@dtu.ac.in

✉ Sudhir G. Warkar
sudhirwarkar@gmail.com

[1] Department of Applied Chemistry, Delhi Technological University, Delhi, India

Published online: 29 November 2023

Basically, hydrogels that exhibit a response to different stimuli, such as pH and temperature, are commonly referred to as stimulus-sensitive hydrogels or smart hydrogels [15, 16]. Among the array of stimuli that have been extensively studied for drug delivery, pH has emerged as a key determinant in facilitating drug delivery [17]. The pH-sensitive hydrogels exhibit either weakly acidic (anionic) or weakly basic (cationic) properties, depending on the ionizable moieties present in their polymer backbones [18]. The swelling or shrinking behavior of these hydrogels in response to pH is primarily governed by the ionization of the functional groups within the polymer structure [6]. In recent decades, extensive research has been dedicated to the fabrication of pH-sensitive hydrogels using synthetic polymers, including polyvinyl pyrrolidone [19], polyacrylamide [20] etc.

Polyacrylamide (PAM), a synthetic polymer, has gained lot many attention in drug delivery applications due to its non-toxicity, exceptional mechanical strength as well as ability to exhibit pH-dependent swelling [20–22].

Poly(vinylpyrrolidone) (PVP) is another non-toxic synthetic polymer, widely used in drug delivery applications [23]. Structurally, this amphiphilic polymer consists of a pyrrolidone ring along with functional groups like C = O and C–N, which contribute to its hydrophilic nature. Furthermore, the presence of an alkyl group enhances its hydrophobic properties [24, 25]. The hydrophobic component of PVP enables favorable interactions with hydrophobic drugs, thereby expected to improve the loading of hydrophobic as well as hydrophilic drugs such as ampicillin sodium [19], theophylline [23], and ibuprofen [26].

To the best of our knowledge, the XG in combination with PVP and PAM have been employed for first time for the development of a novel hydrogel matrix, specifically designed for the release of ibuprofen.

The current study involved the synthesis of pH-sensitive XG/PAM and XG/PAM/PVP-based hydrogels, for the release of ibuprofen. Ibuprofen was loaded into the fabricated hydrogel using an ex situ technique. The addition of PVP was implemented to enhance the loading capacity of the hydrophobic drug (ibuprofen). To confirm the thermal stability, synthesis, surface morphology, and crystalline nature, the hydrogels were characterized using TGA, ATR-FTIR, SEM, and PXRD techniques. The swelling behavior of the developed hydrogels was analyzed at pH 1.2 and 7.4. Sol–gel and porosity studies were also conducted for synthesized hydrogels to analyze the sol–gel fraction and porous nature of the hydrogels. The study mainly focused on estimating the percentage of ibuprofen release through the in vitro release method. To elucidate the drug release mechanism, several kinetic models such as First-Order, Higuchi, Hixson-Crowell, Korsmeyer-Peppas, and Zero-Order were studied.

Materials and methods

Materials

N,*N*′-methylene bis(acrylamide) (MBA, Merck, Germany), polyvinylpyrrolidone K-30 (PVP, CDH, New Delhi, India), potassium persulfate (KPS, Fischer Scientific, Mumbai, India), acrylamide (AM, CDH, New Delhi, India), and xanthan gum (XG, s d fine chem. Ltd., Mumbai, India) were used as obtained. Ibuprofen was courteously provided by Unicure India Pvt Ltd, Noida, India.

Fabrication of xanthan gum/polyacrylamide (XG/PAM) hydrogel and xanthan gum/polyacrylamide/polyvinyl pyrrolidone (XG/PAM/PVP)-based hydrogels

The fabrication of XG/PAM hydrogel (U-1) and series of XG/PAM/PVP hydrogels (U-2 to U-11), with the varied composition of the initiator KPS (25–40 mg), crosslinker MBA (15–30 mg), and biopolymer XG (0.02–0.05 g) was carried out via free radical polymerization. Table 1 details the specific amount of MBA, XG, PVP, KPS, and AM utilized during the fabrication of hydrogels. For the synthesis of XG/PAM hydrogel (U-1), the first fixed amount of XG and AM was dissolved in distilled water followed by continuous stirring on a magnetic stirrer for 1 h. Secondly, KPS and MBA in requisite amounts were added while stirring for an additional 1 h to ensure complete mixing. Then, the mixture was carefully poured into glass test tubes and transferred to a water bath maintained at 60 °C. After 1 h, the tubes were removed from the water bath, and the hydrogel was extracted from the tubes and cut into disks. The obtained hydrogels were then washed with distilled water to remove unreacted chemicals and impurities, if any. Finally, the obtained hydrogel was dried in an oven at 60 °C until a consistent weight was achieved [23, 27].

For the synthesis of a series of XG/PAM/PVP hydrogels (U-2 to U-11), an additional polymer PVP was also introduced simultaneously during the addition of XG and AM, while the rest of the procedure remained the same as outlined above, as presented in Fig. 1 [37].

Swelling studies

To investigate the swelling behavior of the XG/PAM hydrogel and all XG/PAM/PVP hydrogels, gravimetric analysis was conducted. Initially, the pre-dried hydrogels were precisely weighed (M_d) and immersed in an HCl-KCl buffer (pH 1.2) and phosphate buffer saline (pH 7.4) for a duration of 24 h. At intervals of 1 h, the swollen hydrogel disks were carefully extracted and gently blotted with filter paper to eliminate any adhered liquid droplets on the surface. The swelled hydrogels were then weighed again (M_s) and dipped back in the buffer

Table 1 Various compositions of hydrogels with their swelling ratios, gel fraction, and porosity data

Formulation	XG (g)	AM (g)	PVP (g)	MBA (mg)	KPS (mg)	Swelling ratio (%)		Gel fraction (%)	Porosity (%)
						pH 1.2	pH 7.4		
U-1	0.05	2.2	00	15	35	968.88	1049.96	84.27	69.21
U-2	0.02	2.2	0.12	15	35	932.20	1011.14	82.99	69.50
U-3	0.03	2.2	0.12	15	35	957.12	1035.13	86.96	72.35
U-4	0.04	2.2	0.12	15	35	975.28	1070	90.01	75.56
U-5	0.05	2.2	0.12	15	35	997.75	1100.31	91.88	78.35
U-6	0.05	2.2	0.12	15	25	927.91	1034.90	83.44	73.62
U-7	0.05	2.2	0.12	15	30	962.95	1061.88	86.48	75.61
U-8	0.05	2.2	0.12	15	40	975.92	1080.87	87.52	76.59
U-9	0.05	2.2	0.12	20	35	963.02	1062.10	92.39	76.42
U10	0.05	2.2	0.12	25	35	936.98	1027.68	94.40	73.55
U-11	0.05	2.2	0.12	30	35	912	981.13	95.51	71.60

solutions [17, 28]. The entire study was conducted in triplicate to ensure reliability, and the swelling ratio was determined using the provided equation.

$$\text{Swelling Ratio } (\%) = \frac{M_s - M_d}{M_d} \times 100 \quad (1)$$

Sol–gel analysis

To investigate the uncrosslinked part of the XG/PAM hydrogel and all fabricated XG/PAM/PVP-based hydrogels, a sol–gel analysis was conducted. Dried hydrogels were weighed (M_i) and dipped in distilled water for 48 h at room temperature. Subsequently, the extracted swollen hydrogel disks were subjected to oven drying at 45 °C till a consistent weight (M_e) was reached [29, 30]. The sol and gel fraction was quantitatively assessed using the below equation:

$$\text{Sol Fraction } (\%) = \frac{M_i - M_e}{M_e} \times 100 \quad (2)$$

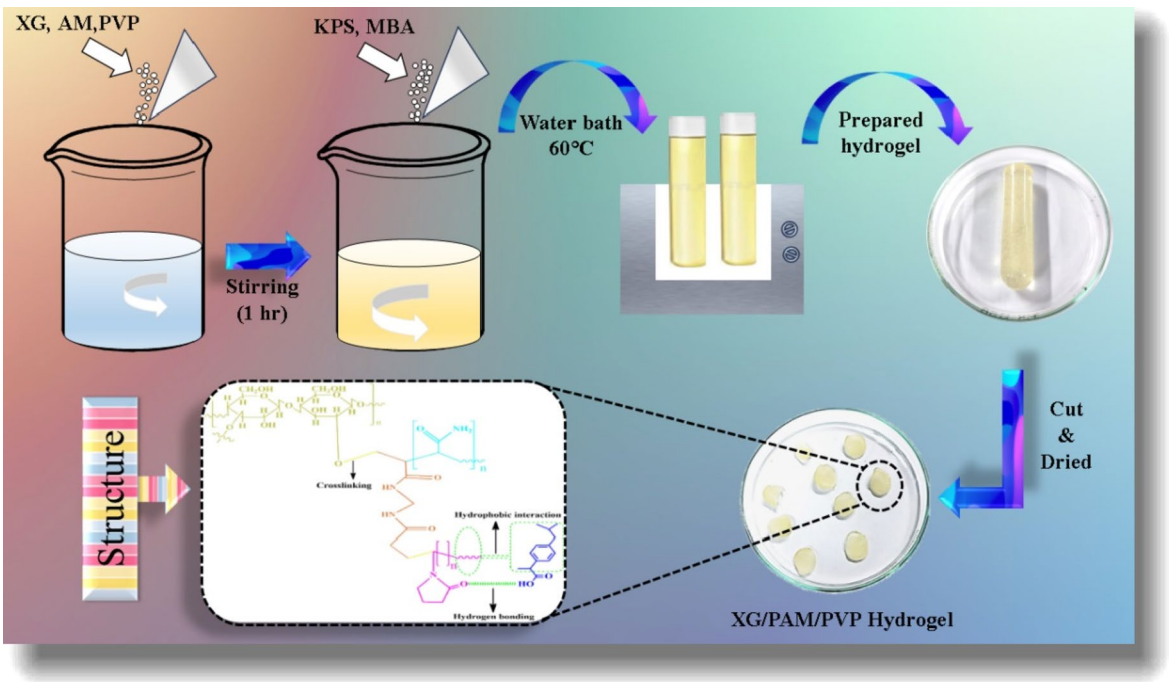


Fig. 1 Schematic illustration of the fabrication of XG/PAM/PVP-based hydrogels

$$\text{Gel fraction (\%)} = 100 - \text{Sol fraction} \quad (3)$$

Porosity measurement

The porosity analysis of all fabricated hydrogels was performed using the solvent displacement method. The initial weighed dry hydrogel disks (M_i) were placed in 20 ml of ethanol overnight. After 24 h, the hydrogel disks were taken out from the ethanol, extra liquid on their surfaces was removed by blotting with filter paper, and the weight of the final dry hydrogel disks (M_f) was recorded. To calculate the porosity of the hydrogel disks, the below formula was used [31]:

$$\text{Porosity (\%)} = \frac{|M_i - M_f|}{\rho V} \times 100 \quad (4)$$

Here, ρ denotes the density of ethanol (0.7826 g/ml) and V denotes the volume of the cylindrical hydrogel disk ($V = \pi r^2 h$, where r is the radius of the disk, and h represents the height of the disk, measured with the help of a vernier caliper) [32].

Ibuprofen loading and entrapment efficiency

Among all the formulated XG/PAM/PVP-based hydrogels, the optimized XG/PAM/PVP (U-5) hydrogel demonstrated the highest swelling capacity. Therefore, drug loading was carried out on this specific formulation, and it was selected for comparison with the control XG/PAM (U-1) hydrogel, which does not contain PVP. To assess the drug loading and encapsulation efficiency of ibuprofen-loaded XG/PAM and XG/PAM/PVP hydrogel, the weight method was employed. The fixed amount of dried hydrogel disks (M_i) were immersed in 100 ml of pH 7.4 buffer solution containing a specific amount (M_D) of the ibuprofen for a period of 24 h at room temperature. After the designated time, the hydrogel disks were carefully removed from the buffer solution and kept in the oven at 40 °C until a constant weight (M_f) was reached [33–36]. The below-mentioned equations were utilized to find out the drug loading (DL%) and drug encapsulation efficiency (DEE%).

$$\text{DL (\%)} = \frac{M_f - M_i}{M_i} \times 100 \quad (5)$$

$$\text{DEE (\%)} = \frac{M_f - M_i}{M_D} \times 100 \quad (6)$$

In vitro ibuprofen release studies

An in vitro release study of ibuprofen was conducted using ibuprofen tablet 200 mg, ibuprofen-loaded XG/PAM hydrogel, and ibuprofen-loaded XG/PAM/PVP hydrogel in an orbital incubator shaker maintained at 37 °C. The fixed amount of the ibuprofen-loaded hydrogel disks was dipped in 100 ml of pH 1.2 and 7.4 buffer solutions [37]. The 3 ml aliquot was collected from each buffer solution at regular time intervals (1 h), and an equal volume of fresh buffer solutions was added to ensure constant volume. The solubility of ibuprofen is extremely low at pH 1.2; hence to determine the quantity of drug release, the withdrawn sample was diluted using an equivalent volume of ethanol. The sample was analyzed using a UV–visible spectrophotometer (model: Cary 300 UV–Vis), recording the absorbance at the λ_{max} 223 nm. The ibuprofen amount in the released samples was determined via a calibration curve constructed using various concentrations of drug solution in pH 7.4 buffer. To ensure accuracy, the experiment was done in triplicate [38].

Kinetic modeling

The various mathematical models, such as First-Order, Zero-Order, Higuchi, Hixson-Crowell, and Korsmeyer-Peppas, were employed to elucidate the drug release mechanism. The release data acquired from the in vitro studies were fitted to each of these models. The goodness of fit was assessed by comparing the regression coefficient (R^2) values associated with each model. The model exhibiting an R^2 value close to 1 is termed the most suitable and best-fitted model for describing the drug release mechanism [39].

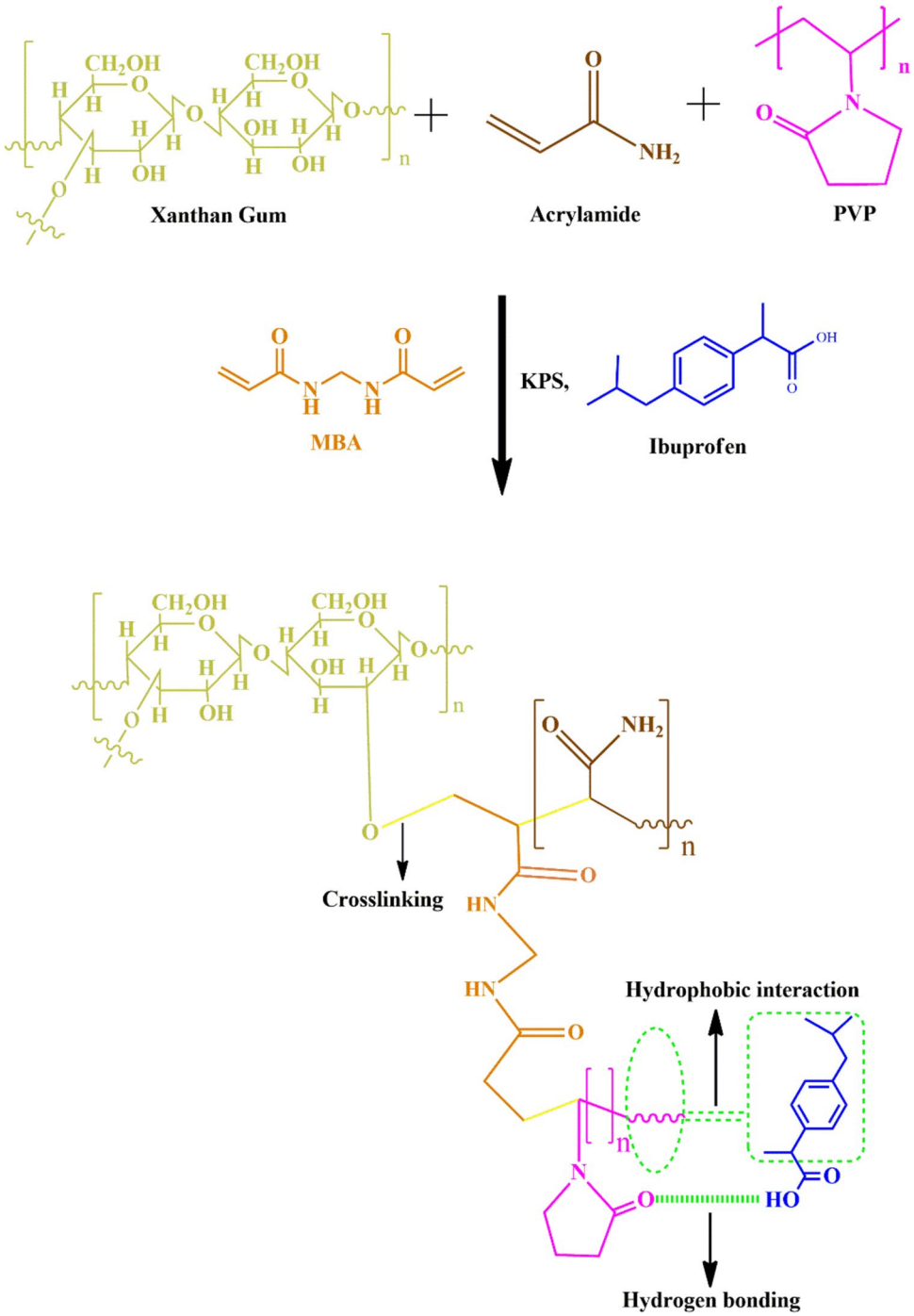
Characterization

The PXRD measurements were done using a Bruker D8 diffractometer, with Cu Kα radiation ($\lambda = 1.5418$ Å), over a 2θ of 5–70° using a Gobel mirror. The ATR-FTIR spectra were recorded using an ATR-FTIR spectrophotometer (model: PerkinElmer spectrum 2) within a range of 450–4000 cm^{-1}. Thermogravimetric analysis (TGA) was conducted using a PerkinElmer TGA 4000 instrument under a nitrogen atmosphere in the temperature from 30 to 900 °C at a heating rate of 10 °C/min. The surface morphology was examined via SEM (model: JEOL Japan Mode: JSM 6610LV).

Result and discussion

Mechanistic pathway of ibuprofen-loaded XG/PAM/PVP hydrogel formation

In the synthesis of all ibuprofen-loaded XG/PAM/PVP hydrogels using a free radical mechanism, KPS (initiator) and MBA (crosslinker) played a key role. Scheme 1 depicts the proposed mechanism of the hydrogel fabrication. At



Scheme 1 Mechanism for ibuprofen-loaded XG/PAM/PVP hydrogel

60 °C, the KPS decomposes, releasing sulfate radicals ($SO^{4\bullet-}$). The highly reactive radicals initiate the polymerization of acrylamide monomers by breaking the vinyl double bond of the acrylamide and generating the free radical on the acrylamide. This sulfate radical also abstracts a hydrogen atom from XG and PVP and generates a reactive free radical. To introduce crosslinks and form a three-dimensional network, a crosslinking agent (MBA) is added. The MBA molecules consist of two reactive ends that react with active radical sites of XG, PAM, and PVP. This results in the formation of covalent bonds between the polymeric chains, leading to the formation of a crosslinked network structure of XG/PAM/PVP hydrogel. Moreover, in ibuprofen-loaded XG/PAM/PVP hydrogel, the hydrophobic interaction exists between the alkyl group of PVP and the aryl ring of ibuprofen, as well as the formation of hydrogen bonds between the –O atom of pyrrolidone ring of PVP and –OH of ibuprofen took place [24, 40].

Swelling studies

The swelling ratio of hydrogel was influenced by the amount of the biopolymer, initiator, and crosslinker, as depicted in Fig. 2. The swelling data of all synthesized hydrogels in pH 1.2 and pH 7.4 buffer were illustrated in Table 1.

Influence of biopolymer

The impact of varying XG amounts (U-2 to U-5 formulation) on the swelling ratio was illustrated in Fig. 2a. The swelling ratio of hydrogels rises as the amount of XG increases from 0.02 to 0.05 g. Notably, the highest swelling 1100.31% was observed at pH 7.4. However, at pH 1.2, a swelling ratio of 997.75% was noticed when using 0.05 g of XG. This can be attributed to an increase in hydrophilicity as the XG quantity increases. As the XG content rises, the number of hydrophilic groups such as –OH attached to XG also increases; consequently, high swelling is observed. Furthermore, a solution with an XG amount exceeding 0.05 g faced challenges due to the resulting highly viscous solution, rendering stirring difficult. Hence, the formation of hydrogel with XG above 0.05 g was not attempted [17, 40].

Influence of initiator

The highest swelling occurred with 0.05 g of XG; hence, this value amount of XG was kept constant in all the attempted formulations and the effect of varying KPS (initiator) concentration (U-6 to U-8) on swelling is observed, as depicted in Fig. 2b. At a KPS amount of 35 mg, the highest swelling of approximately 1100.31% occurred at pH 7.4, whereas pH

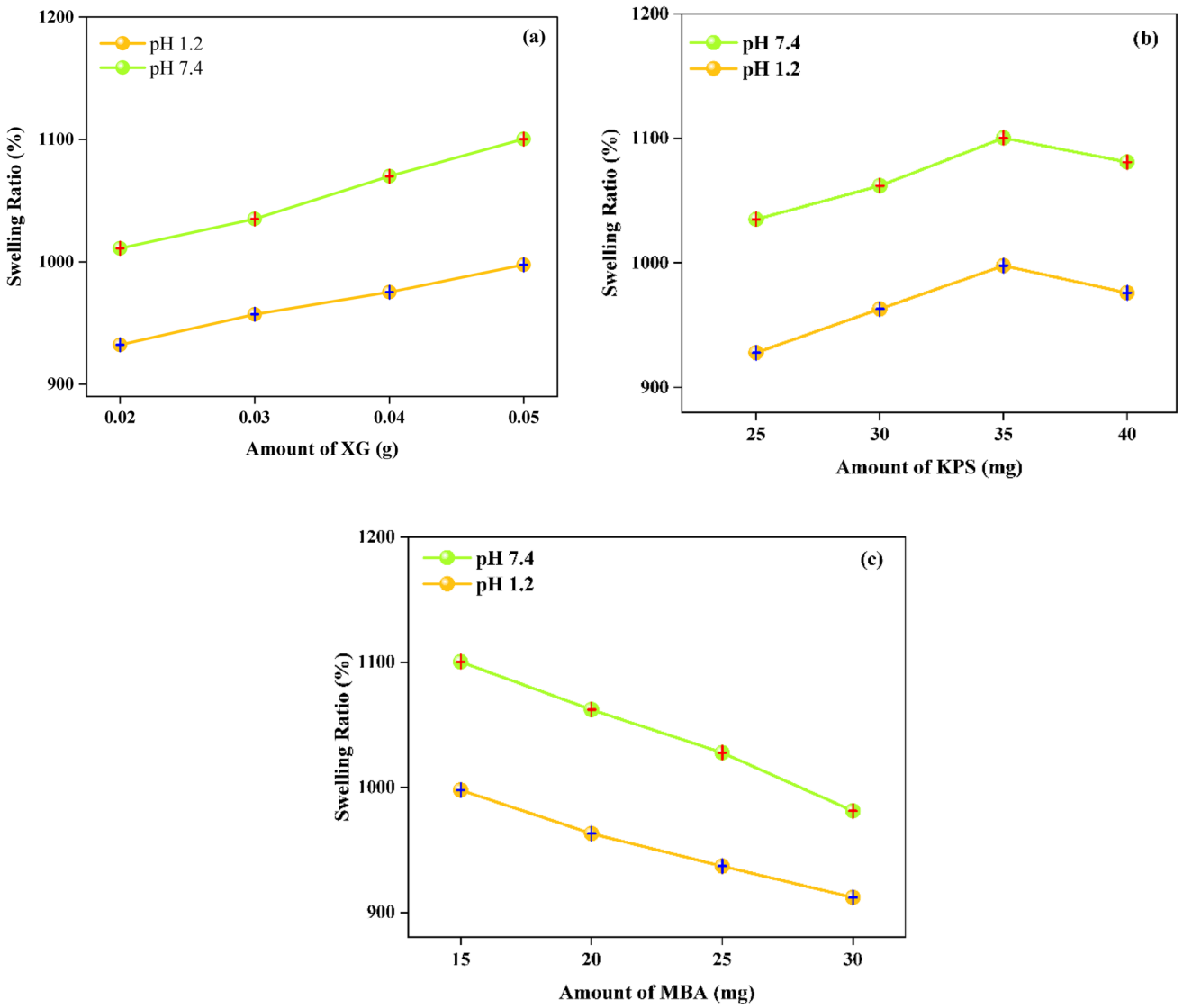


Fig. 2 Influence of the **a** XG, **b** KPS, and **c** MBA on the swelling ratio of the hydrogel

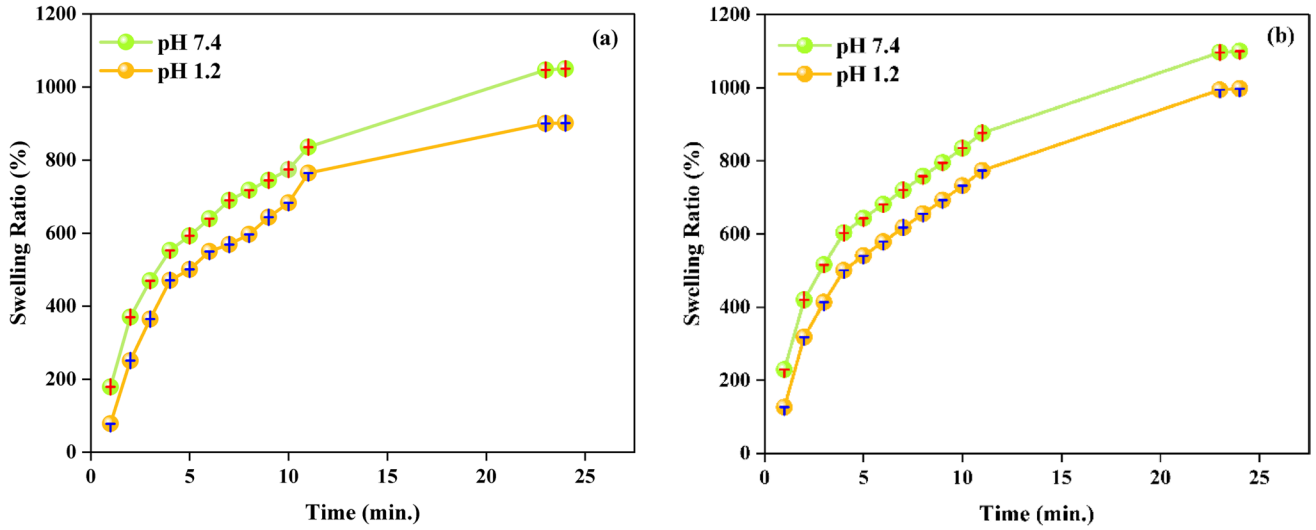


Fig. 3 Swelling studies of **a** XG/PAM hydrogel and **b** XG/PAM/PVP hydrogel in pH 7.4 and 1.2

1.2 exhibited a swelling of 997.75%. The swelling ratio was noted to rise as the amount of KPS increased from 25 to 35 mg, and subsequently, it decreased beyond 35 mg. The initial rise in swelling could be linked to a higher number of free radicals along the polymeric chain, leading to more polymerization, and greater swelling. However, the decrease in swelling results from the bimolecular collision which took place as the reaction rate increased, which increases the crosslinking density and decreases swelling [17].

Influence of crosslinker

The influence of crosslinker amount (U-9 to U-11) on the swelling ratio was examined. As depicted in Fig. 2c, a higher MBA amount yields greater swelling. The increased crosslinker amount from 15 to 30 mg reduces the available space between the polymeric chains. Consequently, the resulting highly rigid crosslinked network cannot expand and hence swelling decreases [17]. The highest swelling of 1100.31% at pH 7.4 and 997.75% at pH 1.2 is attained at an MBA concentration of 15 mg.

In all the formulations, it was noticed that the hydrogel exhibited greater swelling at pH 7.4 in comparison to pH 1.2. This can be attributed to the deprotonation of the COOH group present in XG in the pH 7.4 buffer solution. The repulsion took place between the deprotonated COO^- ions, resulting in chain relaxation, which generates additional free space within the polymeric network and consequently, the hydrogel exhibited more swelling [41]. On the other hand, the presence of hydrogen bonding interactions among PAM, PVP, and XG, at pH 1.2, causes the hydrogel to shrink, resulting in a lesser swelling compared to the alkaline pH [28].

In comparison of the XG/PAM/PVP (U-5) hydrogel to XG/PAM (U-1) hydrogel, the XG/PAM/PVP (U-5) hydrogel shows notably higher swelling, as observed in Fig. 3. It is primarily attributed to the presence of PVP, an amphiphilic polymer containing hydrophilic moieties such as C = O and C–N, responsible for promoting greater swelling. XG/PAM/PVP hydrogel exhibits a swelling ratio of 1100.31% at pH 7.4 and 997.75% at pH 1.2. However, the XG/PAM hydrogel exhibited a swelling ratio of 1049% at pH 7.4 and 901% at pH 1.2.

Sol–gel analysis

The gel fraction and sol fraction for all fabricated hydrogels are illustrated in Fig. 4, and their gel fraction values are presented in Table 1. As the amount of XG (U-2 to U-5) and KPS (U-6, U-7) increases, a noticeable rise in gel fraction is observed. This is attributed to the generation of a higher number of free radicals, leading to a greater number of active sites available for crosslinking. Consequently, the crosslinking density within the polymer network rises, resulting in an increase in gel fraction. However, the further increase in the concentration of KPS (U-8) prompts the generation of soluble chains (oligomer) instead of interconnected crosslinked networks; consequently, the gel fraction decreases. Moreover, the rise in MBA amount (U-9, U-10, U-11) increases crosslinking within the polymeric chain, which corresponds to a rise in gel fraction.

In a comparison of XG/PAM/PVP (U-5) hydrogel with XG/PAM (U-1) hydrogel, the result shows that an addition of polymer (PVP) results in a decrease of the sol fraction and an increase of the gel fraction. This can be attributed to the presence of higher concentration of polymer chains containing free radicals, on incorporation of PVP. Consequently, numerous active sites for crosslinking are generated, leading to more crosslinked parts, subsequently increasing

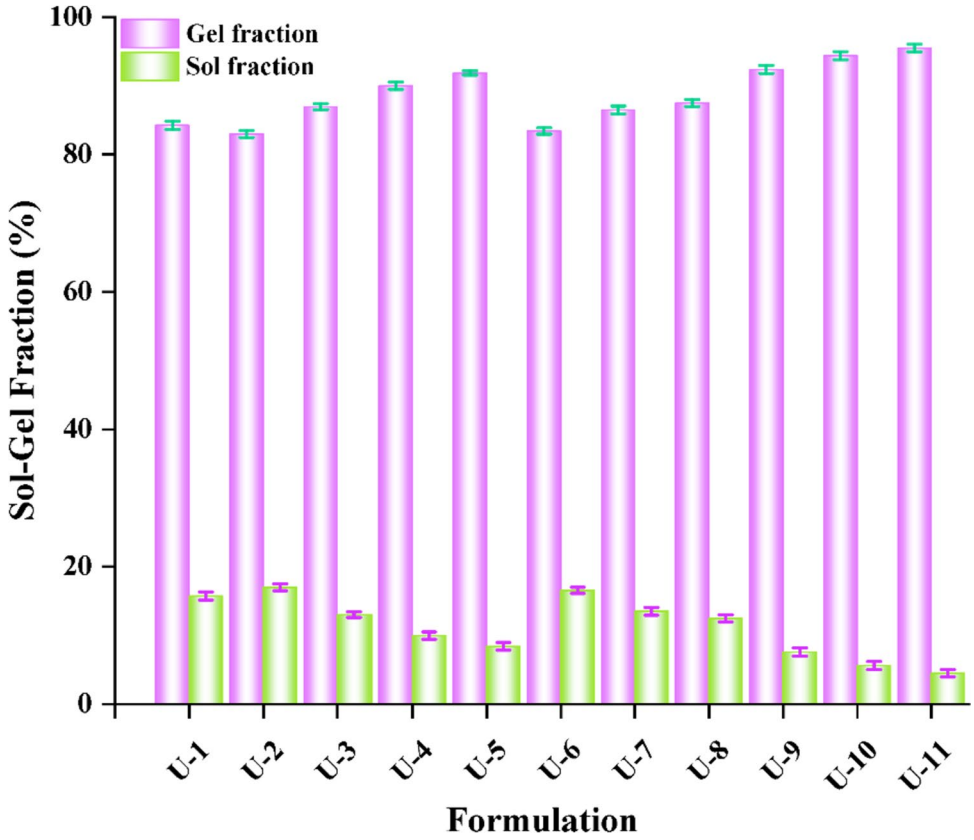


Fig. 4 Sol–gel studies of all fabricated XG/PAM/PVP hydrogels and XG/PAM hydrogel

the gel portion. However, as the gel fraction increases, the sol fraction decreases due to their inverse relationship [42].

Porosity measurement

The analysis of the porosity data presented in Table 1 illustrates that as the amount of the biopolymer XG (U-2 to U-5) is increased, porosity also increases, due to the rise in solution viscosity. The presence of a viscous solution effectively hinders the escape of bubbles, leading to the formation of an interconnected network and consequently contributing to the observed increase in porosity, as depicted in Fig. 5. Furthermore, initially on increasing the initiator, KPS (U-6, U-7), more radicals are generated, leading to higher polymerization rate. This rapid polymerization results in the generation of more voids within the hydrogel structure, thus increasing porosity. However, on further increase in KPS (U-8), the polymerization rate becomes too fast. This causes polymer chains to crosslink too quickly, leading to denser and less porous structures; hence, porosity decreases. Moreover, increasing the amount of crosslinker, MBA (U-9 to U-11), the entanglement between the polymeric chain of XG, PVP, and PAM increases, resulting in a denser structure and lesser porosity [43]. The XG/PAM hydrogel (U-1) exhibits lower porosity in comparison to the XG/PAM/PVP hydrogel (U-5). This revealed that the addition of PVP enhances the viscosity of the solution, resulting in the entrapment of bubbles; consequently, porosity increases [30].

Characterization

PXRD analysis was employed to analyze both the crystallinity of the hydrogel and the distribution of the drug within the polymeric matrix, as depicted in Fig. 6. In the PXRD pattern of ibuprofen, the characteristic peaks corresponding to ibuprofen were observed at 2θ values of 6.1, 12.3, 16.7, 20.1, 22.4, and 27.5 [2]. The distinct and well-defined peaks of ibuprofen indicated its crystalline nature [28]. While the broad peaks observed in the XG/PAM/PVP (U-5) and XG/PAM (U-1) hydrogel diffractograms suggested that the hydrogel matrix lacked a well-defined crystalline structure, however exhibited amorphous characteristics. Moreover, in the ibuprofen-loaded XG/PAM/PVP hydrogel, the peaks of ibuprofen were absent. This absence indicated the uniform dispersion of ibuprofen within the hydrogel [44].

The ATR-FTIR spectra of MBA, XG/PAM hydrogel (U-1), XG/PAM/PVP (U-5) hydrogel, and ibuprofen-loaded XG/PAM/PVP hydrogel are illustrated in Fig. 7, while their corresponding ATR-FTIR data are summarized in Table 2. In the ATR-FTIR spectrum of MBA, the peaks at 1304 cm^{-1} corresponding to C–N stretch shifted to 1318 cm^{-1}, 1319 cm^{-1}, and 1320cm^{-1} in XG/PAM, XG/PAM/PVP, and

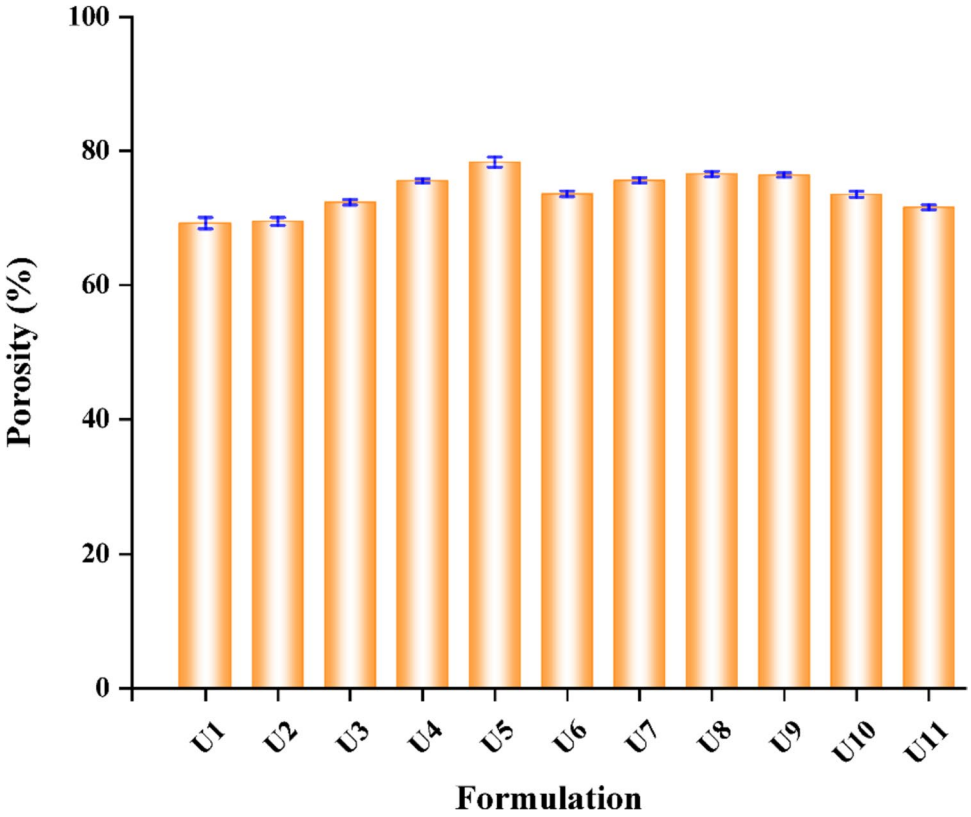


Fig. 5 Porosity plot of all fabricated XG/PAM/PVP hydrogels and XG/PAM hydrogel

ibuprofen-loaded XG/PAM/PVP hydrogel, respectively, due to crosslinking. Moreover, the peak representing the C = C stretch shifted to 1649 cm^{-1}, 1650 cm^{-1}, and 1651 cm^{-1} in XG/PAM, XG/PAM/PVP, and ibuprofen-loaded XG/PAM/PVP hydrogel, respectively, confirming the crosslinking. These results provide confirmation of the successful synthesis of crosslinked hydrogels [40].

The XG/PAM/PVP hydrogel exhibited almost identical spectra to XG/PAM hydrogel due to the overlapping of the characteristic peaks (C = O, N–H, C–H) of PVP with PAM, resulting in no additional visible peaks upon the addition of PVP. Only changes in intensity and little shifts in wave number due to the interaction of PVP were observed.

In the case of ibuprofen-loaded XG/PAM/PVP hydrogel, no new peaks appeared or disappeared as compared to the spectra of unloaded XG/PAM/PVP hydrogel. This observation suggests the absence of chemical interaction, rather, only a physical interaction, such as hydrogen bonding, occurred between the ibuprofen and hydrogel [28].

TGA was employed to assess the thermal stability of the synthesized polymeric network and to investigate interactions between the hydrogel and the drug, as depicted in Fig. 8. The TGA curve showed four weight loss stages for all three types of hydrogels: XG/PAM (U-1), XG/PAM/PVP (U-5), and ibuprofen-loaded XG/PAM/PVP hydrogel. In the TGA curve of XG/PAM hydrogel, the first stage involved water vaporation and initial decomposition of polymer chains, leading to a weight loss of 18.3% between temperatures of 30-270 °C. The second stage, within the temperature range of 270–373 °C, marked the degradation of the polymer backbone, including carboxymethyl and hydroxyl groups, resulting in a weight loss of 19.3%. The third stage, occurring between 373 and 450 °C, indicated the breakdown of the polymeric network's crosslinking, leading to a weight loss of 33.1%. Finally, the fourth stage, observed between 450 and 900 °C, indicates the complete degradation of the hydrogel, causing a weight loss of 26.4%.

In both XG/PAM/PVP hydrogel and ibuprofen-loaded XG/PAM/PVP hydrogel, the initial stage involved the vaporation of water molecules and initial degradation of polymeric chains, leading to weight losses of 14.1% and 12.9%, respectively, within the temperature range of 30–272 °C. The second stage denotes the degradation of the polymer backbone, particularly carboxymethyl and hydroxyl groups, in XG/PAM/PVP hydrogel and ibuprofen-loaded XG/PAM/PVP hydrogel, resulting in weight losses of 18.7% and 17.8%, respectively, within the temperature range of 272–375 °C. The third stage indicated the disintegration of the crosslink bond present in the hydrogel, as weight losses of 36.5% and 37.84% within the temperature range of 375–478 °C. The fourth stage denoted the entire hydrogel decomposition, leading to weight losses of 27% and 26.6% within the temperature range of 478–900 °C [46].

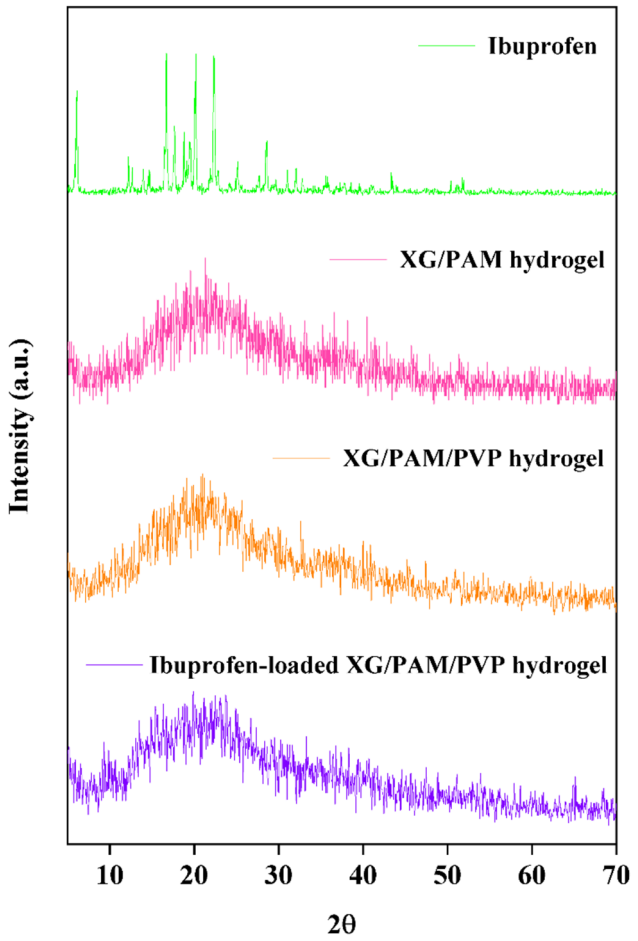


Fig. 6 PXRD plot of ibuprofen, XG/PAM hydrogel, XG/PAM/PVP hydrogel, and ibuprofen-loaded XG/PAM/PVP hydrogel

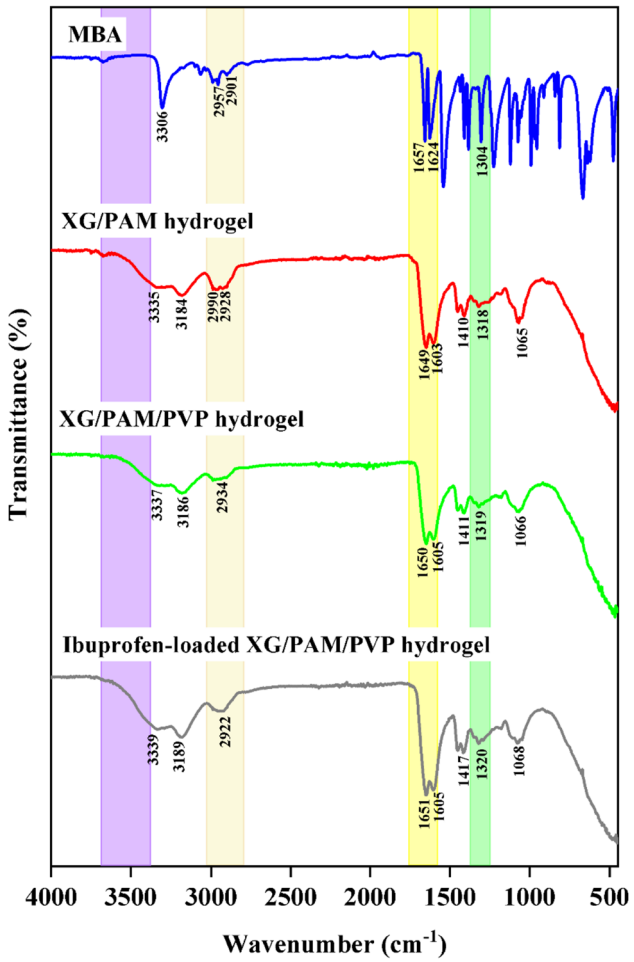


Fig. 7 ATR-FTIR spectra of MBA, XG/PAM hydrogel, XG/PAM/PVP hydrogel, and ibuprofen-loaded XG/PAM/PVP hydrogel

Notably, the TGA curves of XG/PAM/PVP hydrogel and ibuprofen-loaded XG/PAM/PVP hydrogel were very similar. This similarity suggests that ibuprofen does not significantly affect the hydrogel's stability, implying an absence of chemical interaction. Instead, it suggests that there might be a physical or weak interaction, like hydrogen bonding, between ibuprofen and the hydrogel.

SEM micrographs of XG/PAM/PVP (U-5) and ibuprofen-loaded XG/PAM/PVP hydrogel were depicted in Fig. 9. The XG/PAM/PVP hydrogel exhibits an irregular surface morphology and a porous structure, revealing the hydrogel's favorable fluid retention and drug loading capacity.

Moreover, the micrographs of ibuprofen-loaded XG/PAM/PVP hydrogel display a little smooth surface with fewer visible pores as compared to the unloaded hydrogel. This observation suggests that the ibuprofen has successfully been encapsulated into the hydrogel matrix, resulting in reduced interstitial voids within the hydrogel structure [47].

Ibuprofen loading and entrapment efficiency

The ibuprofen loading (DL%) and entrapment efficiency (DEE%) were calculated as per Eqs. (5) and (6), respectively, for ibuprofen-loaded XG/PAM hydrogel and ibuprofen-loaded XG/PAM/PVP hydrogel, and their data displayed in Table 3. Notably, ibuprofen-loaded XG/PAM/PVP hydrogel exhibited higher DL% and DEE% values than ibuprofen-loaded XG/PAM hydrogel. This can be attributed to the hydrophobic interaction between the alkyl group of PVP and the aryl ring of ibuprofen, responsible for increasing the loading of hydrophobic drug ibuprofen in XG/PAM/PVP hydrogel [24].

In vitro ibuprofen release studies

The in vitro release analysis was conducted for pure ibuprofen tablet 200 mg, ibuprofen-loaded XG/PAM hydrogel, and ibuprofen-loaded XG/PAM/PVP hydrogel as shown in Fig. 10. The ibuprofen-loaded XG/PAM/PVP hydrogel

Table 2 ATR-FTIR data of MBA, XG/PAM hydrogel, XG/PAM/PVP hydrogel, and ibuprofen-loaded XG/PAM/PVP hydrogel

Functionality	Wavenumber (cm^{-1})				Ref.
	MBA	XG/PAM hydrogel	XG/PAM/PVP hydrogel	Ibuprofen-loaded XG/PAM/PVP hydrogel	
C–N	1304	1318	1319	1320	[40]
C = C and N–H bending overlap	1624	1649	1650	1651	[40]
C = O	1657	1603	1605	1605	[17]
N–H and O–H stretching overlap	–	3335 and 3184	3337 and 3186	3339 and 3189	[27]
COO^- stretching	–	1410	1414	1417	[27]
C–H stretching	2901 and 2957	2928 and 2990	2934	2922	[39]
C–O–C	–	1065	1066	1068	[39]
N–H stretching	3306	–	–	–	[45]

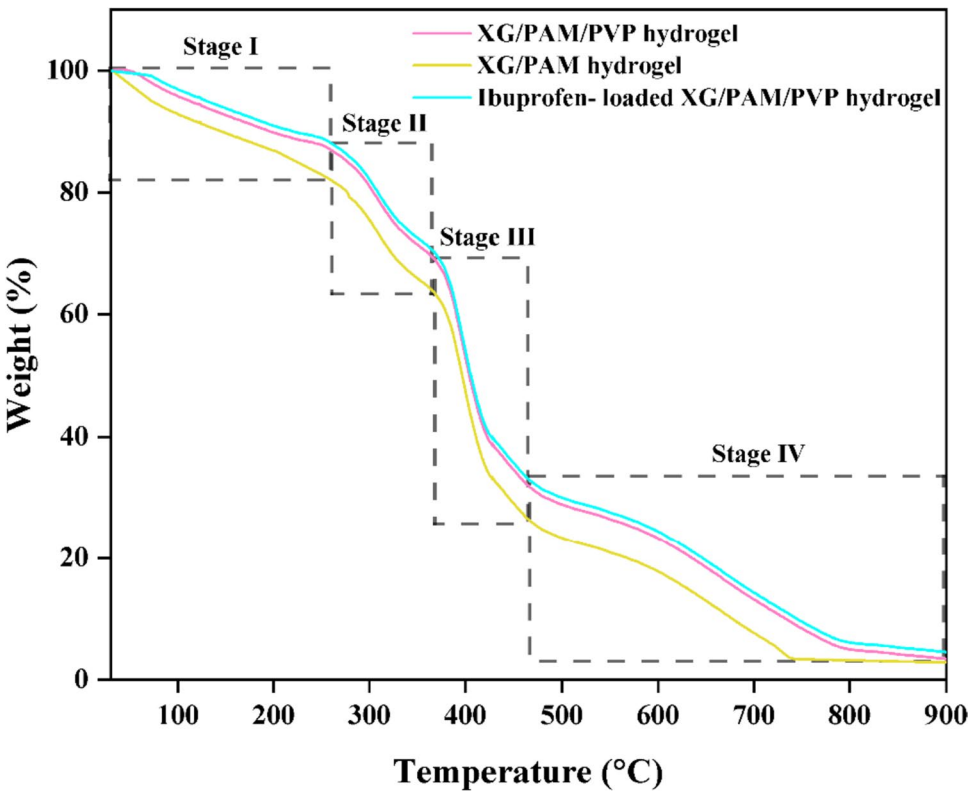


Fig. 8 Thermogram of XG/PAM hydrogel, XG/PAM/PVP hydrogel, and ibuprofen-loaded XG/PAM/PVP hydrogel

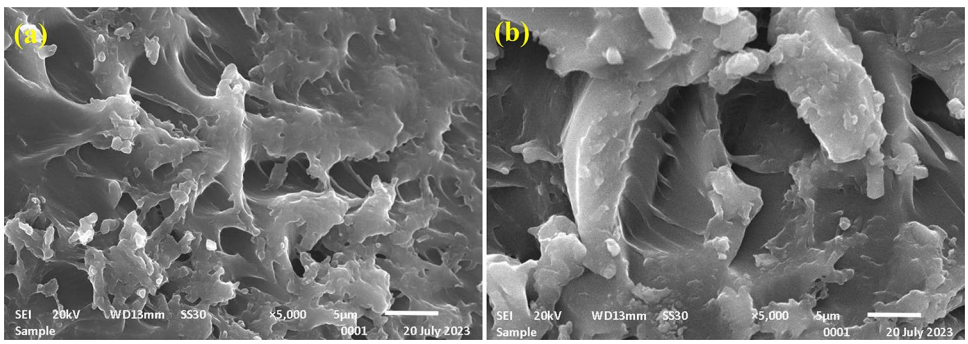


Fig. 9 SEM micrographs of **a** XG/PAM/PVP hydrogel and **b** ibuprofen-loaded XG/PAM/PVP hydrogel

Table 3 Drug loading (DL%) and entrapment efficiency (DEE%) of ibuprofen-loaded XG/PAM hydrogel and ibuprofen-loaded XG/PAM/PVP hydrogel

Hydrogel	Drug loading (DL%)	Drug entrapment efficiency (DEE%)
Ibuprofen-loaded XG/PAM	16.53	55.42
Ibuprofen-loaded XG/PAM/PVP	21.38	62.27

possesses higher drug release (%) than ibuprofen-loaded XG/PAM hydrogel due to higher swelling that observed in the XG/PAM/PVP hydrogel. This notable difference is mainly due to the addition of the PVP, which consists of hydrophilic constituents such as C = O and C–N groups, which increases the swelling; as a result, drug release also increases. The ibuprofen release percentages were 69.3% and 80.2% in XG/PAM and XG/PAM/PVP hydrogel at pH 7.4, respectively, while 17.6% and 20.6% in XG/PAM and XG/PAM/PVP hydrogel at pH 1.2, respectively.

For both formulations, the ibuprofen release was higher at pH 7.4, due to deprotonation of carboxylic acid (–COOH) present in XG, leading to anionic repulsion between COO^- anion, causing polymeric network expansion, increased swelling, consequently higher drug release. However, at pH 1.2, –COOH group protonation led to hydrogel bonding between XG, PAM, and PVP; as a result, hydrogel shrinks, hence lesser swelling and reduced drug release [23]. Hence, fabricated XG/PAM/PVP hydrogel exhibited good potential for sustained release of hydrophobic drug ibuprofen at pH 7.4.

In comparison to a marketed ibuprofen 200 mg tablet, the synthesized ibuprofen-loaded XG/PAM/PVP hydrogel demonstrates notable differences in its release behavior. Under stomach conditions (pH 1.2), the marketed tablet releases 25.68% of ibuprofen within 2 h, while the hydrogel exhibits a slower release, with only 8.06% released in the same time. Moreover,

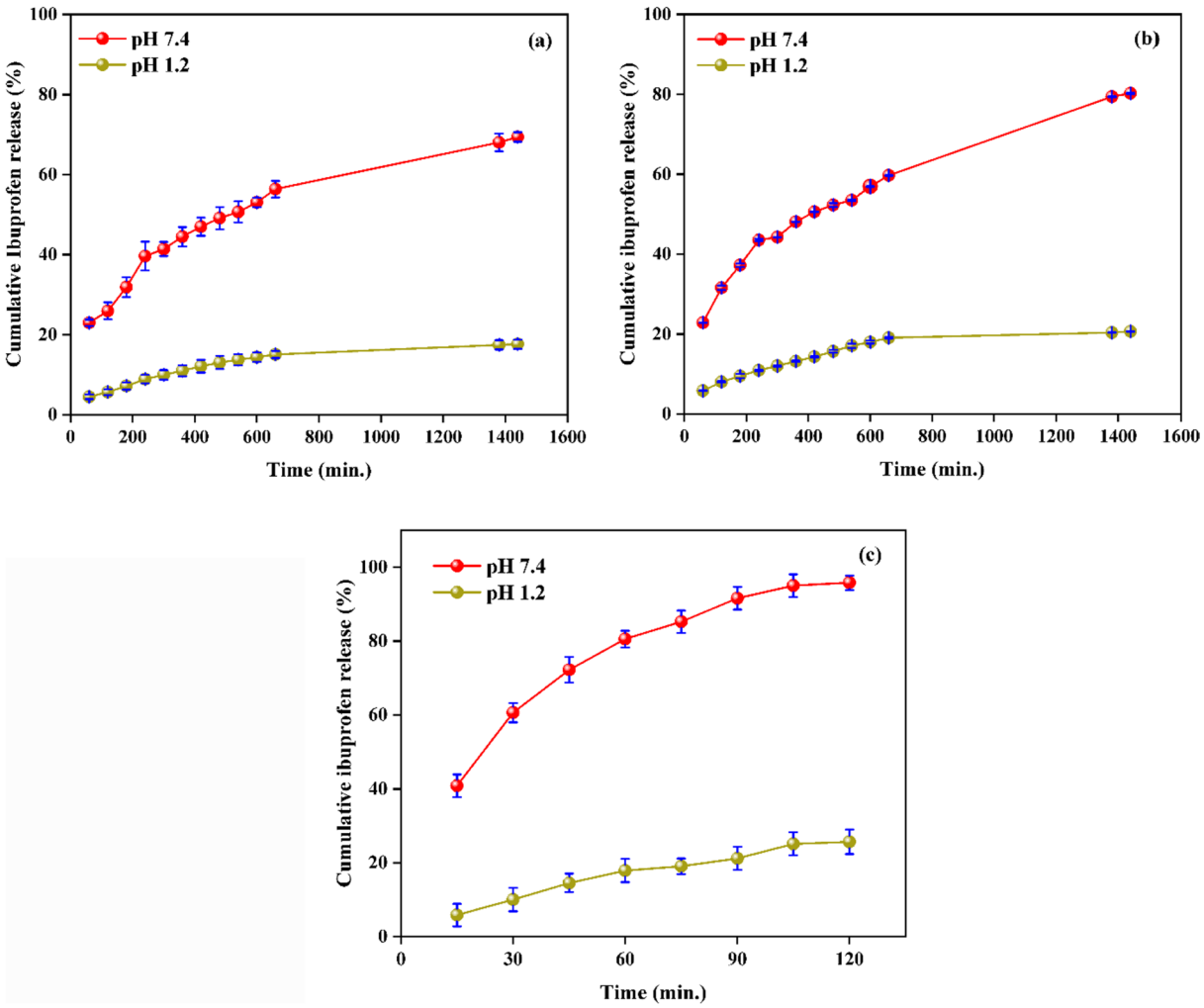


Fig. 10 Cumulative ibuprofen release from ibuprofen-loaded **a** XG/PAM hydrogel, **b** XG/PAM/PVP hydrogel, and **c** ibuprofen tablet in pH 7.4 and 1.2

Table 4 Kinetic modeling data of ibuprofen release from ibuprofen-loaded XG/PAM and XG/PAM/PVP hydrogel

Model	Equation	Ibuprofen-loaded XG/PAM hydrogel				Ibuprofen-loaded XG/PAM/PVP hydrogel				Ref
		pH 7.4		pH 1.2		pH 7.4		pH 1.2		
		n	R^2	*n*	R^2	*n*	R^2	*n*	R^2	
Zero-order	$M_t = M_\infty + k_0 t$	–	0.8650	–	0.7966	–	0.9254	–	0.7561	[48]
Higuchi	$M_t/M_\infty = k_H t^{1/2}$ k_H = kinetic constant	–	0.9780	–	0.9294	–	0.9612	–	0.8982	[46]
First-order	$\text{Log } M_t = \text{Log } M_\infty + \frac{kt}{2.303}$ k = first-order rate constant	–	0.7325	–	0.6531	–	0.7784	–	0.6421	[49]
Hixson-Crowell	$(M_t)^{1/3} - (M_\infty)^{1/3} = k_{HX} \cdot t$ k_{HX} = Hixson-Crowell constant	–	0.7800	–	0.7039	–	0.8358	–	0.6839	[49, 50]
Korsmeyer-Peppas	$M_t/M_\infty = kt^n$ k = kinetic constant n = diffusion exponent	0.3796	0.9858	0.5046	0.9737	0.4160	0.9887	0.4480	0.9997	[10]

the hydrogel continues to release ibuprofen over a 24-h period, reaching a maximum release of 20.6%. However, under intestinal conditions (pH 7.4), the marketed tablet releases 95.79% of ibuprofen within 2 h, whereas the synthesized hydrogel releases 31.61% in the same time, showcasing a more sustained release profile. Over a 24-h period, the hydrogel achieves a total release of 80.2%. These findings emphasize the hydrogel's capacity for a slower and sustained release of ibuprofen, offering potential advantages in terms of prolonged drug action and reduced risk of adverse effects associated with rapid drug release, making it a promising drug delivery system.

Kinetic modeling

Different kinetic models were utilized to analyze the release data and understand the mechanism of drug release from ibuprofen-loaded XG/PAM and XG/PAM/PVP hydrogel, as observed in Table 4. Among the first-order, Hixson-Crowell, zero-order, Higuchi, and Korsmeyer-Peppas models, the Korsmeyer-Peppas model exhibited the highest R^2 value for both hydrogels formulation, as given in Table 4. In the ibuprofen-loaded XG/PAM hydrogel, Fickian diffusion was observed at pH 7.4 with *n* value as 0.3796, while a non-Fickian diffusion was noted at pH 1.2 with *n* value of 0.5046 (Fig. 11a). Moreover, for ibuprofen-loaded XG/PAM/PVP hydrogel, the diffusion mechanism was determined to be Fickian with *n* values as 0.4480 and 0.4160 at pH 1.2 and 7.4, respectively (Fig. 11b). The Fickian diffusion indicates that drug release from the hydrogel via a diffusion process. However, the non-Fickian diffusion suggests that along with diffusion, there was also a relaxation of the polymeric chains during the drug release process.

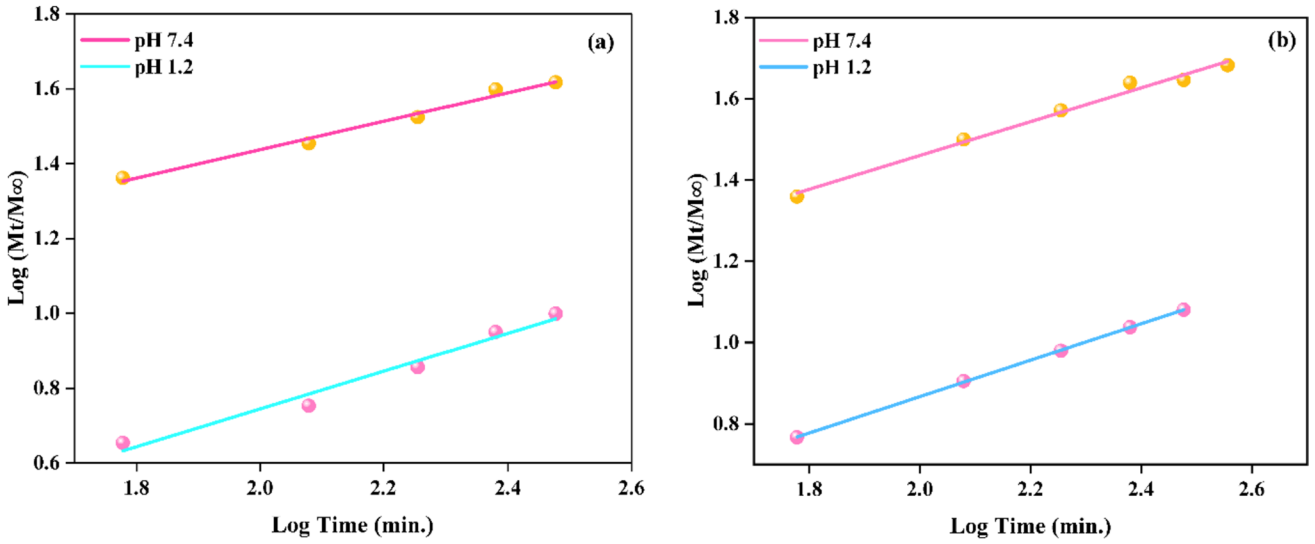


Fig. 11 Kinetic modeling profile depending on the Korsmeyer-Peppas model for ibuprofen-loaded **a** XG/PAM hydrogel and **b** XG/PAM/PVP hydrogel

Conclusion

In this work, xanthan gum-based hydrogels were successfully formulated using a free radical mechanism. The hydrogel exhibited excellent swelling properties and was consequently employed as an effective carrier for delivering the hydrophobic drug ibuprofen. In addition, the content of MBA, KPS, and XG significantly affects the gel fraction, swelling, and porosity of the synthesized hydrogels. The introduction of PVP in hydrogel results in an increase in the percentage of gel fraction, swelling, porosity, and drug loading efficiency with XG/PAM/PVP showing higher drug loading efficiency in comparison to XG/PAM hydrogel. The XG/PAM/PVP hydrogel possesses high swelling and drug release at pH 7.4, in comparison to pH 1.2, hence suitable for targeted drug delivery in the colon region of the gastrointestinal tract. The drug release mechanism for ibuprofen-loaded XG/PAM/PVP hydrogels was determined to be non-Fickian at both pH (1.2 and 7.4) while ibuprofen-loaded XG/PAM hydrogel follows Fickian diffusion at pH 7.4 and non-Fickian diffusion at pH 1.2. In conclusion, the synthesized pH-sensitive hydrogel acts as a potential carrier for efficient and prolonged delivery of the hydrophobic drug, ibuprofen.

Acknowledgements The authors Priyanka Meena, Poonam Singh, and Sudhir G. Warkar thank the Delhi Technological University, New Delhi, India, for its infrastructural facilities and Priyanka Meena also thanks the Council of Scientific & Industrial Research (CSIR) for providing senior research fellowship.

Author contribution Conceptualization: SGW, PM; methodology: PM; formal analysis and investigation: PM; writing—original draft preparation: PM; writing—review and editing: SGW, PS; resources: PM; supervision: SGW, PS.

Availability of data and materials The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical approval Not applicable.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Nguyen HT, Nguyen LT, Ha AC, Huynh PD (2023) Evaluation of ibuprofen prolonged release of biomedical PLA-PEG-PLA hydrogel via degradation mechanism. 2023. https://doi.org/10.1155/2023/5005316
2. Shinde AK et al (2019) Enhancement solubility and dissolution rate of paracetamol and ibuprofen by coamorphous particles using microwave technique. Asian J Pharm Clin Res 12:155–162. https://doi.org/10.22159/ajpcr.2019.v12i11.34589
3. Sami AJ et al (2018) Sustained release of ibuprofen by a novel formulated hydrogel containing molecular pharmaceutics & organic process research sustained release of ibuprofen by a novel formulated hydrogel containing graphene oxide. Mol Pharm Org Process Res 6:1000143. https://doi.org/10.4172/2329-9053.1000143
4. Wang XH, Su T, Zhao J et al (2020) Fabrication of polysaccharides-based hydrogel films for transdermal sustained delivery of ibuprofen. Cellulose 27:10277–10292. https://doi.org/10.1007/s10570-020-03503-0
5. Suhail M, Hsieh YH, Khan A et al (2021) Preparation and in vitro evaluation of aspartic/alginic acid based semi-interpenetrating network hydrogels for controlled release of ibuprofen. Gels 7. https://doi.org/10.3390/gels7020068
6. Kulkarni RV, Sa B (2008) Evaluation of pH-sensitivity and drug release characteristics of (polyacrylamide-grafted-xanthan)-carboxymethyl cellulose-based pH-sensitive interpenetrating network hydrogel beads. Drug Dev Ind Pharm 34:1406–1414. https://doi.org/10.1080/03639040802130079
7. Hanna DH, Saad GR (2019) Encapsulation of ciprofloxacin within modified xanthan gum- chitosan based hydrogel for drug delivery. Bioorg Chem 84:115–124. https://doi.org/10.1016/j.bioorg.2018.11.036
8. Sethi S, Saruchi KBS et al (2020) Cross-linked xanthan gum–starch hydrogels as promising materials for controlled drug delivery. Cellulose 27:4565–4589. https://doi.org/10.1007/s10570-020-03082-0
9. Malik NS, Ahmad M, Minhas MU et al (2020) Chitosan/xanthan gum based hydrogels as potential carrier for an antiviral drug: fabrication, characterization, and safety evaluation. Front Chem 8:1–16. https://doi.org/10.3389/fchem.2020.00050
10. Tushar, Saraswat Y, Meena P, Warkar SG (2023) Synthesis and characterization of novel xanthan gum-based pH-sensitive hydrogel for metformin hydrochloride release. Colloid Polym Sci. https://doi.org/10.1007/s00396-023-05135-9
11. Bhattacharyya R, Chowdhury P (2021) Hydrogels of acryloyl guar gum-g-(acrylic acid-co-3sulfopropylacrylate) for high-performance adsorption and release of gentamicin sulphate. J Polym Res 28. https://doi.org/10.1007/s10965-021-02633-8
12. Sutar PB, Mishra RK, Pal K, Banthia AK (2008) Development of pH sensitive polyacrylamide grafted pectin hydrogel for controlled drug delivery system. J Mater Sci Mater Med 19:2247–2253. https://doi.org/10.1007/s10856-007-3162-y
13. Reddy PRS, Rao KM, Rao KSVK et al (2014) Synthesis of alginate based silver nanocomposite hydrogels for biomedical applications. Macromol Res 22:832–842. https://doi.org/10.1007/s13233-014-2117-7
14. Rizwan M, Yahya R, Hassan A et al (2017) pH sensitive hydrogels in drug delivery: brief history, properties, swelling, and release mechanism, material selection and applications. Polymers (Basel) 9. https://doi.org/10.3390/polym9040137
15. Ahmad U, Sohail M, Ahmad M et al (2019) Chitosan based thermosensitive injectable hydrogels for controlled delivery of loxoprofen: development, characterization and in-vivo evaluation. Int J Biol Macromol 129:233–245. https://doi.org/10.1016/j.ijbiomac.2019.02.031
16. Munim SA, Raza ZA (2019) Poly(lactic acid) based hydrogels: formation, characteristics and biomedical applications. J Porous Mater 26:881–901. https://doi.org/10.1007/s10934-018-0687-z
17. Hosseinzadeh H (2010) Controlled release of diclofenac sodium from pH-responsive carrageenan-g-poly(acrylic acid) superabsorbent hydrogel. J Chem Sci 122:651–659. https://doi.org/10.1007/s12039-010-0100-1
18. García-Verdugo KF, Ramírez-Irigoyen AJ, Castillo-Ortega M et al (2022) A pH/temperature-sensitive s-IPN based on poly(vinyl alcohol), poly(vinyl methyl ether-alt-maleic acid) and poly(vinyl methyl ether) prepared by autoclaving. Macromol Res 30:353–364. https://doi.org/10.1007/s13233-022-0044-6

19. Raza MA, Park SH (2020) Irradiated Ch/GG/PVP-based stimuli-responsive hydrogels for controlled drug release. J Appl Polym Sci 137. https://doi.org/10.1002/app.49041
20. Thirupathi K, Phan TTV, Santhamoorthy M et al (2022) pH and thermoresponsive PNIPAm-co-polyacrylamide hydrogel for dual stimuli-responsive controlled drug delivery. Polymers (Basel) 15:167. https://doi.org/10.3390/polym15010167
21. Martinez-Ruvalcabal A et al (2009) Swelling characterization and drug delivery kinetics of polyacrylamide-co-itaconic acid/chitosan hydrogels 3:25–32. https://doi.org/10.3144/expresspolymlett.2009.5
22. Musa H, Yahaya S (2022) Polyacrylamide hydrogels for application in oral drug delivery. Niger J Sci Res
23. Wei QB, Fu F, Zhang YQ et al (2014) PH-responsive CMC/PAM/PVP semi-IPN hydrogels for theophylline drug release. J Polym Res 21. https://doi.org/10.1007/s10965-014-0453-0
24. Meena P, Singh P, Warkar SG (2023) Development and assessment of carboxymethyl tamarind kernel gum-based pH-responsive hydrogel for release of diclofenac sodium. Eur Polym J 197:112340. https://doi.org/10.1016/j.eurpolymj.2023.112340
25. Lim DK, Cui MH, Nam JM (2011) Highly stable, amphiphilic DNA-encoded nanoparticle conjugates for DNA encoding/decoding applications. J Mater Chem 21:9467–9470. https://doi.org/10.1039/c1jm11150g
26. Afriani K, Sutanti Budikania T (2020) Synthesis and characterization of hydrogel of chitosan-poly(N-vinyl-2-pirrolidone) (PVP)-alginate for ibuprofen release. J Pure Appl Chem Res 9:201–211. https://doi.org/10.21776/ub.jpacr.2020.009.03.558
27. Chauhan D, Kumar A, Warkar SG (2022) Synthesis, characterization and metal ions sensing applications of *meta*-benziporphodimethene-embedded polyacrylamide/carboxymethyl guargum polymeric hydrogels in water. Environ Technol 43:991–1002. https://doi.org/10.1080/09593330.2020.1812730
28. Tan LS, Tan HL, Deekonda K et al (2021) Fabrication of radiation cross-linked diclofenac sodium loaded carboxymethyl sago pulp/chitosan hydrogel for enteric and sustained drug delivery. Carbohydr Polym Technol Appl 2:100084. https://doi.org/10.1016/j.carpta.2021.100084
29. Rahmani S, Olad A, Rahmani Z (2022) Preparation of self-healable nanocomposite hydrogel based on gum arabic/gelatin and graphene oxide: study of drug delivery behavior. Polym Bull 80:4117–4138. https://doi.org/10.1007/s00289-022-04247-6
30. Naeem F et al (2017) pH responsive cross-linked polymeric matrices based on natural polymers: effect of process variables on swelling characterization and drug delivery properties. Bio-Impacts 7:177–192. https://doi.org/10.15171/bi.2017.21
31. Kalantari K, Mostafavi E, Saleh B et al (2020) Chitosan/PVA hydrogels incorporated with green synthesized cerium oxide nanoparticles for wound healing applications. Eur Polym J 134:109853. https://doi.org/10.1016/j.eurpolymj.2020.109853
32. Kumar PTS et al (2014) Flexible, micro-porous chitosan-gelatin hydrogel/nanofibrin composite bandages for treating burn wounds. RSC Adv 4:65081–65087. https://doi.org/10.1039/c4ra11969j
33. Khan R, Zaman M, Salawi A et al (2022) Synthesis of chemically cross-linked pH-sensitive hydrogels for the sustained delivery of ezetimibe. Gels 8. https://doi.org/10.3390/gels8050281
34. Safari JB, Bapolisi AM, Krause RWM (2021) Development of pH-sensitive chitosan-g-poly(acrylamide-co-acrylic acid) hydrogel for controlled drug delivery of tenofovir disoproxil fumarate. Polymers (Basel) 13. https://doi.org/10.3390/polym13203571
35. Rashid H, Ahmad M, Minhas MU et al (2015) Synthesis and characterization of poly(hydroxyethyl methacrylate-co-methacrylic acid) cross linked polymeric network for the delivery of analgesic agent. J Chem Soc Pakistan 37:999–1007
36. Tulain UR, Ahmad M, Rashid A et al (2018) Fabrication of pH-responsive hydrogel and its in vitro and in vivo evaluation. Adv Polym Technol 37:290–304. https://doi.org/10.1002/adv.21668
37. Niu B, Jia J, Wang H et al (2019) In vitro and in vivo release of diclofenac sodium-loaded sodium alginate/carboxymethyl chitosan-ZnO hydrogel beads. Int J Biol Macromol 141:1191–1198. https://doi.org/10.1016/j.ijbiomac.2019.09.059
38. Jana S, Sharma R, Maiti S, Sen KK (2016) Interpenetrating hydrogels of O-carboxymethyl tamarind gum and alginate for monitoring delivery of acyclovir. Int J Biol Macromol 92:1034–1039. https://doi.org/10.1016/j.ijbiomac.2016.08.017
39. Varahala Setti ML, Ratna JV (2009) Preparation and evaluation of controlled release tablets of carvedilol. Asian J Pharm 3:252–256. https://doi.org/10.4103/0973-8398.56307
40. Malik R, Warkar SG, Saxena R (2023) Carboxy-methyl tamarind kernel gum based bio-hydrogel for sustainable agronomy. Mater Today Commun 35:105473. https://doi.org/10.1016/j.mtcomm.2023.105473
41. Mali KK, Dhawale SC, Dias RJ (2017) Synthesis and characterization of hydrogel films of carboxymethyl tamarind gum using citric acid. Int J Biol Macromol 105:463–470. https://doi.org/10.1016/j.ijbiomac.2017.07.058
42. Suhail M, Khan A, Rosenholm JM et al (2021) Fabrication and characterization of diclofenac sodium loaded hydrogels of sodium alginate as sustained release carrier. Gels 7:1–16. https://doi.org/10.3390/gels7010010
43. Nawaz S, Khan S, Farooq U et al (2018) Biocompatible hydrogels for the controlled delivery of anti-hypertensive agent: development, characterization and in vitro evaluation. Des Monomers Polym 21:18–32. https://doi.org/10.1080/15685551.2018.1445416
44. Elzayat EM, Abdel-Rahman AA, Ahmed SM et al (2016) Studying the impact of formulation and processing parameters on the release characteristics from hydroxypropyl methylcellulose matrix tablets of diclofenac. Acta Pol Pharm - Drug Res 73:439–452
45. Abdelkader R, Mohammed B (2016) Green synthesis of cationic polyacrylamide composite catalyzed by an ecologically catalyst clay called maghnite-H+ (Algerian MMT) under microwave irradiation. Bull Chem React Eng Catal 11:170–175. https://doi.org/10.9767/bcrec.11.2.543.170-175
46. Khushbu Warkar S G, Thombare N (2022) Controlled release and release kinetics studies of boron through the functional formulation of carboxymethyl tamarind kernel gum-based superabsorbent hydrogel. Polym Bull 79:2287–2303. https://doi.org/10.1007/s00289-021-03634-9
47. Ekici S, Saraydin D (2007) Interpenetrating polymeric network hydrogels for potential gastrointestinal. Polym Int 56:1371–1377. https://doi.org/10.1002/pi.2271
48. Goonoo N et al (2014) Naltrexone: a review of existing sustained drug delivery systems and emerging nano-based systems. J Control Release 183:154–166. https://doi.org/10.1016/j.jconrel.2014.03.046
49. Das D, Das R, Mandal J et al (2014) Dextrin crosslinked with poly(lactic acid): a novel hydrogel for controlled drug release application. J Appl Polym Sci 131:1–12. https://doi.org/10.1002/app.40039
50. Kajal KR, Meena P, Warkar SG (2023) Development and characterization of pH-responsive CMTKG/PAM/PEG hydrogel for oral administration of etophylline. Colloid Polym Sci. https://doi.org/10.1007/s00396-023-05152-8

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.

Frugal Design: An Empirical Conceptualization and Measurement

Anuradha Kumari
Research Scholar
Department of Design, Delhi Technonological University,
Delhi, India
anuradha_2k19phddes02@dtu.ac.in

Dr. Ravindra Singh,
Assistant Professor,
Department of Design, Delhi Technological University,
Delhi, India
ravindra@dtu.ac.in

Prof. Lalit Kumar Das,
Honorary Distinguished Faculty,
Department of Design, Delhi Technological University,
Delhi, India
lalitdas@dtu.ac.in

***Abstract*- Designers/Engineers must practice Frugal design to address rising cost-conscious consumption. However, It has been observed that the conception of frugal design and its measurement scale have not yet been created. A three-stage mixed-method research approach was used to identify and validate the frugal design measuring attributes. Stage 1 determined initial frugal design attributes using existing literature research, word frequency count analysis, and expert review content analysis. Survey research (n = 250) was undertaken in Stage 2, and a Principal Component Analysis (PCA) was carried out to refine the preliminary attributes. The final stage involved a survey with a sample of (n=500) Brazilians. To validate the final attribute for gauging frugal design, the Delphi consensus method was used). The frugal design scale provides a framework for organizations to assess frugality based on minimizing inequalities amongst diverse income groups and maximizing well-being by delivering efficient solutions for all groups of people.**

Keywords: Frugal Design, Mixed approach, Content validity, Attributes.

I. Introduction

According to the United Nations Organization (UNO), there are approximately 8 billion people on the planet, of which 648 million people in emerging nations are uncovered to acute poverty and struggle for their livelihood. It is anticipated that this population number will rise to 9.7 billion by 2050 [1]. This extremely rapid population growth will undoubtedly lead to many problems related to the overconsumption of resources and a decline in quality of living. As the demand for resources continues to rise, so will energy consumption, resulting in environmental challenges, i.e., biodiversity loss and climate change. Also, the long-term availability of resources required to maintain the expanding population may become a concern [2]. Expansion of global markets will raise the costs of essential resources and inaccessible products, causing social stratification and inequality and an unsafe environment for a living [3].

Various solutions, such as the 12R doctrine, circular economy, shared economy, etc., have been developed to tackle these burning issues, as mentioned above, but none of them appear to be effective in solving them effectively [4]. Frugal design is a comparatively new type of approach that gained significant growth in an era of scarce resources [5]. It is a concept that emerges regardless of financial, technological, human, and other resource constraints. It produces economic outcomes catering to marginalized society's fundamental needs [6]. A few examples of Frugal Design are Selco: solar energy; in India, solar energy is provided to people at the bottom of the social and economic pyramid; Mitti cool; a refrigerator that runs without battery, Voda-phone mobile: solar power mobile, M-Pesa, etc. [7].

Although frugal design has increasingly gained scholarly attention as an approach to developing affordable solutions, there is a gap in the literature regarding a comprehensive understanding of this approach [8]. The literature has mainly concentrated on grassroots movements until now when describing this process [9]. It is worth mentioning that frugal design may include attributes other than just low-cost innovation and material constraints method. A specific requirement is to investigate and validate frugal design conception from a multi-attribute aspect. Earlier investigations, such as[10], have highlighted dimensions of frugal design; however, their research on frugal design was primarily qualitative. Therefore, the question remains: what attributes are required in the early design phase for creating the frugal design? The purpose of this study is to construct and validate a measurement scale to give an appropriate response to this question. Based on the practical importance of frugal design and the discovered gaps in the literature about the attributes of frugal design, the current study tries to fill the gap by empirically conceptualizing frugal design.

The study adds a significant contribution to the existing reservoir of knowledge on frugal design. The study creates a consistent and valid multi-attribute frugal design measurement scale. Additionally, the study explains how a frugal design is conceptualized empirically from customers' perspectives. The present study uses a mixed-method approach with multiple stages and a variety of methodologies to investigate and validate the assessment of frugal design. Finally, defining the

attributes can help designers and engineers develop or enhance their frugal design process at an early stage of design to benefit from frugality-based advantages.

II. Process of Frugal Design Scale Development

The present study used a mixed-method approach to construct a reliable and accurate scale to assess frugal design in the industry and companies, following the scale development process [11]. A three-stage research method with different studies included in each stage is shown in Fig. 1.

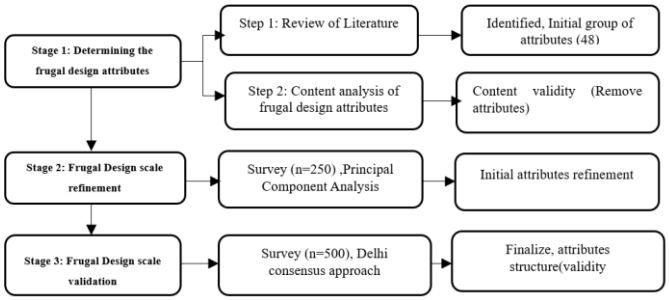


Fig. 1. Process for Frugal Design Scale Development

In stage 1, a preliminary group of measuring attributes of frugal design was identified. The study at this stage integrated inclusion and exclusion criteria [12]. A critical analysis of the literature on frugal design, in general, served as the foundation for the deductive method. Consequently, frugal attributes were discovered, and their indicators were selected depending on an adequate definition of the conceptual areas [13]. These indicators were derived from scales already existing in the literature. Furthermore, at this stage, inductive methods were used on qualitative data derived from a frugal design expert group assessment, preliminary analysis, and content analysis; in stage 2, The measurement was refined by utilizing reliability testing and principle component analysis (PCA) and the quantitative survey. In stage 3, the measure was verified using a Delphi consensus approach in an expanded quantitative survey.

III. Stage 1: Determining the frugal design attributes

A. *Step 1: Review of Literature*

The study conducts a systematic literature review, which covers several studies done in the past. The literature was analyzed according to methodologies, trending topics, and industrial and geographical contexts. The initial outcomes from searching for "Frugal Innovation," "Frugal Design," and "Frugality" in the database, a total of 511 articles were retrieved, i.e., ABI/INFORM Complete 63, Emeralds 49, EBSCO 32, IEEE Explore 15, Inder Sciences 2, Scopus 44, Science Direct 99, Sage Premier 30, Taylor & Francis 70, Web of Science 63, and Wiley 44. The study focused on peer-reviewed publications to achieve the frugal design definition and higher levels of quality control. The article searches were limited to 2000-2023 because 2000 was the first year that the term frugal design was introduced in the literature. The articles that were chosen were written in English to prevent erroneous interpretations based on the inclusion and exclusion criteria. 149 duplicate articles from well-known databases were eliminated throughout the search process. Furthermore, 43 out of 362 remaining articles were removed because they were outside our assessment's scope. A total of 319 articles have been found in the literature.

The word-counting Software Sobolsoft https://www.sobolsoft.com was utilized *for* word-frequency count. For each attribute, the article-specific frequency percentage was computed. This task was carried out in the following parts.

1. A total 4,02,776 no of words were extracted from all the published articles related to design.
2. A list of the 1,294 most commonly used attributes related to design was selected.
3. To condense the database cluster, remove the attributes that are not directly related to the design. The attributes that did not belong to the product feature or specification were excluded (e.g., potential, perspective, validate, etc.). Similar attributes were eliminated, and finally, the attributes whose word frequency was less than five were eliminated. Finally, the forty-eight attributes were generated to develop the initial list of Frugal Design attributes, as shown in Table 1.

B. *Step 2: Content analysis of frugal design attributes*

The content validity analysis method was carried out to validate the above attributes of Frugal Design[14]. This pretest was designed to find and eliminate any unclear or missing information. A group of ten experts (professors) with similar backgrounds (i.e.,

design) evaluated the attributes using a four-point Likert scale, where 1 and 4 represent not essential and highly essential, respectively—further calculated the Item-Content Validity Index (I-CVI) and Content Validity Ratio (CVR) using equation no 1.

$$CVR = \frac{Ne - N/2}{N/2} \quad (1)$$

N denotes the overall number of experts, whereas N_e represents the number of experts who consider each attribute essential.

Thirty-nine attributes were deemed relevant for frugal design as they had a high CVR (>0.62) and an I-CVI score (>0.8). Other attributes, i.e., manufacturing, marginalized, disruptive, socially, energy efficient, and Desirability, were removed as they had a low CVR (<0.62) and I-CVI score (<0.8), indicating that the majority of experts did not consider them relevant to frugal design [15] (Lawshe, 1975; Zeraati and Alavi, 2014) as shown in Fig. 2

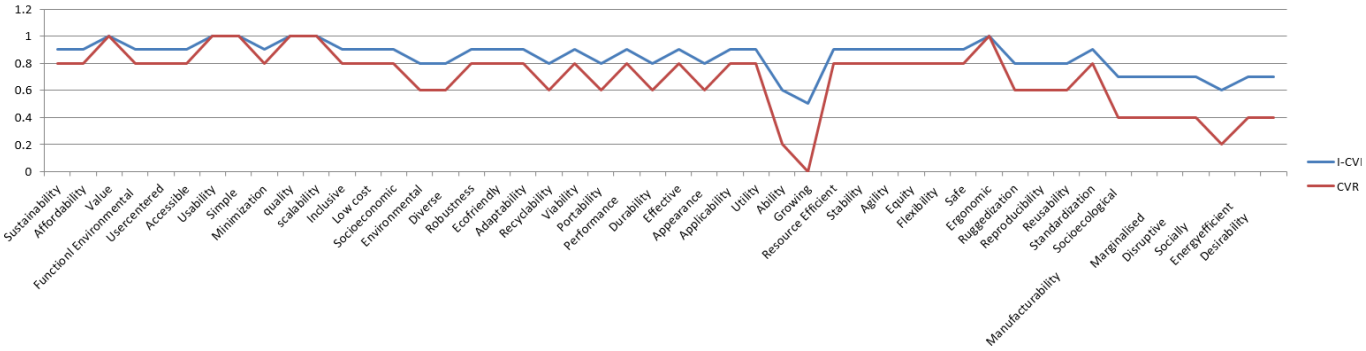


Fig. 2. Item Content Validity Index (I-CVI) and Content Validity Ratio (CVR) and of attributes.

IV. Stage 2: Frugal Design scale refinement

This stage aims to refine the database of attributes identified from the previous steps. A questionnaire survey was carried out from customers from India, since it is a developing nation, there are institutional gaps, resource limitations, and a substantial number of customers at the bottom of the pyramid (BOP). The researcher provided the participants with examples of frugal products before distributing the questionnaire; participants were expected to fully understand the frugal design.

The first part of the survey questionnaire was intended to gauge the participant's knowledge of the subject (Frugal Design). The opinions from participants were gathered in the next section using five-point Likert scales to the identified attributes of frugal design. Due to incomplete sdata and disengaged responses, 23 observations were dropped during data cleaning. In total, 250 participants constituted the sample for this study, achieving a valid response rate of 91.57%. We performed a conducted using principal component analysis (PCA) on the set of 39 attributes to look at the attributes' structure and perform a preliminary evaluation of the validity and reliability of the scales [16] (Joshi et al., 2015). Further, the obtained attributes were standardized using equation no 2.

$$Z_i = \frac{Xi - X}{S} \quad (2)$$

Where X and S are the Mean and Standard Deviation of the attribute X_i and Z_i is the corresponding transformed attribute.

Principal Component Analysis (PCA) identifies the final list of Frugal Design attributes. PCA is a dimensional reduction technique that analyses the attributes' covariance and determines the important attributes for characterizing the Frugal Design. It captures the maximum amount of variation in data and transforms the existing dimension into a smaller space [17]. Table 1 lists the attributes and their eigenvalues, arranged in decreasing order. Also, it indicates each attribute's strength (in percentage) in expressing the variation in data. From this, we can observe that twelve attributes obtained a cumulative strength of 100%. Sustainability is the most important attribute as it provides the maximum strength, 20.821%, in capturing the variation in data.

TABLE 1. List of most frequently used attributes of Frugal Design and Eigen value of the selected attributes

Attributes	Frequency	Frequency %	% of Variance	Cumulative %	Attributes	Frequency	Frequency %	% of Variance	Cumulative %
Sustainability	1305	24.40	20.821	20.821	Portability	23	0.43	0.662	98.132
Affordability	611	11.42	12.547	33.367	Performance	21	0.39	0.498	98.630
Value	358	6.69	9.165	42.532	Durability	17	0.31	0.466	99.095
Inclusive	305	5.70	2.819	85.233	Effective	17	0.31	0.394	99.489

Simple	261	4.88	4.452	72.264	Appearance	23	0.43	0.214	99.703
Accessibility	385	7.20	5.654	62.911	Applicability	14	0.26	0.123	99.826
Functional	135	2.52	7.768	50.301	Resource efficient	11	0.20	0.077	100.000
User-Centered	155	2.89	6.955	57.256	Stability	11	0.20	0	100.000
Usability	215	4.02	4.901	67.811	Agility	19	0.35	0	100.000
Quality	109	2.03	3.461	79.318	Equity	8	0.14	0	100.000
Minimization	132	2.46	3.594	75.858	Flexibility	8	0.14	0	100.000
Scalability	191	3.57	3.096	82.414	Safe	9	0.16	0	100.000
Environmental	280	5.23	1.692	91.241	Ergonomic	8	0.14	0	100.000
Low-cost	118	2.20	2.350	87.584	Ruggedization	8	0.14	0	100.000
Utility	74	1.38	0.097	99.923	Reproducibility	7	0.13	0	100.000
Ability	71	1.32	-	-	Reusability	7	0.13	0	100.000
Growing	69	1.29	-	-	Standardization	8	0.14	0	100.000
Socioeconomic	67	1.25	1.965	89.549	Socioecological	8	0.14	-	-
Recyclability	62	1.15	0.831	96.759	Manufacturability	8	0.14	-	-
Diverse	57	1.06	1.405	92.646	Marginalized	5	0.09	-	-
Robustness	42	0.78	1.272	93.918	Disruptive	7	0.13	-	-
Eco-friendly	36	0.67	1.050	94.969	Socially	5	0.09	-	-
Adaptability	26	0.48	0.959	95.928	Energy efficient	6	0.11	-	-
Viability	18	0.33	0.711	97.470	Desirability	7	0.13	-	-

The Scree Plot (Fig. 3) represented the results, as listed in Table 1, graphically. The eigenvalues are shown on the y-axis, and the attributes are on the horizontal x-axis. This graph considered the attributes up to the twelfth as eigenvalues of these attributes are higher than one as per Kaiser-Meyer-Olkin criteria [18].

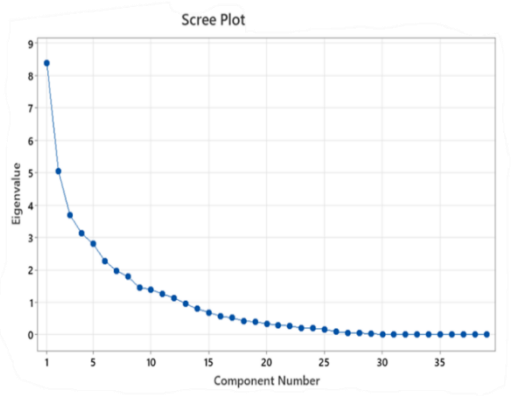


Fig. 3. Screen plot of Principal components and Eigenvalue

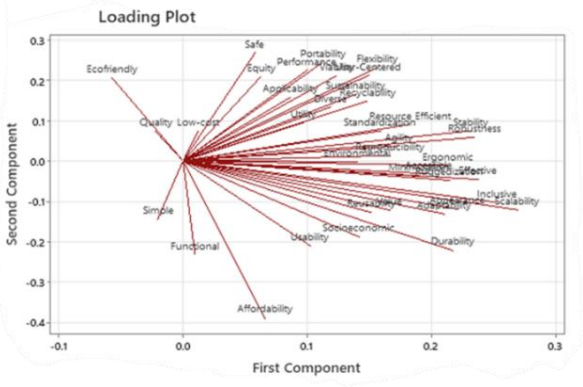


Fig. 4. Loading Plot of Highly Influenced Attributes

Loading plot helps to determine which attributes have the most significant influence on each other. Attributes loadings near -1 or 1 indicate a substantial impact on the component. If the loading is near 0, the variable weakly influences the component, as shown in Fig 4. According to this loading plot, the attributes Sustainability, Affordability, Value, Functional, User-Centered, Accessible, Usability, Simple, Minimization, Quality, Scalability, and Inclusive have a strong influence and positive correlation with each other, which indicates that these attributes primarily contribute to Frugal Design.

The study identified the twelve attributes of frugal design (Sustainability, Affordability , value, functionality, User-centered design, accessibility, usability, simplicity, and minimization. Quality and scalability These attributes work together to develop true frugal designs accessible by a wide range of people.

V. Stage 3: Frugal Design scale validation

A. Qualitative Assessment of Scale

The objective of this study was to validate the scale developed in the previous stage objectively. To prevent any discrepancies from outside effects, the refined scale is proposed to be re-evaluated with data gathered from a new sample in the final stage of the multi-attributes scale development process [19]. As a result, we specifically chose various samples during both of the stages.

Furthermore, the Delphi consensus approach was used in three rounds to test the frugal design features on customers in five Brazilian cities (Recife, Brasilia, Guarulhos, Manaus, and Salvador)[20]. In the first round of the Delphi process, the criteria were first evaluated by being

marked as either "important" or "not important." A 5-point Likert-type scale (5 = highly important to 1 = not important) was used in Delphi round two to rate the attribute's importance. In the final round, The Delphi process validated the frugal design attribute's importance level using a 5-point Likert-type scale. The study obtained 500 valid responses at the end of the data collection.

Table 2 shows the results of the Delphi consensus approach; some participants in the first round of the Delphi approach examined some attributes of the frugal design scale as not important; Nevertheless, the majority of the participants believed that all of the improved attributes needed to be included in the validation process in Delphi Round 2 to determine the degree of relevance and affirm the level of importance in the final round of the Delphi approach. Sustainability was rated important by 480 participants (96%). Comparable to 495 participants (99%) thought the affordability attribute was important, etc. Three attributes were determined to be unimportant in the second and third rounds of the Delphi approach (User-Centered, $\bar{x}$ = 2.74, Accessible, $\bar{x}$ = 2.83, minimization $\bar{x}$ = 2.31). All of the attributes that participants deemed to be "important" fulfilled the consensus principles ($\bar{x}$ > 3 and consensus rate > 75%), according to the Delphi consensus results. According to the Delphi consensus results, all the criteria assessed as "important" by participants have met the consensus principles (mean value > 3 and consensus rate > 75%) [21]. As a result, the Delphi consensus technique produced nine legitimate attributes: sustainability, affordability, value, functionality, usability, simplicity, quality, scalability, and inclusivity.

TABLE 2: Frugal design scale attributes validation by Delphi consensus approach.

Validated Attributes	**Delphi round one**		**Delphi round two**		**Delphi round three**		**Delphi Consensus**	
	N	**%**	$\bar{x}$	σ	$\bar{x}$	σ	%	C
Sustainability	480	96	4.05	0.442	4.36	0.48	96.65	✓
Affordability	495	99	4.14	0.436	4.414	0.493	100.00	✓
Value	415	83	3.34	0.456	4.120	0.32	91.30	✓
Functional	470	94	3.91	0.91	4.23	0.5079	95.65	✓
User-Centered	285	57	2.63	0.76	2.74	0.61	×	×
Accessible	240	48	2.29	1.02	2.83	0.48	×	×
Usability	385	77	3.36	0.74	4.120	0.3255	86.96	✓
Simple	465	93	3.91	0.91	3.9879	0.765	93.65	✓
Minimization	255	51	2.53	0.61	2.91	0.41	×	×
Quality	390	78	3.28	0.58	3.6092	0.4884	94.33	✓
Scalability	410	82	3.33	0.79	4.082	0.321	86.83	✓
Inclusive	430	86	3.49	1.24	4.36	0.48	96.65	✓

VI. Discussion and Conclusion

A. *Discussion*

The findings of the study added to the body of literature in numerous ways. First, the study fills a research gap by identifying the frugal design attributes [22]. Our study is the first to construct a multi-attribute frugal design scale from a consumer standpoint, using extensive analytical techniques and a variety of samples. The study's findings show nine attributes of frugal design: Sustainability, Affordability, Value, Functional, Usability, Simple, Quality, Scalability, and Inclusive. The current study differs from earlier studies on frugal design as it also considers sustainability and value, inclusiveness, and usability attributes. The inclusion of these attributes in a product has a favorable effect on consumers' buying decisions [23], their adoption of the product [24], and their perception of the brand [25]. Second, our study found evidence that frugal design influences consumers' opinions toward utilizing frugal products as well as their decision to purchase these kinds of goods [26]. This empirical research contributes to the field of frugal design scale development research, which has typically not evaluated the predictive validity of generated scales [22]. Additionally, frugal design attributes can help organizations tailor their products according to regional and context preferences and allow them to be accepted at a mainstream level. Finally, the study provides a framework for organizations to assess frugality based on minimizing inequalities amongst diverse income groups and maximizing well-being by providing efficient solutions for all. Frugal design attributes can be viewed as an essential step toward achieving the Sustainable Development Goals of the United Nations. Sustainability encourages conscientious utilization and development of renewable energy sources. Additionally, it highlights the importance of promoting sustainable production and consumption practices that decrease waste, save resources, and mitigate the negative environmental impact of human activities.

B. *Conclusion*

Researchers are increasingly interested in understanding and developing the Frugal Design to identify and alleviate problems efficiently and effectively. It remains a challenge to empirically conceptualize and define Frugal Design. This study

aimed to bridge that gap by developing a frugal design measurement scale. According to our study, a frugal design should have nine attributes: Sustainability, Affordability, Value, Functional, Usability, Simple, Quality, Scalability, and Inclusive. Thus, these attributes serve as a blueprint for producing affordable and accessible products for all sets of users and conserving resources throughout their lifetime. Finally, the findings also contribute to sustainable development by developing the fugal design measurement scale, potentially addressing social, ecological, and economic sustainability goals simultaneously to develop low-cost products, utilize fewer resources, and improve physical and social needs. Frugal design holds significant potential for the future, particularly when combined with sustainable development goals and parameters such as People, Planet, Peace, Partnership, and Prosperity. By incorporating these parameters, Frugal Design can be further strengthened, offering even greater benefits and opportunities for the future.

Reference

[1] United Nations Department of Economic and Social Affairs, Population Division. "World Population Prospects 2022: Summary of Results. UN DESA/POP/2022/TR/NO. 3, 2022.

[2] X. Xing, J. Wang,T. Liu,H. Liu. And Y. Zhu,"How energy consumption and pollutant emissions affect the disparity of public health in countries with high fossil energy consumption," International Journal of Environmental Research and Public Health, 16(23), p.4678,2019.

[3] Y. Ding, "Impacts of affluence and overexploitation of natural resources," Environment and Development,1-20, 2003.

[4] G.V.A., Vasantha, R. Roy, A. Lelah, and D. Brissaud. "A review of product–service systems design methodologies," Journal of Engineering Design, 23, no. 9: 635-659, 2012

[5] P. Soni, and T. Krishnan.. "Frugal innovation: aligning theory, practice, and public policy," Journal of Indian Business Research, 6, no. 1: 29-47, 2014

[6] D. Roiland, "Frugality, a positive principle to promote sustainable development," Journal of Agricultural and Environmental Ethics, 29: 571-585, 2016.

[7] C. Horn, and B. Alexander. "Strategic directions on innovation management–a conceptual framework," Management research review, *36*(10), pp.939-954.2013.

[8] P. Upadhyay, and R. M. Punekar, "A framework for designing frugal innovations in marginalised contexts." Journal of Cleaner Production, 410, p. 137170, 2023

[9] A. Brem, C. Wimschneider, A. R. de Aguiar Dutra, A. L. V. Cubas, and R. D. Ribeiro, "How to design and construct an innovative frugal product? An empirical examination of a frugal new product development process," Journal of Cleaner Production 275, p. 122232, 2020

[10] D.E. Rossetto., F.M. Borini, R.C. Bernardes, and G.L Frankwick, "Measuring frugal innovation capabilities: An initial scale proposition,"Technovation, 121, p.102674,2023

[11] G.A. Churchill Jr, " A paradigm for developing better measures of marketing constructs," Journal of marketing research, 16(1), pp.64-73, 1979.

[12] C.M. Patinoand, , J.C. Ferreira, " Inclusion and exclusion criteria in research studies: definitions and why they matter," Jornal Brasileiro de Pneumologia, 44, pp.84-84,2018.

[13] S. Tesic, B. Bajic, I. Cosic and A. Rikalovic , " Frugal innovation in industry 4.0: The Identification of key indicators," Annals of DAAAM & Proceedings, 10(2), 2021.

[14] F. Yaghmaie, "Content validity and its estimation," Journal of medical education, 3(1), 2003.

[15] C. H. Lawshe, "A quantitative approach to content validity," Personnel psychology, 28(4), 563-575. 1975.

[16] A. Joshi, S. Kale, S. Chandel, and D. K. Pal. "Likert scale: Explored and explained," British journal of applied science & technology 7(4),pp, 396,2015.

[17] I.T Joliffe and , B.J.T. Morgan, "Principal component analysis and exploratory factor analysis,"Statistical methods in medical research, 1(1), pp.69-95, 1992

[18] H.F. Kaiser, and J. Rice. "Little jiffy, mark IV," Educational and psychological measurement 34(1),pp. 111-117, 1974.

[19] R. Farooq, "An updated paradigm for developing better measures: a review of scale development practices, "Anvesha, 10(2), 42-53, 2017

[20] L. Giannarou and E. Zervas 2014, " Using Delphi technique to build consensus in practice," International Journal of Business Science & Applied Management (IJBSAM), 9(2), pp.65-82, 2014

[21] A. Habibi, A. Sarafrazi, and S. Izadyar, "Delphi technique theoretical framework in qualitative research," The International Journal of Engineering and Science, 3(4), pp.8-13,2014.

[22] D.E. Rossetto, F.M. Borini, R.C. Bernardes and G.L. Frankwick., "Measuring frugal innovation capabilities: An initial scale proposition," Technovation, 121, p.102674.2023

[23] J. Levänen and S. Lindemann, "Frugal innovations in circular economy: exploring possibilities and challenges in emerging markets," In The international society for ecological economics 2016 conference, transforming the economy: sustaining food, water, energy and justice. Washington DC, USA.2016

[24] M. Dabić ,T. Obradović ,B. Vlačić, S. Sahasranamam and J. Paul, " Frugal innovations: A multidisciplinary review & agenda for future research," Journal of Business Research, 142, pp.914-929, 2022

[25] R. Tiwari and C. Herstatt, "Frugal innovation: A global networks' perspective," Die Unternehmung, 66(3), pp.245-274, 2012

[26] P. Soni and T. Krishnan, R., "Frugal innovation: aligning theory, practice, and public policy," Journal of Indian Business Research, 6(1), pp.29-47, 2014.

Growth responses of *Chlorella* sp. to some selected variants of culture medium and in effluents of a Brewery

Debadyuti Das[1], Sukhamay Maity[2], Debajyoti Kundu[3], Bana B. Jana[3], Suman Pandey[3], Jatindra N. Bhakta[3], Jaigopal Sharma[4]

[1]*Department of Rural Development and Management, Narendrapur Ramkrishna Mission, Kolkata, India*
[2]*Department of Zoology, Palpara College, Palpara, East Midnapore, West Bengal, India*
[3]*Department of Ecological Studies and International Centre for Ecological Engineering, University of Kalyani, Kalyani, West Bengal, India*
[4]*Department of Biotechnology, Delhi Technological University, New Delhi, India*

ARTICLE INFO

Keywords:
Chlorella sp.
Microalgae
pH
CO_2
Brewery
Effluent

Corresponding Author:
Jaigopal Sharma
Email: sharmajaigopal@gmail.com

ABSTRACT

Global warming poses a major threat to our civilization that has led to unsustainable development worldwide. Though there are different strategies for global warming mitigation, carbon sequestration in water column via microalgae is the eco-friendly, cost effective and sustainable tool to capture and utilize carbon in a beneficial way. Microalgae, *Chlorella* sp. was cultured in *ex situ* using (a) standard basal medium as control, (b) culture medium aerated, (c) exogenously introduced liquid carbon dioxide into the culture medium. The growth of *Chlorella* was also examined under different pH (4 - 11) of the culture medium as well as in various dilutions (0, 25, 50%) of effluents of a Brewery under autoclaved and non- autoclaved conditions. There was no marked difference in growth of *Chlorella* sp. between 0 - 166 hours and between 218 - 272 hours either, but significant difference ($P < 0.05$) in growth was clearly discernable during the peak period among the treatments. Exogenously introduced CO_2 or aeration of the culture medium did not significantly improve the growth over basal medium suggesting that optimal conditions of carbon or oxygen have prevailed in the basal medium of culture. There was a sharp dichotomy of pH effects on the growth of Chlorella as growth was distinctly higher at pH 8 to 11 compared to the remaining pH (4 - 6) showing no significant differences ($P > 0.05$) among themselves. *Chlorella* sp. grown in brewery effluent showed the maximum growth in 50% and 25% dilution of non autoclaved and autoclaved effluent of beverage factory respectively suggesting that autoclaving saved 25% of bacteria driven nutrient demand.

1. Introduction

Global warming poses major threat to our civilization that has led to unsustainable development worldwide. The primary reasons have been traced to the increasing trends of emission of greenhouse gases (GHGs) from industrial and anthropogenic sources into the atmosphere. These GHGs cause depletion of ozone layer thatprotects the atmosphere against UV radiation, thereby warming the atmosphere. As a result, the average concentration of CO_2 increased from 315 ppm in 1960 to 380 ppm in 2007 (IPCC, 2007).Further, there has been more than 35% increase in CO_2 emission worldwide since 1990.This concern has become the main stimulus for much research being carried out worldwide. Though there are different major strategies such as trapping of atmospheric $CO_2$from the atmosphere, long term oceanic storage and biofuel or biochar production, sequestration in water column via microalgae and macrophytes and carbon sink in bottom sediments of wetlands have been identified as a promising tool for mitigation of global warming. The acquisition of CO_2 through the biological process is eco-friendly, cost effective and sustainable to capture and utilize carbon in a profitable manner (Jana et al., 2016). It has several advantages for creation of mild conditions for CO_2 fixation and trapped $CO_2$is incorporated into carbohydrates and lipids and algal biomass can be transformed into biofuel and biochar (Peterhansel et al., 2010; Solovchenko and Goldber, 2013; Nogia et al., 2016) or natural algal food for fish and human consumption. This is a win-win strategy towards carbon mitigation and production of low carbon foot print fish, especially in tropical countries.

All photoautotrophs have to produce their own food through photosynthesis though their several modifications with xanthophylls and caretenoids occurs in higher plants (Eonseon et al., 2003).Microalgae are mostly photosynthetic and have chlorophyll embedded in the chloroplasts of microalgae(Hosikian et al, 2010).

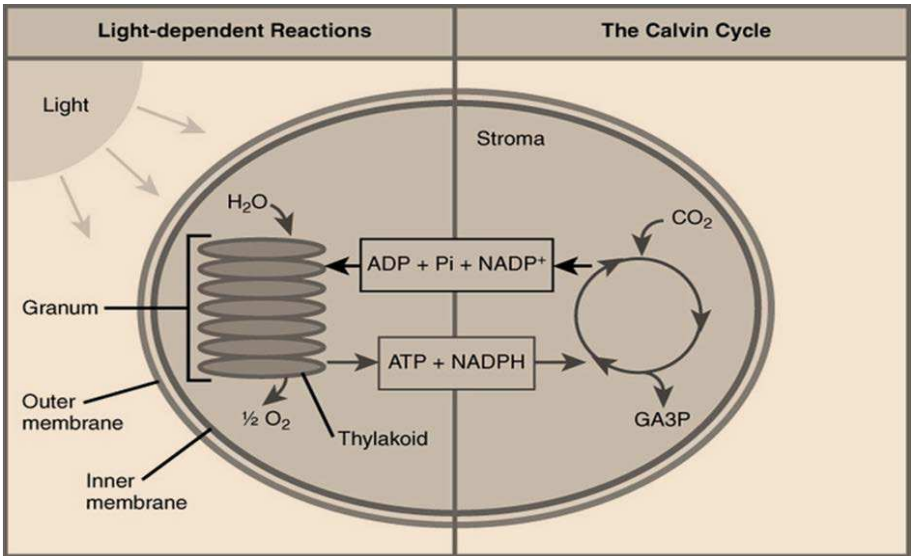
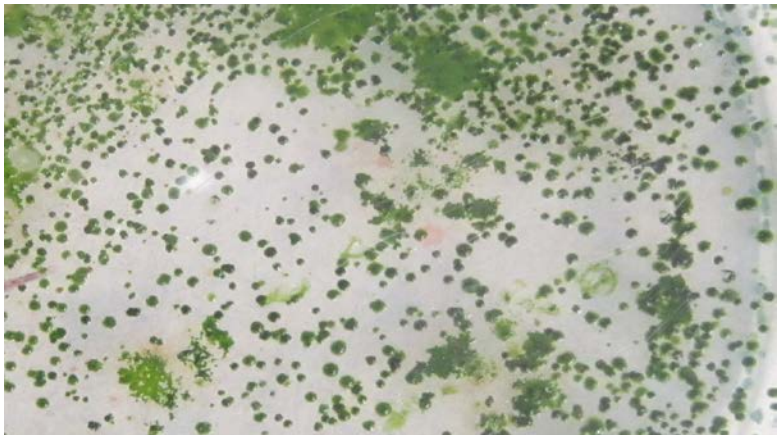


Fig. 1 A schematic representation of the photosynthetic process in microalgae

Photosynthesis is a process for the conversion of radiant energy into chemical energy with the help of $CO_2$and water. Altogethersix kinds of photosynthetic pigments (chlorophyll a, b, c, d, e and f) and many accessory pigments such as blue, red, brown, and gold are present in photosynthetic plant and algae. Microalgae are unique example where all kinds of photosynthetic pigments including chlorophyll e are present. Chlorophyll a is most essential to capture light wavelengths in the spectrum of sunlight. Chlorophyll b augments the ability of chlorophyll a to capture more sunlight. Chlorophyll c occurs mostly in dinoflagellates which also helps chlorophyll a to collect more sunlight, but it does not participate in photosynthesis beyond the initial stage. Chlorophyll d is found in marine red algae and cyanobacteria, and absorbs the infrared light of the electromagnetic spectrum. Chlorophyll e is a rare type found in some golden algae also assists in light capture (Pareek et al., 2017).Recently discovered chlorophyll f found in cyanobacteria and oxygenic bacteria absorbs further in the red (infrared light) than other chlorophylls. Chlorophyll e is a rare type found in some golden algae (Schliep et al., 2013).All types of chlorophyll viz. Chlorophyll a, b, c1, c2, c3, phytollated Chl c, d and f contains Mg. The molecular formula of chlorophyll d is $C_{54}H_{70}O_6N_4Mg$.

Photosynthesis is the conversion of photon energy into chemical energy to form glucose and water and oxygen is released (Fig. 1) as the byproduct. Chlorophylls are made up of lipid-soluble hydrocarbon tail ($C_{20}H_{39}$), a hydrophilic head and a magnesium ion at the centre. Microalgal photosynthesis occurs in two stages: the first stage is the photoactivation of chlorophyll when light energy is initially converted to chemical energy in the form of two compounds: NADPH, a source of energized electrons; and ATP, the versatile energy currency. The second stage is called Calvin cycle where sugar is produced (Fig. 1).The Calvin cycle starts with the capture of CO_2or carbon fixation from air into the organic molecule in the chloroplast (Raines, 2003).In this cycle when fixed carbon is reduced to carbohydrate (CH_2O) by the addition of electron provided by NADPH and ATP (Johnson, 2016) which are produced by light reactions of photosynthesis. In essence, both the steps are coordinated with each other because none of the steps require light directly.

Fig. 2 Culture of *Chlorella* sp. in agar medium



Chlorella is a freshwater single-celled green alga spherical in shape, about 2 to 10 µm in diameter, and is without flagella (Fig. 2). It contains the green photosynthetic pigments-chlorophyll a and b in its chloroplast. *Chlorella* is rich in protein and a good source of several B- complex vitamins and iron, minerals, antioxidants including vitamin C and carotenoids. *Chlorella* is available in a variety of dietary supplement and detoxification regimen in the form of liquid extracts, tablets and powders. It has immense medicinal properties to improve cholesterol and blood sugar levels in human being. Further, it is reported that chlorophyll may bind with specific cancer-causing agents to form molecular complexes and may be useful in blocking the carcinogenic effects of these agents (Solymosi and Mysliwa-Kurdziel, 2017).

With the help of carbon dioxide, water, sunlight, and a small amount of minerals, *Chlorella* multiplies rapidly in water. Because of this unique property, *Chlorella* has often been selected as carbon sequester (Cheng et al., 2006; Chiu et al., 2009; de Morais and Costa, 2007) at different temperatures and for cleaning the effluents of several industries. During outdoor cultivation with solar as the light source, biomass productivity is strongly affected by environmental factors such as irradiation, temperature (Ugwu et al., 2007), CO_2% and pH of the medium.

Despite considerable interest for *Chlorella* for as dietary supplement, medicinal properties, and mitigation options for global warming, no systematic studies have been focused on its mass production as influenced by different factors like pH, CO_2, nutrient level so as to harvest the beneficial option for different uses. The purpose of the present study is to examine the growth responses of *Chlorella* sp. in different pH of the medium under aerated and controlled conditions.

2. Materials and methods

Growth responses of *Chlorella* were examined in the laboratory using (a) standard basal medium as control, (b) culture medium aerated, (c) exogenously introduced liquid carbon dioxide into the culture medium (Fig. 3). *Chlorella* used in the study was procured from the stock culture maintained in the laboratory of ICEE, Kalyani University and grown in the basal algal culture medium(Hi Media) having the basal requirements (Table 1). All the methods for preparation of media, inoculation of *Chlorella* sp. were done following standard procedures of autoclaving of the media, serial dilution of the stock culture of *Chlorella* sp., in the range of 10^{5-6} andculture in 500 ml conical flasks in triplicate under artificial light conditions for 24 hours. The counts of *Chlorella* grown in different dilutions were enumerated in a microscope at different time intervals. The growth of *Chlorella* was also recorded in terms of density of culture medium determined in a Spectrophotometer at 680 nm against control.

Table 1. Composition of Bold Basal media.

Constituents	gm/ L
$NaNO_3$	25.0
$MgSO_4.7H_2O$	7.5
NaCl	2.5
K_2HPO_4	7.5
KH_2PO_4	17.5
$CaCl_2.2H_2O$	2.5
$ZnSO_4.7H_2O$	8.82 g
$MnCl_2.4H_2O$	1.44 g
MoO_3	0.71 g
$CuSO_4.5H_2O$	1.57 g
$Co(NO_3)_2.6H_2O$	0.49 g
H_3BO_3	11.42 g
EDTA	50.0 g
KOH	31.0 g
$FeSO_4.7H_2O$	4.98 g
H_2SO_4 (conc)	1.0 ml

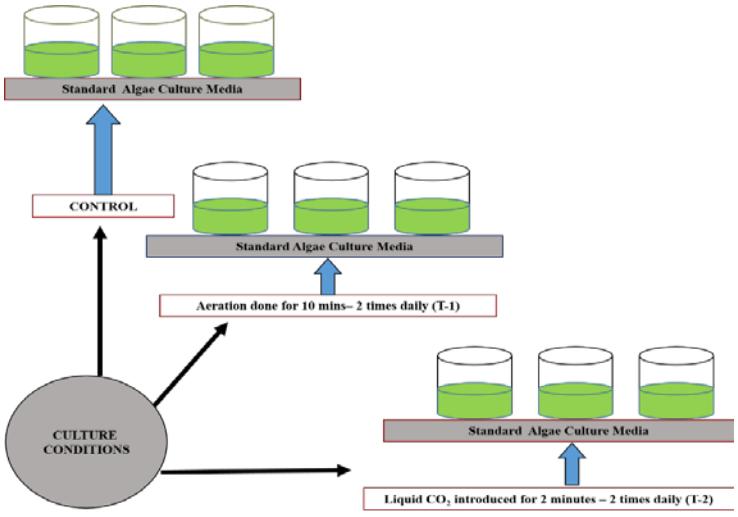


Fig. 3 Experimental design followed for growth of C*hlorella* sp. in algae culture media under aerated, exogenously introduced liquid carbon dioxide conditions

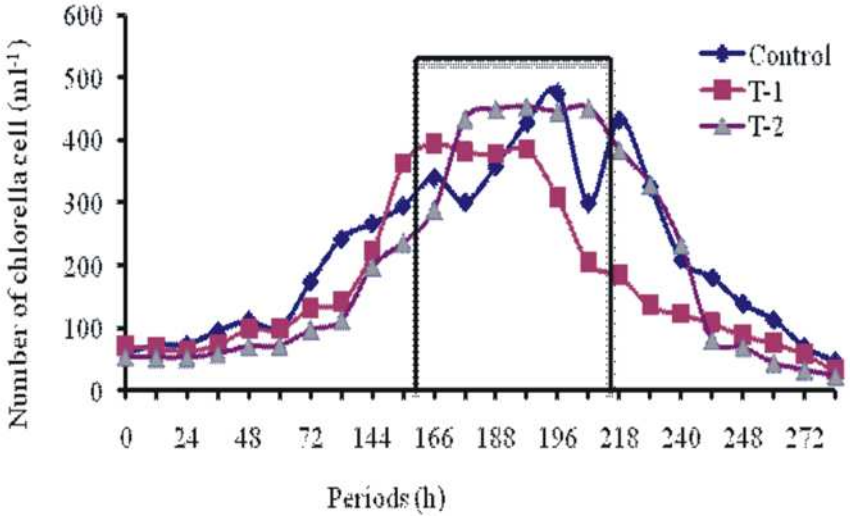


Fig. 4 Growth of *Chlorella* sp. under aerated, exogenously introduced liquid carbon dioxide conditions in algae culture media

2.1. Effect of pH of the culture medium

The effect of different pH of the culture medium was also determined; initial pH of the medium was measured with a digital pH meter. Different pH of the culture medium (4-11)was adjusted using 1 N HCl and N/44 NaOH. The culture was maintained in500 ml conical flasks. The growth of *Chlorella* sp. and the change of pH in the medium were recorded every 24 hours and adjustment of pH, if necessary, was done accordingly.

2.2. Growth of *Chlorella* in the effluents of beverage factory

In situ growth performance of *Chlorella* sp. was determined using different dilutions of the effluents of a Brewery factory. Effluent of beverage factory was collected and half portion of it was autoclaved under 15 lb pressures at 121°C for 20 minutes and then cooled to the ambient temperature where half part was not autoclaved as served as control. Both autoclaved and non autoclaved effluents werediluted by25% and 50% in conical flasks and volume was made up to 500 ml. Dilution was made by adding sterile water where one set of wastewater was kept as raw effluent. Culture of *Chlorella* was then inoculated in different dilutions of brewery effluents and allowed to grow under artificial light conditions in the laboratory. Growth of the algae species was measured at 680 nm in spectrophotometer (Shimadzu UV-visible Spectrophotometer, Model UV-1800) after every 24 hours from the culture till the growth cycle of *Chlorella* terminated.

3. Results and discussion

3.1. Growth of *Chlorella* sp.

There was no marked difference in growth of *Chlorella* sp. between 0 - 166 hours and between 218 - 272 hours either, but significant difference ($P<0.05$) in growth was clearly discernable during the peak period of growth ranging from 166 to 218 hours (Fig. 4). This was followed by a declining phase in all the three conditions including the control. The results of the study showed that exogenously introduced CO_2 or aeration of the culture medium did not significantly improve the growth over basal medium suggesting that optimal conditions of carbon or oxygen have prevailed in the basal medium of culture.

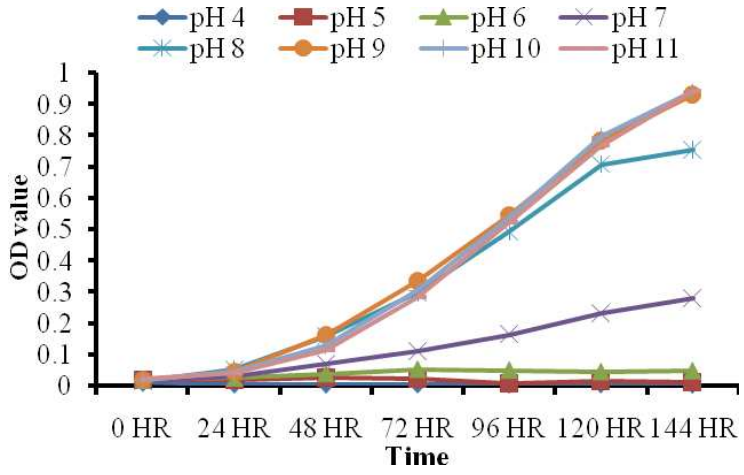


Fig. 5 Growth of *Chlorella* in different pH of culture media

Int J Env Tech Sci - www.journalijets.org

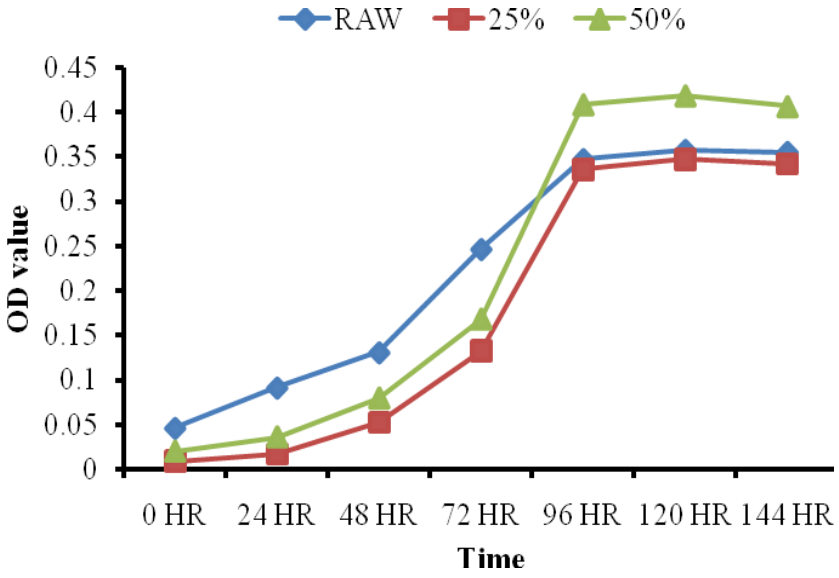


Fig. 6 Growth of *Chlorella* in non autoclaved effluent of Beverage factory

3.2. Growth of *Chlorella sp.* at different pH

There was a strong dichotomy of pH effects on the growth of *Chlorella*. There were clear-cut effects of pH of the culture medium on the growth of Chlorella sp. as growth was higher at pH 9 to 11 compared to the remaining pH showing no significant differences (P>0.05) among themselves. There was no difference during the lag phase of growth, but the differences tended to become more pronounced gradually after 48 hours when a strong dichotomy in growth was registered. The growth under pH of 8 - 11 remained distinctly higher than the rest of the pH of the medium tested. The growth at pH 7.0 remained higher than pH 4.0 - 6.0 (Fig. 5).

Optimal pH for growth of several algal species is reported to be in the range 7.5-8.5 (Acién Fernández et al., 2001; Chisti, 2008; Marcel et al., 2003; Molina et al., 2001), although there are species that can also grow at much higher pH (Ogbonna et al., 2000). pH values of beyond 10 can easily be reached in algal cultures in the absence of significant buffer or CO_2 supply and can lead to inhibition of growth (Qiang et al., 1996; Richmond and Cheng-Wu, 2001; Zhang and Richmond, 2003).Since algae can utilize $CO_2$for photosynthesis, rise in pH of water during intense photosynthesis is expected (Larsdotter et al., 2007). Besides, the precipitation reactions of cations with P at high pH, algal growth rate and species composition may also be affected by pH, especially in mixed populations in wastewater treatment systems. Though the growth of several algal species have been reported in the range of 7.5 - 8.5 (Marcel et al., 2003; Molina et al., 2001), *Chlorella* sp. in the present study and some other algae (*Rhodobacter sphaeroides*, *Chlorella sorokiniana* and *Spirulina platensis*) (Ogbonna et al. (2000) were able to grow well at higher pH. This shows that *Chlorella* utilized bicarbonate or carbonate as a source of carbon for their growth because of the fact that aquatic cyanobacteria and eukaryotic algae are able to use all forms of dissolved inorganic carbon (free CO2-CO_3-HCO_3) the dominance of which varied depending upon the pH of water. For example, bicarbonate (HCO_3) is most dominant (> 50%) at pH between 6.4 and 10.3, whereas carbonic acid (H_2CO_3) and carbonate (CO_3) are dominant at pH < 6.4 and > 8.3, respectively (Jana 1998; Wetzel, 2006).

It is likely that in *Chlorella* captured carbon is pumped into the cell by bicarbonate transporters present in both the plasma membrane and in the chloroplast envelope as reported for eukaryotic algae (Spalding et al., 2008). Inside the chloroplast, bicarbonate is concentrated, dehydrated spontaneously or by carbonic an hydrase through Calvin-cycle activity, finally yielding algal biomass.

3.3. Growth of *Chlorella* sp. in non autoclaved and autoclaved waste water

There were differences in growth of *Chlorella* between autoclaved and non autoclaved water of a brewery factory. Whereas 50% dilution showed higher growth of *Chlorella* than raw or 25% dilution in non-autoclaved effluent (Fig 6), it was 25% dilution of autoclaved effluent (Fig.7) had the same growth response of Chlorella. This suggests that microbial driven nutrient demandwas reduced by 25% in the autoclaved ones due to death of bacteria in the brewery effluent.

The algal biomass recovered from the wastewater can be used as feedstock for several products including biodiesel, biomethane, fertilizer, and nutraceuticals. Several algal strains have been reported produce neutral lipids (triacylglycerides) that can be converted into biodiesel (Hu et al., 2008; Olaizola, 2003; Schenk et al., 2008; Sheehan, 1998).

4. Conclusion

The conditions prevailed in the standard Bold algal culture medium was adequate for growth of *Chlorella* as no positive impacts of exogenously introduced CO_2 is or aeration were registered. Effect of pH was clearly visible as the growth of *Chlorella* was distinctly higher at pH 8 to 11 compared to pH 4.0- 6.0. Autoclaving of Brewery effluent saved around 25%of the nutrients for chlorella growth.

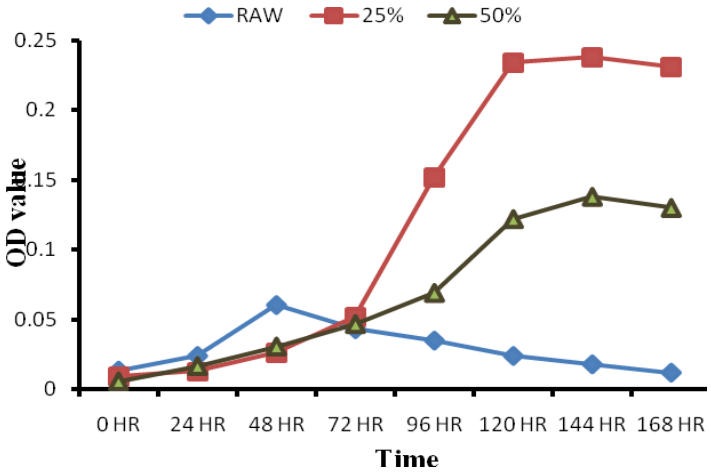


Fig. 7 Growth of *Chlorella* in autoclaved effluent of Beverage factory

References

Acién Fernández FG, Fernández Sevilla JM, Sánchez Pérez JA, Molina Grima E, Chisti Y (2001) Airlift-driven external-loop tubular photobioreactors for outdoor production of microalgae: assessment of design and performance. Chem Eng Sci 56(8):2721-2732.

Cheng L, Zhang L, Chen H, Gao C (2006) Carbon dioxide removal from air by microalgae cultured in a membrane-photobioreactor. Sep Purif Technol 50(3):324-329.

Chisti Y (2008) Biodiesel from microalgae beats bioethanol. Trends Biotechnol 26(3):126-131.

Chiu SY, Kao CY, Chen CH, Kuan TC, Ong SC, Lin CS (2008) Reduction of CO_2 by a high-density culture of *Chlorella* sp. in a semicontinuous photobioreactor. Bioresour Technol. 99(9):3389-3396.

de Morais MG, Costa JAV (2007) Carbon dioxide fixation by *Chlorella kessleri*, *C. vulgaris*, *Scenedesmus obliquus* and *Spirulina* sp. cultivated in flasks and vertical tubular photobioreactors. Biotechnol Lett 29(9):1349-1352.

Eonseon J, Polle JEW, Lee HK, Hyun SM, Chang M (2003) Xanthophylls in microalgae: from biosynthesis to biotechnological mass production and application. J Microbiol Biotechnol 13(2), 165–174

Hosikian A, Lim S, Halim R, Michael KD (2010) Chlorophyll extraction from microalgae: a review on the process engineering aspects. International Journal of Chemical Engineering 2010:1-11

Hu Q, Sommerfeld M, Jarvis E, Ghirardi M, Posewitz M, Seibert M, Darzins A (2008) Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. Plant J 54(4):621-639.

Jana BB 1998.The State-of-the-art of lakes in India: An overview. Archives fur Hydrobiologie (Suppl) (Monograph Studies) 121: 1 – 89.

Jana BB, Kundu D, Dutta D, Lahiri S, Bag S, Sarkar D, Bhakta JN (2016) Evidences of manure driven and C:N regulated enhanced carbon status and microalgal productivity in managed aquatic system under simulated green house conditions. Journal of Earth Science and Climate Change 7: 336.

Johnson MP (2016) Photosynthesis. Essays Biochem 60: 255–273.

Larsdotter K, Jansen JC, Dalhammar G (2007) Biologically mediated phosphorus precipitation in wastewater treatment with microalgae. Environ Technol 28(9):953-960.

Marcel J, Johannes T, Luuc RM, René HW (2003) Enclosed outdoor photobioreactors: light regime, photosynthetic efficiency, scale-up, and future prospects. Biotechnol Bioeng 81(2):193-210.

Marcel J, Johannes T, Luuc RM, René HW (2003) Enclosed outdoor photobioreactors: light regime, photosynthetic efficiency, scale-up, and future prospects. Biotechnol Bioeng 81(2):193-210.

Molina E, Fernández J, Acién FG, Chisti Y (2001) Tubular photobioreactor design for algal cultures. J Biotechnol 92(2):113-131.

Molina E, Fernández J, Acién FG, Chisti Y (2001) Tubular photobioreactor design for algal cultures. J Biotechnol 92(2):113-131.

Moroney J. V., Ynalvez, R. A., 2010. Proposed Carbon Dioxide Concentrating Mechanism in *Chlamydomonas reinhardtii*. Eukaryotic cell 6 (8), 1251-1259

Nogia P, Kaur S, Mehrotra R, Mehrotra S (2016)Capturing atmospheric carbon: biological and nonbiological methods. International Journal of Low-Carbon Technologies 11(2, 1):266–274.

Ogbonna JC, Yoshizawa H, Tanaka H (2000) Treatment of high strength organic wastewater by a mixed culture of photosynthetic microorganisms. J Appl Phycol 12(3-5):277-284.

Olaizola M (2003) Commercial development of microalgal biotechnology: from the test tube to the marketplace. Biomol Eng 20(4-6):459-466.

Pareek S, Sagar NA, Sharma S, Kumar V, Agarwal T, Gustavo A, Gonzalez-Aguilar, Yahia EM (2017) Chlorophylls: chemistry and biological functions. In: Yahia EM (ed) Fruit and vegetable phytochemicals: chemistry and human health, 2nd edn, Wiley-Blackwell, UK, pp 269-284.

Peterhansel C, Horst I, Niessen MBC, Kebeish R,Kürkcüoglu S, Kreuzaler F (2010) Photorespiration. The American Society of Plant Biologists 8:2-24.

Qiang H, Hugo G, Amos R (1996) A flat inclined modular photobioreactor for outdoor mass cultivation of photoautotrophs. Biotechnol Bioeng 51(1):51-60

Raines CA (2003) The Calvin cycle revisited. Photosynth Res 75: 1-10.

Richmond A, Cheng-Wu Z (2001) Optimization of a flat plate glass reactor for mass production of *Nannochloropsis* sp. outdoors. J Biotechnol 85(3):259-269.

Schenk P, Thomas-Hall S, Stephens E, Marx U, Mussgnug J, Posten C, Kruse O, Hankamer B (2008) Second Generation Biofuels: High-Efficiency Microalgae for Biodiesel Production. Bioenergy Res 1(1):20-43.

Schliep M, Cavigliasso G, Quinnell RG, Stranger R, Larkum AWD (2013) Formyl group modification of chlorophyll a: a major evolutionary mechanism in oxygenic photosynthesis. Plant Cell Environ 36:521–527

Sheehan J, Dunahay T, Benemann J, Roessler P (1998) A Look Back at the U.S. Department of Energy's Aquatic Species Program- Biodiesel from Algae. National Renewable Energy Laboratory (NREL), Golden CO. NREL/TP580-24190. Web. https://www.nrel.gov/docs/legosti/fy98/24190.pdf. Accessed January 17 2016.

Solovchenko A, Khozin-Goldberg I (2013) High-CO_2 tolerance in microalgae: possible mechanisms and implications for biotechnology and bioremediation. Biotechnol Lett 35(11):1745-52.

Solymosi K, Mysliwa-Kurdziel B (2017) Chlorophylls and their derivatives used in food industry and medicine. Mini-Rev Med Chem 17(13):1194-1222.

Ugwu CU, Aoyagi H, Uchiyama H (2007) Influence of irradiance, dissolved oxygen concentration, and temperature on the growth of *Chlorella sorokiniana*. Photosynthetica 45(2):309–311.

Ugwu CU, Aoyagi H, Uchiyama H (2007) Influence of irradiance, dissolved oxygen concentration, and temperature on the growth of *Chlorella sorokiniana*. Photosynthetica 45(2):309–311.

Wetzel R (2006) Transudation of elevated atmospheric CO_2 and natural light trough altered carbon fluxes and higher trophic level production in aquatic ecosystems.Verh. Internat.Verein. Limnol. 29:1112-1128.

Winston MM, Harold HS (1943) Chlorophyll d, a green pigment of red algae J Biol Chem 151:1-19.

Zhang CW, Richmond A (2003) Sustainable, high-yielding outdoor mass cultures of *Chaetoceros muelleri* var. *subsalsum* and *Isochrysis galbana* in vertical plate reactors. Mar Biotechnol 5(3):302-310.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-16598-x

HateDetector: Multilingual technique for the analysis and detection of online hate speech in social networks

Anjum[1] · **Rahul Katarya**[1]

Received: 12 December 2022 / Revised: 24 April 2023 / Accepted: 21 August 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
With the proliferation of social media platforms that provide anonymity, easy access, and the establishment of online communities and discussion, hate speech identification and monitoring has become a major concern for society, individuals and policymakers, which can be interpreted as hate speech. Many researchers attempted to detect hate speeches in multiple languages from social media, but the research was limited due to high complexity and minimum accuracy. A novel 'HateDetector: Multilingual Hate Speech Detection Technique' has been proposed to overcome these issues. In this technique, Bidirectional Encoder Representations from Transformers (BERT) with Multi-Layer Perceptron (MLP) is developed to identify the nature of the tweets by performing the process of code conversion and similarity check that results in good vector values representing the tweet nature. Additionally, the exact sentiment or nature of a tweet, whether hate or non-hate, is identified using the Profanity Check Technique (PCT), composed of ReLu activation function with a logistic regression classifier that classifies the resultant vectors and its respective emoji to neutral or hate speech. This technique performs all analyses of a tweet. It also auto-detects and easily finds hate speech, even from poorly written and complex text. According to the experiment's findings, the proposed technique performed exceptionally well, with a classification accuracy of 97.9%. Our proposed technique was able to compete with other state-of-the-art models.

Keywords BERT · Multilingual encoder-decode (MBart) · Online Hate speech (OHS) · Log-likelihood test · Logistic Regression · ReLu

✉ Anjum
anjum_2792@yahoo.com

Rahul Katarya
rahuldtu@gmail.com

[1] Big Data Analytics & Web Intelligence Laboratory, Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

Published online: 01 November 2023

1 Introduction

The vast majority of people across the globe are increasingly utilizing social media platforms such as Facebook, WhatsApp, Instagram, and Twitter. These platforms are user-friendly, easy to use, and give individuals the opportunity to have their voices heard. Now, people may readily share their thoughts and information from any location and at any time. Without meaningful restrictions or procedures, anybody can make detrimental and untrue comments in abusive or offensive language against anybody intending to spoil one's image and status in the community. Hate speech is the use of offensive and discriminatory language directed towards someone because of their origin, religion, race, sexual orientation, socioeconomic class, or any other attribute. When such things happen on social media to create content, blogs, or to exploit someone is called Online Hate Speech (OHS) [1]. The enormous number of posts, comments, and messages on these websites make it extremely difficult to monitor the information that is shared on them, despite the fact that they serve as a platform for individuals to express and exchange their ideas and viewpoints [2]. Moreover, many people tend to use aggressive, unwanted, and hateful language when discussing certain backgrounds, cultures, and other aspects [3]. This finally leads to inhumanity which may affect the social media users mentally, as also shown in Fig. 1.

People of all ages use social media, and they speak a variety of languages. These languages need to be translated into a common language, such as English, which is commonly referred to as "code." This is because English is a lingua franca language; more specifically, it is a declarative programming language that is vast, flexible, and standard. As a result, the term "multilingualism" refers to the ability of an individual to speak many languages and the presence of diverse language groups in the same geographic area [5]. Accordingly, in contrast to the general belief, most of the population uses multi-language or two languages at a minimum [6]. The minority population of the World uses a single language which is said to be monolingual. Multilingualism uses several languages as well as code-mixing [7]. Code-mixing is the mashing up of words, sentences, and phrases from two different grammatical systems within the same speech [8] (e.g., Tamil + English = Tanglish or Hindi + English = Hinglish).

The code-mixing of Hindi, which is the most widely spoken language in South Asia, with English is known as "Hinglish," a portmanteau derived from the two language's names. Hinglish differs significantly from its parent languages in syntax, phonetics, grammar, and even punctuation. The accent and sentiments are Hindi, while the vocabulary

Fig. 1 Hate Speech content on Twitter [4]

is made up of several English (Roman) transliterations of Hindi words, as well as a few English terminologies was the mixing of words, phrases, and sentences from two distinct grammatical (sub)systems within the same speech event [9]. With the wide-reaching popularity of social media platforms, code-mixing has emerged as one of the significant linguistic phenomena among multilingual communities that switched languages [10]. Thus, to detect hate speech, the existing presentations used a meta-learning approach based on metric-based and optimization-based (MAML and Proto-MAML) methods [9], Sentiment reversal analysis [11], Small-sized Transformer model [12], used several layers of classifiers, RGWE method for sentiment analysis [13], Deep convolution neural network [14], Lateral semantics analysis [15] and so on. The above-said methods quickly adapt and generalize new languages with labeled data points and obtain their objective. However, improvement is needed to analyze the data set given out as output because many techniques use low-resource languages and not high-resource languages, many classifiers to bring out its performance, and some concentrate only on some parts of hate speech. The major contribution in this paper is given as follows:

1. We have proposed a novel HateDetector that uses BERT-based word embedding with MLP in Multilingual Hate Speech Detection.
2. A 'Profanity Check Technique' is proposed to extract the emotion of the tweet for a better classification of hate speech.
3. The proposed model is trained and tested on three datasets against various models, which yields high performance under different experimentation settings.

The flow of this paper is as follows. Section 2 of this study briefly describes the literature review of various past studies related to detecting and analyzing OHS. Section 3 presents a detailed methodology of our proposed approach for the detection and analysis of OHS. Section 4 of this article describes the experimental results obtained after training the proposed model. Section 5 concludes our research work and presents the scope of future developments possible.

2 Literature survey

Classifying and detecting hate speech has recently gained considerable attention, and many researchers have proposed various approaches using ML and Deep Learning models. In this section, we have explored the various research papers for solving the problem of OHS. Some of the proposed models for OHS detection have been reviewed with its drawback.

2.1 Deep learning-based convolution neural network model

In standard machine learning, features are manually extracted, but in deep learning, this step is skipped in favour of directly incorporating data into a deep learning algorithm like a Convolutional Neural Network (CNN), which then makes additional predictions about the item. CNN is a subclass of DNN (deep neural networks). CNN is mainly used in the area of analyzing visual imagery. The three layers of an image are converted into a vector of suitable size, and then a DNN is trained on it. Their other applications include video understanding, speech recognition, and natural language processing. CNN was utilized by the author [16] to locate racist and sexist statements. The proposed model was tested by tenfold

cross-validation and gave a 78.3% f-score. The author [17] employed text features, i.e., surface-level features, linguistic features, and sentiment features in deep learning classifiers, and then implemented an ensemble-based novel approach. The author found an accuracy of 0.918 with the novel approach. The authors provided the CNN-based web browser plugin to visualize the online hostility on Twitter and Facebook [18]. By monitoring Tweets before they are shared among other Twitter users, the authors have constructed the Deep Convolutional Neural Network (DCNN), an automated approach to filter postings related to hate speech [19]. A database of 500 million tweets per day was difficult to process though it utilized GloVe embedding vector to filter the tweets with the help of a convolution operation. However, it cannot achieve better accuracy in the current dataset extension as it grows bigger daily. The authors [15] have presented a model utilizing Long–Short-Term Memory (LSTM), Ada Boost, and BERT. The pre-trained data set was given to BERT, and the classifier outcomes were given to AdaBoost. This model was used to identify the hate speech in the Bengal language along with Benglish content. Using L-Boost efficacy, it attained 95.11% accuracy. However, they did this only for one language with many classifiers.

2.2 Transformer methods

The transformer [20] is the latest innovation that has taken the natural language processing domain by storm. The transformer can account for long-term dependencies, but unlike LSTMs, transformers do not process data sequentially as done in the case of RNN's and LSTMs. Instead, to account for the position of each word is added to its embedding. The transformer was first introduced for machine translations (Sequence to Sequence Model), and thus it has two components, an encoder and a decoder. Though only the encoder is relevant in text classification tasks such as Hate speech detection. In an encoder, the inputs are first fed into a self-attention layer which generates an embedding taking into account all other words in a sentence and depicts the relevance of each word for a particular word. The embeddings obtained from the self-attention layer are fed into a neural network. This process is repeated many times, i.e., Many layers of self-attention and neural networks are stacked to form the encoder. The decoder of a transformer is very similar to the encoder except for an Encoder-Decoder attention layer, which is added to find the inputs relevant to a particular output [20].

In the context of hate speech detection, embedding obtained from the pre-trained model such as BERT (Bidirectional Encoder Representations from Transformers) has been widely used. BERT is a transformer trained using the masked LLM technique. Masked LM technique [21] requires 15% of the words in the sentence to be masked, and the transformer then attempts to predict these words from context during the training process. In the paper [22], the authors showed the efficacy of fine-tuning the BERT in the context of hate speech detection. Comparing pre-trained models for hate speech detection explores and compares various multilingual transformers such as mBART (multilingual encoder-decode). The research authors [23] contend that to obtain cutting-edge accuracy for the multi-class classification problem of online hate speech, transformers must be utilised instead of classic machine learning, RNN-based deep learning, or even attention-based RNN models. They propose a streamlined version of BERT, called DistilBERT, which has half the number of parameters with no loss in performance. On comparison and experimentation with various LSTM and BERT-based models, DistilBERT outperforms all the models given on various metrics. In our proposed work, we have introduced a novel technique that uses novel features of mBART, a sequence-to-sequence pre-trained model on large multilingual corpora.

In the paper [12], the author presented a simple transformer model on Twitter-based hate speech. They attained a 0.8426 F1 score and by providing more training data, they achieved a 0.8504 F1 score. However, it can be improved by the fore-coming techniques and other transformer models to extend the evaluation of Hate speech detection.

2.3 Sentiment analysis approach

The author [9] has presented a Meta-learning-based approach to analyze the issues of offensive content and few-shot hate speech in low-resource languages by only evaluating a few-labeled data items in a specific target language at high speed. This research has identified hate speech and offensive language as two different tasks and collected the data into two different datasets, some datasets across eight languages for hate speech and others across six languages for offensive languages. However, the number of languages was restricted, and high-resource languages were not considered. In the paper [24], the author has investigated Sentiment reversal which was used to find the relation between textual information and the sentiment diffusion pattern of Twitter messages. This investigation was done on the iterative algorithm Sentidiff, which yields 5.09% and 8.38% of PR-AUC on the classification of Twitter sentiment tasks. However, this model holds only the sentiment analysis where no other part of Twitter data handling was done. The authors [25] have investigated hate speech and offensive language detection by considering hate speech and similar content. Hate speech and offensive language were related to sentiment and emotion analysis, where the sentiment analysis was evaluated using negative opinions and emotions by pretending or experiencing the anger addressed in the speech. Finally concluded that hate speech and offensive language targeted a specific group or person. However, improvements can be made to identify the given speeches in the content. In the paper [26], the authors have introduced the RGWE method based on the concept of sentiment analysis which was used to find the optimal concept for the sentiment of words under different contexts with sentiment representation of words. However, only the noun concept was done by Microsoft concept graph, whereas verbs, adverbs and so on were not found in the paper. The author [27] have presented a model using Latent Semantic Analysis (LSA) to detect hate speech in social networks, which uses ten binary data sets with different categories. It used 2–4 ngrams to find character and 1 to 5 ngrams to find words and an advanced machine language model, CAT boost, to attain its objective of detecting hate speech, but it did not work for complex and overlapping hate speech. In Table 1, the systematic literature review of the hate speech paper has presented in which the first two column shows the reference and objective of that reference. Next, the classifiers used in the experiments, the dataset reference, and the results and limitations were discussed.

From Table 1, We can determine that the authors have included numerous feature extraction approaches, and by using those features, they trained different ML classification algorithms. Authors have devised various novel models for hate speech detection and have worked on fine-tuning its parameters.

2.4 Limitations of the existing research

1. The research work has been limited to spotting hate in the English Language and few pieces of research in Arabic, Indonesian, Italian, Turkish, Swedish, Albanian Language, and hate content in the rest of the languages like Hindi and Hinglish goes unfiltered.

Table 1 Systematic Literature Review of latest hate speech detection techniques

Ref	Objective	Classifier Used	Features Used	Dataset	Results and Limitations
[28]	Deep multitask framework has proposed to combat the online hate speech	CNN, GRU, LSTM	Unigrams, Bigrams, Trigrams, TF-IDF	Hatebase.org	• A deep multi-task learning framework has been suggested to exploit the information from these various connected tasks • The framework can be extended to include user metadata, user posting patterns that promote harmful speech, and the likelihood that any hazardous language will spread
[19]	Proposed a Deep convolution neural network	DT, LR, NB, SVM, GB, KNN	Self-extraction of contextual features	T. Davidson from GitHub	• The proposed DCNN model makes use of the tweet text and the GloVe embedding vector to convolutionally extract the semantics of the tweets •To detect hate speech patterns, a more comprehensive dictionary could be developed
[29]	Proposed a neural architecture for classifying multilingual hate speech on Twitter	Fasttext Ccrawl, LSTM, GRU	Word embedding, Ngrams, Social network–specific features, Emoji embedding	Data was collected from Zeerak Waseem and Dirk Hovy Dataset from HateSpeechData website by Evalita	• The author has experimented with English, Italian, and German language • They have first discovered a recurrent neural architecture that is fairly reliable and effective across several languages

Table 1 (continued)

Ref	Objective	Classifier Used	Features Used	Dataset	Results and Limitations
[30]	Proposed technique for Classifying texts on whether they are racist or not	SVM	BOW, Bigrams, POS tags	Hate speech data was gattered from Yahoo Directories of (3 million words)	• The text classification Using using SVM, problems like overfitting can be prevented • Proposed techniques save time and labour • Training and testing was performed on comparatively small dataset
[31]	Detection of intent in a Twitter Post	Random Forest, Naive Bayes, Decision Tree	Lexicon Features, Cohen's Kappa coefficient, NLP	Data collected from Tumblr (3228 text posts)	• They discovered numerous factors that could result in unjustified bias • As a result of their focus on a small number of annotators, the number of posts examined in this paper is rather small
[32]	Study of Twitter data for racial bias	Logistic regression	BOW, TF-IDF	Blodgett et al. 2016 (59.2 million tweets)	• Classifiers trained with each of the five datasets exhibited racial bias • Distinct algorithms and global features may produce variable outcomes • Only one facet of racial bias was studied in this study
[33]	Proposed a Novel algorithm to detect hate speech messages of Twitter	Random Forest	Kappa coefficient	Data collected using Twitter APIs (20,880 tweets)	• Their performance prediction contributed to the regulation of consistent nodes, which resulted in a smaller tree that was not overfitted • The dataset is confined to attacks in London

Table 1 (continued)

Ref	Objective	Classifier Used	Features Used	Dataset	Results and Limitations
[34]	Identifying hate speech in posts using RNN	RNN	Character Embeddings, GRU	HatEval (19,800 tweets)	• The approach presented in this paper was able to manage posts with large noises • Several variants were assessed to establish the ideal model
[35]	Detection of hate speech using CNN deep learning model	CNN	Character 4-g, Word2vec, N-grams, BOW	Waseem (6656 tweets)	• Various types of hate speech have been classified • Various features such as Bert and combination techniques such as LSTMs or transformers may be incorporated in the future

2. In order to furnish research in the field of OHS, a multimodal and multilingual dataset should be developed.
3. One of the constraints is the lack of publicly available data towards the progress of online hate speech detection.
4. Very few give high-accurate performance profanity check techniques have been introduced.
5. Lastly, to reduce the complexity of words caused by incoming datasets while performing detection is not explored.

From the existing reviewed proposals, it was viewed that they used many classifiers, concentrated only on a single analysis, hate speech detection was done only for one language, and required improved performance in those models. Thus, a novel technique must be implemented to overcome the abovementioned issues. Our research has tried to devise a novel technique using Bidirectional Encoder Representations from Transformers with multi-layer Perceptron to identify the nature of the tweets by performing the process of code conversion and similarity check that results in good vector values representing the tweet nature. Additionally, we exacted the sentiment or nature of the tweet, whether it is positive or hateful, is identified using PCT that is composed of ReLu activation function with a logistic regression classifier that classifies the resultant vectors and its respective emoji's. In next Sect. 3, we discussed the proposed methodology of research work.

3 Proposed methodology for HateDetector

The "HateDetector: Multilingual Hate Speech Detection" model consists of several steps that help in detecting and analyzing hate speech. Firstly, the model uses a Bidirectional Encoder Representation of Transformer (BERT) to pre-process the tweets from Twitter media. BERT is a powerful natural language processing model that can understand the context of the text and produce high-quality embeddings for text. The word embeddings produced by BERT are then given to the mBART model, which converts the code-mixed tweet to English. This conversion helps in ensuring that the model can detect hate speech in multiple languages and code-mixed languages. After the tweet has been converted to English, it is given to Bag of Words (BOW) with N-gram. BOW is a technique that splits the given sentence into words and weighs each word using numbers by mapping the sentence provided using vector equations. This technique helps in detecting and identifying the frequency of the occurrence of specific words that may be associated with hate speech. The output of BOW is then fed to the similarity checker to check the spelling and grammar of the tweet. This step ensures that the model can detect and analyze hate speech that may contain incorrect spelling or grammar. Finally, the model employs the Profanity Check Technique (PCT) with Rectified Linear Unit (ReLu) activation function to recognize the emotion of the sentence. PCT is a technique that uses a set of profane words to identify the emotion associated with a sentence. The output of the PCT layer is then given to a sigmoid function that works between the zeros and ones. The sigmoid function provides binary outcomes that indicate whether the sentence is positive or negative. Then finally the logistic regression layer transmits the final output by analyzing the nature of the sentence, whether it is positive or negative, and identifying any hate speech in the sentence. The proposed model provides an effective solution to the challenges faced by Twitter in managing and detecting hate speech in the open space. The model employs various techniques such

as BERT, Multilayer Perceptron-based word embedding, and Profanity Check Technique (PCT) to detect and analyze hate speech in multiple languages, code-mixed language, and identify the emotions associated with the sentence. The model's ability to detect and analyze hate speech in various languages and code-mixed language makes it a valuable tool in promoting a safe and healthy social media environment.

Figure 2 shows the process flow of the proposed model. Initially, a code-mixed tweet is given to the BERT with Multilayer Perceptron (MLP), which consists of several encoders

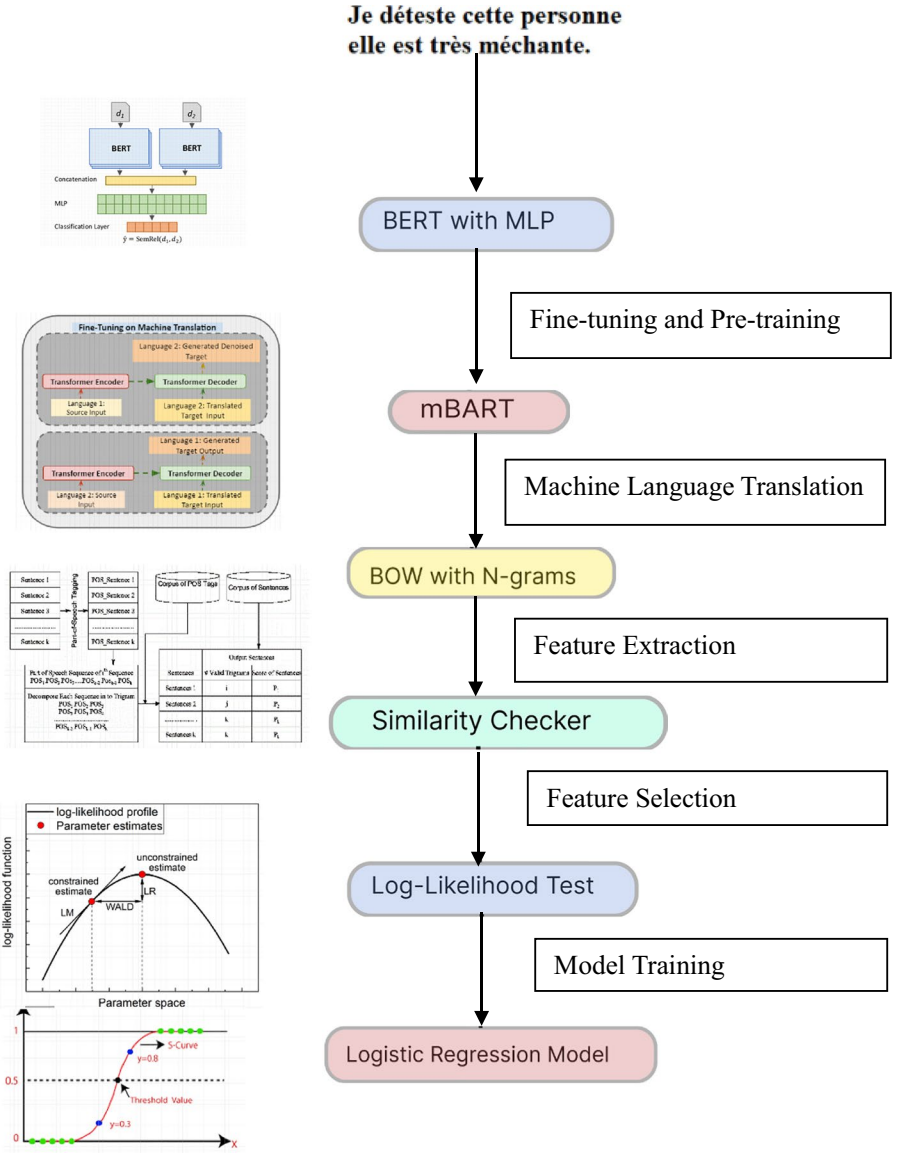


Fig. 2 Flow chart of the proposed technique

and decoders to pre-process the mixed tweets and remove all the unnecessary stopping words and tags. All the hidden layers are used to compress both text and image datasets. The pre-processed data is entered into the mBART (multilingual encoder-decode), which de-noise the key-phrased dataset again. The BOW with n-gram maps the vectors in the dimensional feature space and splits according to the gram. A similarity checker is used to gamble the vocabulary that collects the data, analyses the meaning confirms it with the trained data available and investigates the speech. The logarithmic value of the log-likelihood test performs the process of sensing the Positive sentiments of the words. However, it could not reach the correct sense of the sentence, so PCT with activation function and classifier is performed with the sigmoidal flow of zeros and once, which precisely identifies the nature of the sentence and detects the hate speech. In the next section, we discussed our proposed methodology in a detailed fashion.

3.1 Multilingual hate speech detection technique

A multilingual hate speech detection that holds up a BERT with multi-layer perceptron-based word embedding. It detects hate speech of tweets with non-linguistic concepts. This is done with the help of several tests and functions performed within the multilingual hate speech detection technique, which we will discuss in the below sections. The classification performance is divided into a trained, test, and validation set. This training set is used as a reference for the pre-processed dataset, the test set is to optimize the proposal's approach, and the validation set is to evaluate the proposed technique.

3.1.1 Bidirectional encoder representation of transformer with MLP

The proposed model consists of BERT [21] with a 12-layer transformer network that pre-trains the data set and fine-tunes the tweet. The pre-trained tokens are provided to each layer, a token from each layer is used as a token for the next layer, and this word embedding uses a neural network. Multi-layer Perceptron embedded within the BERT is a feed-forward artificial neural network with one input layer and 768 hidden layers and one output layer that processes the input data by backpropagation.

In Fig. 3, the data set assigned for the task is given into the input layer of the BERT, where the input data is fine-tuned and pre-trained in the hidden layers composed within

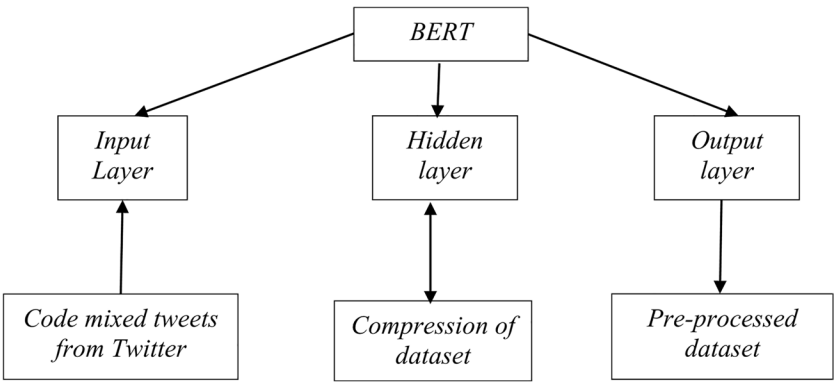


Fig. 3 Pre-processing of tweets in Twitter datasets by BERT

the BERT, and the key phrases are given out through the output layer. Here, the input layer sends the data set, the upcoming hidden layer compresses the tweets fed into it, and the pre-processed key phrases are given out in the output layer. The BERT holds many pre-trained tasks that helps to perform well. Algorithm 1 shows the pre-processing flow where the tweet is cleansed by converting the lower to upper case and removing the stopping notations, URLs, tags. The tweet is split into each character, and the BERT model starts reading each character compared with the pre-trained data set.

Input: Tweet text

Output: Cleaned tweet tex

1. Convert all characters to lowercase

2. Remove URLs and mentions (words that start with '@')

3. Remove hashtags but keep the words after them

4. Replace contractions (e.g., "can't" -> "cannot")

5. Replace slang and informal words with their standard equivalents (e.g., "u" -> "you")

6. Remove non-alphabetic characters and punctuation, except for exclamation and question marks

7. Tokenize the text into individual words

8. Remove stop words (common words that don't carry much meaning, such as "the" and "and")

9. Perform stemming or lemmatization to reduce words to their base form

10. Combine the remaining words back into a single string and return the cleaned text.

Algorithm 1 Pre-processing of Tweets for the Hate speech detection

In the pre-processing section, the given data set is processed before it is given as input to the mBART. The dataset consists of tweets which further consist of Emojis to express emotions, tags, stopping notations, uppercase letters, and URLs are cleansed initially from each tweet by the above algorithm 1. Whether the text contains hate or not, these embellishments add nothing to the text. The next step of pre-processing the tweet is tokenization and lemmatization, where another specific symbol replaces the sensitive data without giving up the security level and the tokens are lemmatized and finally, the output is stemmed and key-phrased where plurals are removed and forms as a key-phrase sentence that is stored and used when it is needed.

In Fig. 4, the complete pre-processing technique of the Twitter data set is shown with an illustration. In a sense, how a tweet is converted into Tokens and key phrases into a sentence. Here, a tweet from the Twitter page is used. It is first tokenized into a single word, then lemmatized, which means the keyword is arranged in the pre-processed data set, shortcuts of words are enlarged, then key-phrased into a sentence. A brief illustration of the implementation of BERT with MLP is given below.

1. Preprocess the input data: We have tokenized the input into individual words or subwords, converted it into numerical sequences, and padded it to a fixed length. It involves breaking the input text into individual words or subwords. We have used a tokenizer

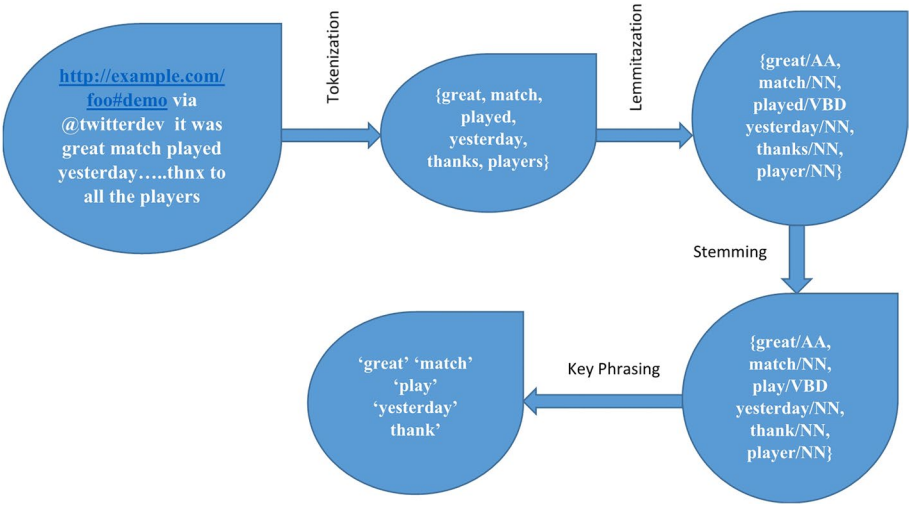


Fig. 4 Systematic representation of Preprocessing of the Twitter Dataset

library such as Tokenizer from the Transformers library in Python. We have created an instance of the tokenizer class and used its encode method to tokenize the input text into a list of tokens.

2. Load the pre-trained BERT encoder: Then we used TensorFlow to load the pre-trained BERT model.
3. Extract features using BERT encoder: In this step, we have passed the preprocessed input sequences through the BERT encoder to obtain the output features for each token. BERT outputs a sequence of vectors of dimension d, where d is the hidden size of the BERT model.
4. Feed features into an MLP classifier: We flatten the sequence of output features to obtain a fixed-size feature vector for the entire input sequence and feed this vector into an MLP classifier. The MLP consists of one or more hidden layers, each with a nonlinear activation function such as ReLU, and a final output layer with the appropriate number of units for the task-specific prediction.
5. Train and evaluate the BERT-MLP model: We have trained the model on the labeled training data using an optimization algorithm such as Adam and evaluated its performance on a held-out validation set. Then model is fine-tuned by adjusting the hyperparameters such as learning rate, number of hidden layers, and batch size.

3.1.2 mBART: Multilingual bidirectional encoder representation of transformer

The pre-processed and pre-trained output from BERT is given as an input to mBART [36] that denoises the multilingual text by encoding it sequence by sequence, Hinglish is converted into English since English is the code for the NLP, which fine-tune any of the supervised or unsupervised pair of languages without any specific modification in task with one trained set of parameters for all languages even if it is complex.

The process flow of mBART is shown in Fig. 5, an autoencoder converting the output from BERT to the reconstructed code output by compressing the key phrases into their

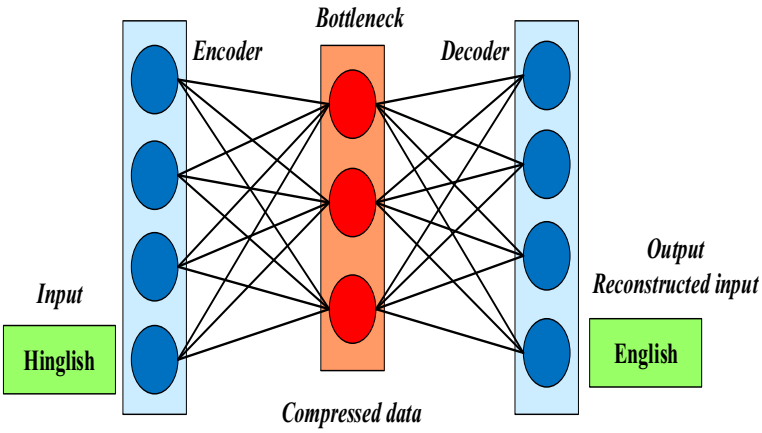


Fig. 5 Process flow of mBART [37] for proposed technique

original language and then converting it into the code language understandable by the machine.

Example String: "Je déteste cette personne, elle est très méchante." This is a French sentence that translates to "I hate this person, she is very mean." in English. Here we have used mBART model to detect whether this sentence contains hate speech or not. So, first, we have performed the preprocessing on the above text to convert it into a format that can be fed to the mBART model. This involves tokenizing the text, converting it into numerical sequences, and padding the sequences to a fixed length. In this research, we have tokenized the text into individual words, converted each word into its corresponding numerical index in the mBART vocabulary, and pad the sequence to a fixed length of 50 tokens.

Tokenized text: [’Je’, ’déteste’, ’cette’, ’personne’, ’,’, ’elle’, ’est’, ’très’, ’méchante’, ’.’].

Numerical sequence: Next, we converted the tokenized sequence into a numerical sequence by mapping each subword to its corresponding numerical index in the mBART vocabulary. This gives us the following numerical sequence: [27, 30,888, 618, 4480, 18, 299, 51, 529, 31416].

Since mBART processes sequences of fixed length, we need to pad the numerical sequence to a fixed length. In our experiment, we used the padding of 10. Then we can pad the sequence with the special padding token, which has a numerical index of 1, to get the following padded sequence: [27, 30,888, 618, 4480, 18, 299, 51, 529, 31,416, 1]. Now, we fed the padded numerical sequence into the mBART encoder network to get the fixed-length representation of the input sequence. The encoder network applies a series of self-attention and feedforward layers to the input sequence to compute a sequence of hidden states. We took the final hidden state, which summarizes the entire

input sequence, as the fixed-length representation. We denoted the fixed-length representation as p.

p = Encoder([27, 30,888, 618, 4480, 18, 299, 51, 529, 31,416, 1]).

Finally, we fed the fixed-length representation z into the mBART decoder network to generate the output sequence in the target language. The decoder network applies a series of self-attention, encoder-decoder attention, and feedforward layers to the input representation to compute a sequence of hidden states, which are then used to generate the output sequence token by token. Since the input sentence is in French, we have set the target language to be English, which is the code language for the mBART model. We denoted the output sequence as y.

y = Decoder(z, [2])

Here, [2] is the numerical index for the special start-of-sequence token in English. The decoder network generates the output sequence by predicting the next token in the sequence given the previous tokens and the input representation z. The output sequence is terminated by the special end-of-sequence token, which has a numerical index of 0 in English.

3.1.3 BOW with N-gram

In this step output from mBART is given into BOW with N-gram that splits words from the sentence. Features are extracted by BOW that tokenizes the sentence into words and then these tokens are taken as an array of the frequency called it as bow vector; finally, these vectors are mapped at a fixed dimension in the space given. In simple words, a positive, negative or neutral sentence is converted to numbers by the following methodology of mapping the words.

Let us take all the sets of finite sequences of words as A and pre-trained dictionary labels as D. Mapping of finite set A to the dimensional feature space is done by

$$\varnothing : A \rightarrow R^M \tag{1}$$

The sentiment label set is taken as $\gamma = \{1, \ldots .., K\}$, assuming $K = 2$ shows whether it is positive or negative. The labeled trained data set trains γ and the input dataset is given as $x = (w1, \ldots .wN)$ where N is the input sequence length. In bags of n-gram $\varnothing(x)$ maps x to $M = |\Gamma|$ where M is dimensional latent space where Γ is the vocabulary of n-grams and Γ is evaluated using $|\Gamma| = O(|D|^n)$.

The embedding of the γ_j is evaluated by,

$$P_{\gamma_j} = h(F \times z_{\gamma_j} \tag{2}$$

where F: projection matrix

$$h(.) = \tanh(.)^3$$

$$\gamma_j = (w_j, w_{j+1}, \ldots .., w_{j+n-1})$$

F: projection matrix

z_{γ_j} is an operator string of trained set

Finally, the vector representation of a document is represented as

$$\varnothing(x) \equiv d_x = 1/N \sum_{j=1}^{N} P_{\gamma_j} \tag{3}$$

where, $d_x \in \mathbb{R}^M$ and $x = (w1, \ldots, wN)$

The reason for formulating this evaluation is to fix the length n of the sentence with the number of phrases.

A simple N-gram calculation is done by

$$Ngram_K = X - (N - 1) \tag{4}$$

where,

$Ngram$: Weightage

K: Given sentence

X: Number of words

Equations 3 and 4 help in the easy and correct calculation of n-grams. The output of the n-gram is an n-gram dictionary with an actual valid n-gram id, n-gram phrases, document frequency, length of n-gram, global term frequency, and other information [18]. We use these details to find the output in the next step of our process.

3.1.4 Similarity checker and log-likelihood test

The similarity checker detects, removes, and corrects the spelling mistakes in the respective input from the n-gram dictionary. We use a computer-programmed similarity checker, Grammarly and plag checker, that compares the present vocabulary with the already stored one. Furthermore, to evaluate whether the output of the checker is a good one or not, Log-Likelihood Test is performed. If the log-likelihood value is higher than the dataset better suits the model. It ranges from a Negative infinity value to a Positive infinity value. Values are obtained by Eq. 5,

$$l(\theta) = lnL(\theta) \tag{5}$$

where

$L(\theta)$:the output of the similarity checker

$l(\theta)$: the output of the log-likelihood test

From Eq. (5), the validated vocabularies are given out as output, but it is not precise in its emotion analysis as the classification of words may fail to detect the nature of the sentence correctly. Thus "Profanity Check Technique" is implemented which detects the correct emotion of the sentence, whether it is sad or happy, or angry.

3.2 Profanity check technique

We proposed a method called Profanity Check Technique (PCT) that uses a neural network to identify negative or profane language in a given string of text. The

neural network is built using Rectified Linear Unit (ReLU), a popular activation function used in deep learning models. The PCT method utilizes a library of words and phrases that are known to be associated with negativity or profanity. The input string is compared with this library to determine whether it contains any of these words or phrases. If the input string is found to have negative or profane language, it is flagged as such. The neural network used in PCT consists of multiple layers and is trained using sigmoidal function and logistic regression. The sigmoidal function is used to activate the vectors provided to the neural network. However, this function can sometimes suffer from vanishing gradient issues, which can lead to slow convergence during training. To mitigate this problem, We have used ReLU in the calculation, which does not use any transforms and thus avoids vanishing gradient issues.

ReLu function $f(x)$ is calculated by

$$f(x) = max\ (0, x) \tag{6}$$

From Eq. (6) the output of the function is found. As if it returns 0 it is negative, or else it is positive. Though, complex functions cannot be learned using this ReLu function which may lead to regression problems. A generalized linear classifier model known as logistic regression is implemented to overcome this issue. That undergoes supervised learning with a labeled dataset and analyzes pre-trained binary data where the output is continuous with the parameters and summing of input and brings out the relation between independent and dependent variables. Logistic regression uses the sigmoid activation function to take back the label's probability to code any value between 0 and 1. The S curve shows that the output is negative, where it cannot go behind 1. This is calculated using the Eq. 7 below.

We have used the Loglikelihood test along with the Profanity Check Technique to improve the performance of the hate speech detection model. While the PCT is useful in identifying profane words or phrases, it may not be enough to accurately identify hate speech since not all hate speech contains profanity. The log-likelihood ratio test is a statistical method that can help determine the likelihood of a given text being hate speech based on the frequency of certain words or phrases in hate speech versus non-hate speech. By combining the PCT with the log-likelihood ratio test, the model can potentially achieve a higher accuracy in detecting hate speech. We have proposed the novel PCT algorithm 2 for the OHS problem.

$$Loglikelihood\ test = 2\log_e\left\{L_s(\theta)/L_g(\theta)\right\} \tag{7}$$

where,

s :pre-processed sample

g :pre-trained parameters

From equation [7] the output of the log-likelihood test is evaluated by giving the sample attained with the pre-trained dataset available.

Input: *twt*: A text string

Output: *decision:* A binary label indicating whether the text contains profanity or not

Begin

1. Input: A string of text

2. Perform pre-processing on the input text:

a. Convert the text to lowercase

b. Remove all punctuation marks

c. Tokenize the text into individual words

d. Remove stop words

e. Perform stemming and lemmatization

3. **Use a pre-defined list of profanity words and compare each word in the pre-processed text to the list of profanity words.**

4. If any profanity words are found, assign a score of 1 for each profanity word.

5. Calculate the loglikelihood score for the input text using a pre-trained language model:

a. Tokenize the pre-processed text into a sequence of tokens

b. Use the language model to predict the probability distribution of the next token given the previous tokens.

c. Calculate the log-likelihood score of the input text based on the predicted probability distribution

6. Combine the PCT score and the loglikelihood score to classify the input text:

a. If the PCT score is greater than a pre-defined threshold AND the loglikelihood score is negative, classify the input text as hate speech

b. If the PCT score is greater than a pre-defined threshold but the loglikelihood score is positive, classify the input text as profanity

c. If the PCT score is below the pre-defined threshold, classify the input text as non-offensive

7. Output the classification of the input text.

Example: for the above Algorithm

Input: "I hate this person, she is very mean"

Pre-processed text: "hate person mean"

Profanity words: ["hate", "mean"]

PCT score: 2

Loglikelihood score: -7.5

Threshold: 1.5

Since the PCT score is greater than the threshold and the loglikelihood score is negative, the input text is classified as hate speech.

Algorithm 2 PCT for OHS

The final output from the above algorithm is then passes to the logistic regression algorithm. It is a machine learning algorithm that is commonly used for binary classification tasks. It takes a set of input features and learned a set of weights for those features that can predict the probability of a given example belonging to a particular class (in this case, the class of hate speech vs. non-hate speech). Once the input text has been preprocessed and features have been extracted using Bag-of-Words

with Ngram, logistic regression is trained on these features to learn the weights that best predict the probability of a given example being classified as hate speech. During training, the logistic regression algorithm iteratively adjusted the weights of the features until it achieves the highest possible accuracy on the training data. Once the model has been trained, it is used to predict the probability of new examples belonging to each class. In the case of hate speech detection, the logistic regression model would output a probability score for each input text indicating the likelihood that it is hate speech. The output of the PCT with Loglikelihood test is used as input features for the logistic regression algorithm, which can then learn to predict the probability of a given example being classified as hate speech.

So, it is clear that the proposed technique utilizes high resource language for hate speech detection and a much more pre-trained data set is placed in the dataspace with the help of a multi-layer perceptron embedded with BERT. All the forms of text, images, and emojis are taken into account to detect the OHS, and finally the emotion is efficiently predicted with the help of the above profanity check technique.

4 Experimental result

BERT with a multi-layer perceptron-based word embedded system has been recently used by most researchers who undergo research on hate-speech detection. The technique used in this study was a Multilingual hate speech detection technique with word embedding and detected hate speech on the Internet. The proposed technique is used to detect the pre-processed dataset's vocabulary. Good words have been evaluated, and hate speeches have been detected by sensing the exact emotion of the sentence formulation and reporting an individual or a group.

4.1 Implementation setup

Python programming language (version 3.6) [38] and its machine learning (ML) module Scikit-Learn [39] was utilized for all experiments and model construction. All calculations were performed on a system with an Intel(R) Core i7 8750H CPU 2.20 GHz processor, 8 GB RAM, and a 4 GB NVIDIA GeForce 1050 TI Graphics Processing Unit running Windows 10 Home edition. Google Collaboratory Notebook [40] was utilized for certain intensive computations and model training. Further in this section, we have discussed the four datasets collected from the open-sourced repositories, the experimentation results we obtained after training the proposed technique on these Datasets, and a brief discussion and comparison report on the results obtained. This section is further divided into three subsections to incorporate all this information.

4.2 Data set description

The simulation result has been discussed in this section in detail. The hate speech detection uses the Multilingual Hate Speech Technique, which uses BERT with a multi-layer neural network that is further encoded with PCT to hold the emotion of the data set. The methodology carries tweets on Twitter media as the dataset for performing the methodology.

1 T. Davidson [41]: The dataset consists of 24,783 tweets divided into three categories: Hate speech (5.8%), Offensive Language (77.4%) and Neither (16.8%). This open-sourced dataset was collected from GitHub. The tweets labeled as Offensive or Neither were included in the Non-Hate Category. Furthermore, the data set column labeled as having Hate Speech was included in Hate Category.
2 Hinglish Hate Speech Dataset: The first set of data was drawn from the paper [42]. Two linguists, one fluent in Hindi and the other fluent in English, annotated the tweets that were gathered through the Twitter API. Data from the research classification of Offensive Tweets in the Hinglish Language [43] was used for the second batch of data. As a result of our current task, the dataset has been relabeled into two classes: Not-Abusive and Hate-inducing categories have been labelled as hate, while the non-offensive category has been relabeled as not hate. The final piece of data comes from the HASOC task, which was a collaborative effort.
3 Hot (Hate Offensive Text) [43]: It was collected by using the Twitter Streaming API[3], in which the tweets had more than three Hinglish words. These tweets were then manually annotated. This collection of tweets covered the period spanning November 2017 and February 2018. In order to limit the scope of our data mining to a particular geographic region, we only considered tweets that originated from the Indian subcontinent. It has been shown that in a Hindi Offensive Text (HOT) data set with many numbers, nearly 24,779 tweets are taken for pre-processing. The tweet has been divided into Hate speech, offensive language, and other classes. Then the given tweets are pre-processed by the BERT with a multi-layer perceptron in the Multilingual Hate-speech Detection Technique.

The tweets are pre-processed using BERT, as the upper-case letters are converted into lower-case letters, removal of tags, URLs, and stopping notations makes the machine easy to understand and validate the given tweet. The first part of the processed tweet holds the tweets with the capitalization of letters and the second part comprises pre-processed tweets without capitalization and has no tags like #, @, and so on. This simulation result has been obtained by applying the proposed technique. In the above simulation, the tweets have been pre-processed using BERT. Followingly, the Hinglish comment has been given to the mBART to convert them into English. Then forwarded to the similarity checker where the incomplete or misspelled words are vocabulary and spell-checked using an online checker forwarded to the log-likelihood test where the tweet's negative, positive or neutral nature is verified. Additionally, the sentiment of the tweet has been analyzed by the Profanity check technique. Here, with ReLu, sigmoid function, and logistic regression, the last validated words are compared with the nearby words in the sentence. The emotion has been found out whether it's negative, positive, or neutral. Finally using the above simulation, the input tweet has been classified into three sub-categories. When we come to Hate speech, 4744 words are optimized with 38,419 reference words, and from those, 4259 words are validated.

4.3 Performance evaluation metrics

The performance of the proposed technique has been evaluated in this subsection with parameters [44] such as Precision, Recall, Accuracy, and F1-score. Then this performance evaluation graph comprises the result with other techniques such as Bidirectional Encoder Representation from Transformer- Convolutional Neutral Network (BERT-CNN), Convolutional Neutral Network (CNN), Bidirectional Long-Short Term Memory (Bi-LSTM), Ternary Trans-CNN, Deep Learning (DL) Ensemble Stacked, Embeddings from Language Model (ELMO). Equations 8–11 show the mathematical formulas of these evaluation metrics.

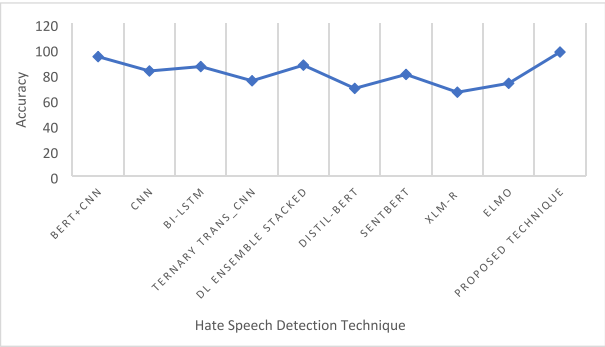


Fig. 6 Accuracy result of the HateDetetctor technique with previously Implemented techniques

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1SCORE = \frac{2 * Precision * Recall}{Precision + Recall} \tag{11}$$

TP: True Positive, FP: False Positive, TN: True Negative, and FN: False Negative.

4.3.1 Evaluation of accuracy

In Fig. 6, the performance of the proposed technique based on accuracy is shown. The degree of achieving the standard level of the calculation is accuracy by using Eq. 8. From the result, it has been clear that the accuracy attained to 97.6% when it has been analyzed with the existing technique like BERT-CNN has 94% accuracy, CNN has 82.7% accuracy,

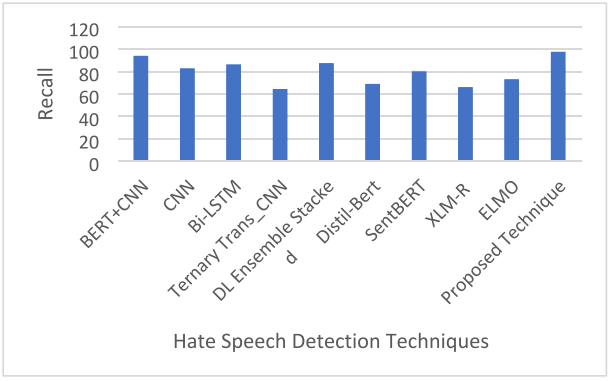


Fig. 7 Recall of the HateDetetctor technique with previously Implemented techniques

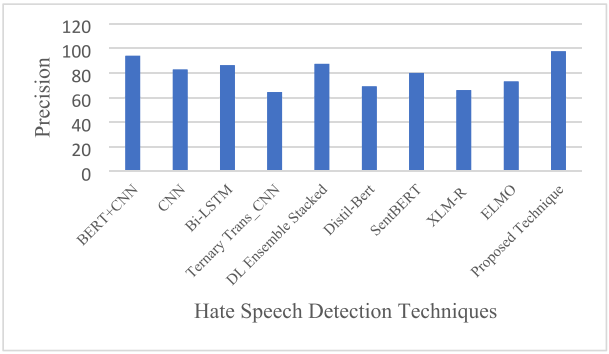


Fig. 8 Precision result of the HateDetetctor technique with previously Implemented techniques

Bi-LSTM has 86.2% accuracy, Ternary Trans-CNN has 75% accuracy, DL Ensemble Stacked has 87.3% accuracy, Disttil-BERT has 69%, SentiBERT has 80%, XLR-R has 66% and ELMO has 73% accuracy. On over-viewing it, the proposed technique reaches high in its accuracy and ELMO Technique was low in its accuracy.

4.3.2 Evaluation of recall

In Fig. 7 the recall evaluation of the proposed technique is shown. The capability to identify the positive tweets is recall. From the result, the recall value by using Eq. 10 of the proposed technique attained is 97.6% when it has been analyzed with the existing technique, BERT-CNN has 94%, CNN has 82.7%, Bi-LSTM has 86.2%, Ternary Trans-CNN has 64.4%, DL Ensemble Stacked has 87.3%, Disttil-BERT has 69%, SentiBERT has 80%, XLR-R has 66%and ELMO has 73%. On over-viewing it, the proposed technique performs high in recall and the Ternary Trans-CNN Technique is low in its recall.

4.3.3 Evaluation of precision

In Fig. 8, the Precision evaluation of the proposed technique is shown. The quality measure of positive samples is called precision. From the result, the precision value by using Eq. 9 of the proposed technique has been attained to 97.5% when it has been analyzed with the existing technique, BERT-CNN has 94%, CNN has 82.7%, Bi-LSTM has 86.2%, Ternary Trans-CNN has 64%, DL Ensemble Stacked has 87.4%, Disttil-BERT has 69%, SentiBERT

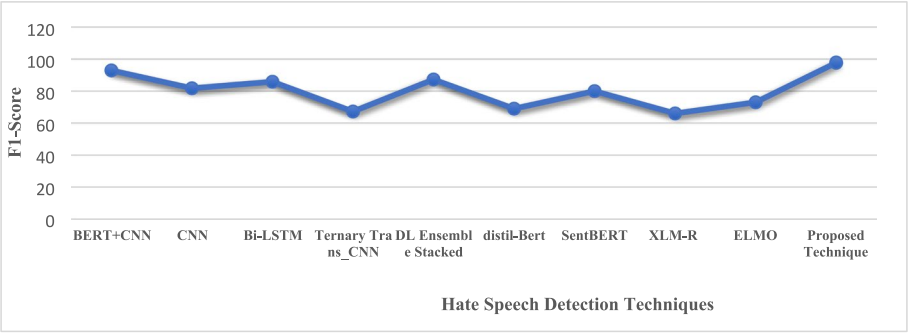


Fig. 9 F1-Score result of the HateDetetctor technique with previously Implemented techniques

Table 2 Comparison Analysis of the Accuracy of HateDetetctor technique with previously Implemented techniques

Hate Speech Models	Davidson dataset	Hinglish Hate dataset	Hot dataset
BERT + CNN [45]	94	94.6	95.6
CNN [46]	84.7	85.7	86.7
Bi-LSTM [47]	86.2	87.2	88.2
Ternary Trans_CNN [48]	84.4	86.4	87.4
DL Ensemble Stacked [49]	87.4	85.4	86.4
Distil-Bert [50]	69	69.1	69.5
SentBERT [51]	80	80.3	80.2
XLM-R [52]	66	66.2	66
ELMO [53]	73	75	78
Proposed Technique: HateDetector	**97.5**	**97.3**	**97.7**

has 80%, XLR-R has 66% and ELMO has 73%. The results show that the proposed technique gives high value of precision and the Ternary Trans-CNN Technique was low in its precision.

4.3.4 Evaluation of F1 score

Figure 9 shows the F1- Score result of the proposed technique, which is calculated using Eq. 11. The proposed technique has achieved 97.8% when it has been analyzed with the existing technique. The other technique like BERT-CNN has 94%, CNN has 82%, Bi-LSTM has 86%, Ternary Trans-CNN has 71%, DL Ensemble Stacked has 88%, Disttil-BERT has 69%, SentiBERT has 80%, XLR-R has 66% and ELMO has 73% on overviewing it, proposed achieved high in F1-Score performance and Ternary Trans-CNN Technique is low in its F1-Score.

4.4 Comparative analysis of classifiers performance

This subsection compares three different types of data sets with proposed and existing Techniques BERT-CNN, CNN, Bi-LSTM, Ternary Trans-CNN, DL Ensemble Stacked, and ELMO.

From Table 2, it shows the comparative analysis of proposed and existing techniques based on three different types of data sets: the Davidson Dataset, the Hinglish Hate Dataset, and the HOT dataset. The three types of datasets maintain the same level of accuracy in the proposed technique when compared with the other existing technique. Each technique taken for comparison has undergone certain deviations from its accuracy value for all three data sets. Here too, the accuracy does not change. The proposed technique has achieved high in its accuracy without any such deviations and the Ternary Trans-CNN and ELMO techniques have undergone more deviations when compared with others.

In Fig. 10 the comparison analysis of proposed and existing technique was viewed based on the above said three data set. The three types of datasets maintain the same level of

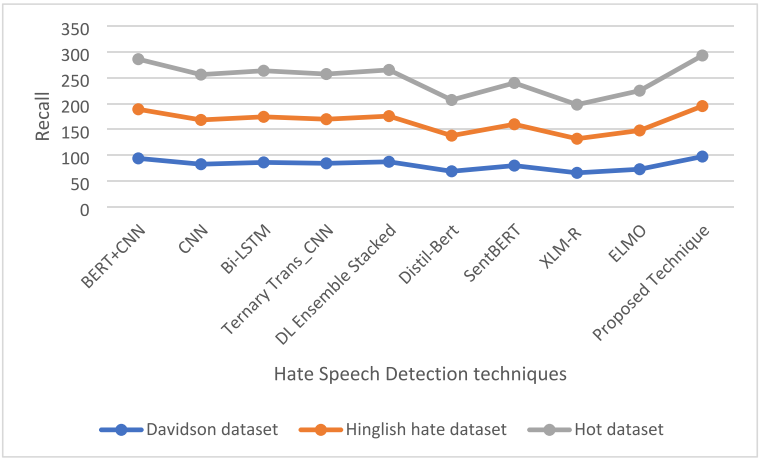


Fig. 10 Comparison Analysis of Recall of HateDetetctor technique with previous Implemented techniques

recall in the proposed when compared with the other existing technique. The proposed has achieved high recall without any such deviations and the CNN, and Bi-LSTM techniques have undergone more deviations when compared with others.

In Fig. 11 the proposed and existing technique is compared with three datasets. The proposed technique has achieved high precision too and all other techniques have highly deviated for all three data sets.

From Fig. 12 it has been clear that the proposed technique does not vary for any dataset given. The output result shows that all the three data set that works under this proposed brings out the output with a high-performance range. All technique has high deviations and Bi-LSTM, Ternary Trans-CNN, and DL Ensemble Stacked technique has equal deviations.

Thus, comparing all the parameters of the proposed technique for all three datasets, it attained the utmost high performance. The proposed technique shows an accuracy level of 97.9%, precision 97.9%, recall 98%and F1-score 97.8%. On comparing with the existing technique, namely Bidirectional Encoder Representation from

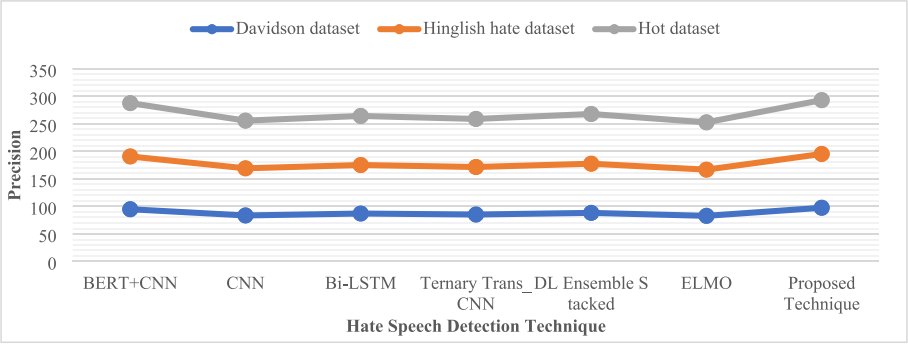


Fig. 11 Comparison Analysis of Precision

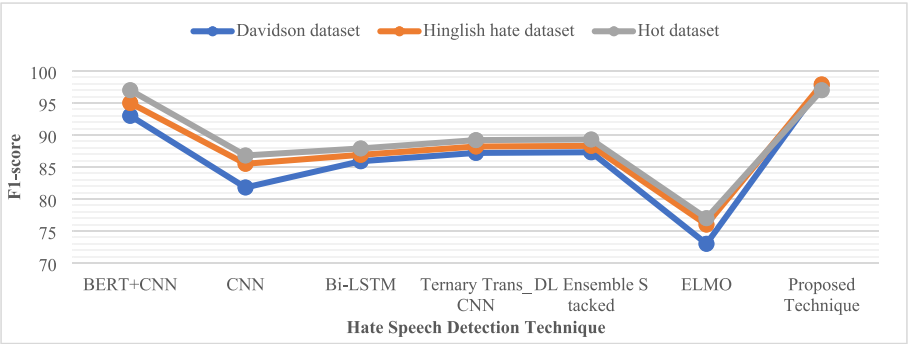


Fig. 12 Comparison Analysis of F1-Score of HateDetetctor technique with previous Implemented techniques

Transformer- Convolutional Neutral Network (BERT-CNN), Convolutional Neutral Network (CNN), Bidirectional Long-Short Term Memory (Bi-LSTM), Ternary Trans-CNN, Deep Learning (DL) Ensemble Stacked, Embeddings from Language Model (ELMO) the proposed technique gives higher and efficient performance.

5 Conclusion and future work

We have worked on the Detection and Analysis of Multilingual OHS in Twitter Using Artificial Intelligence with our proposed techniques. Multilingual Hate Speech Detection and Profanity Check is new as the step-by-step performance of several methods or techniques is carried over within these two proposed techniques. The Hinglish Tweets are pre-processed using BERT and coded using mBART, weighted by numbers using BOW with N-gram, and spell-checked then finally, the emotion is analyzed using the Profanity check technique which comprises of ReLu, sigmoid function, and logistic regression and finally the hate speech is detected. Thus, it is concluded that the performance of the existing technique BERT-CNN, CNN, Bi-LSTM, Ternary Trans-CNN, DL Ensemble Stacked, Disttil-BERT, SentiBERT, XLR-R and ELMO has been taken as a reference in performance evaluation and comparison was low, compared with the performance of the proposed which is ultimately high. This perfectly removes the code-mixing, with fast processing capability, is easy to train, and is highly reliable with a top performance of 97.9% of accuracy, precision, and F1-Score and 98% recall. For future work, the unlabeled data should be examined for the unsupervised machine learning model as labeling data is a very time-consuming task. Therefore, further study of the deep learning model is essential and advantageous to address hate speech problems.

Author Contribution Ms. Anjum
Dr. Rahul Katarya (Corresponding Author)
1. Conceived and designed the analysis
2. Collected the data
3. Contributed data or analysis tools
4. Performed analysis
5. Wrote the paper
6. All other necessary contributions related to this paper

Data availability Data sharing not applicable – no new data generated: Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Declarations

Conflict of interest There is "NO" Conflict of interest for this manuscript and with any author.

References

1. Dinakar K, Reichart R, Lieberman H (2021) Modeling the Detection of Textual Cyberbullying. Proceedings of the International AAAI Conference on Web and Social Media 5(3):11–17. https://doi.org/10.1609/icwsm.v5i3.14209
2. Zhang Z, Luo L (2019) Hate speech detection: A solved problem? The challenging case of long tail on Twitter. Semant Web 10(2019):925–945
3. Pereira-Kohatsu JC, Quijano-Sánchez L, Liberatore F, Camacho-Collados M (2019) Detecting and monitoring hate speech in twitter. Sensors (Switzerland) 19(21):1–37. https://doi.org/10.3390/s19214654
4. Guiora A, Park EA (2017) Hate Speech on Social Media. Philos (United States) 45(3):957–971. https://doi.org/10.1007/s11406-017-9858-4
5. Corazza M, Menini S, Cabrio E, Tonelli S, Villata S (2020) A Multilingual Evaluation for Online Hate Speech Detection. ACM Trans Internet Technol 20(2). https://doi.org/10.1145/3377323
6. Ousidhoum N, Lin Z, Zhang H, Song Y, Yeung DY (2020) Multilingual and multi-aspect hate speech analysis. EMNLP-IJCNLP 2019 - 2019. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing Conference, pp 4675–4684. https://doi.org/10.18653/v1/d19-1474
7. Sreelakshmi K, Premjith B, Soman KP (2020) Detection of Hate Speech Text in Hindi-English Code-mixed Data. Procedia Comput Sci 171(2019):737–744. https://doi.org/10.1016/j.procs.2020.04.080
8. Mandl T (2020) Overview of the HASOC Track at FIRE 2020 : Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German, pp 29–32. https://doi.org/10.1145/3368567.3368584
9. Mozafari M, Farahbakhsh R, Crespi N (2022) Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning. IEEE Access 10:14880–14896. https://doi.org/10.1109/ACCESS.2022.3147588
10. Ridenhour M, Bagavathi A, Raisi E, Krishnan S (2020) Detecting Online Hate Speech: Approaches Using Weak Supervision and Network Embedding Models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12268 LNCS:202–212. https://doi.org/10.1007/978-3-030-61255-9_20
11. Wang L, Niu J, Yu S (2020) SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis. IEEE Trans Knowl Data Eng 32(10):2026–2039. https://doi.org/10.1109/TKDE.2019.2913641
12. Alonso P, Saini R, Kovács G (2020) Hate Speech Detection Using Transformer Ensembles on the HASOC Dataset. In: Speech and Computer, pp 13–21
13. Wang Y, Huang G, Li J, Li H, Zhou Y, Jiang H (2021) Refined Global Word Embeddings Based on Sentiment Concept for Sentiment Analysis. IEEE Access 9:37075–37085. https://doi.org/10.1109/ACCESS.2021.3062654
14. Cao R, Lee RK-W, Hoang T (2020) DeepHate: Hate Speech Detection via Multi-Faceted Text Representations, pp 11–20. https://doi.org/10.1145/3394231.3397890
15. Mridha MF, Wadud MAH, Hamid MA, Monowar MM, Abdullah-Al-Wadud M, Alamri A (2021) L-Boost: Identifying Offensive Texts From Social Media Post in Bengali. IEEE Access 9:164681–164699. https://doi.org/10.1109/ACCESS.2021.3134154
16. Park JH, Fung P (2017) One-step and Two-step Classification for Abusive Language Detection on {T}witter. In: Proceedings of the First Workshop on Abusive Language Online, Aug., pp. 41–45. https://doi.org/10.18653/v1/W17-3006
17. Paschalides D, Stephanidis D, Andreou A, Orphanou K, Pallis G, Dikaiakos MD, Markatos E (2020) MANDOLA: A Big-Data Processing and Visualization Platform for Monitoring and Detecting Online Hate Speech. ACM Trans Internet Technol. https://doi.org/10.1145/3371276

18. Modha S, Majumder P, Mandl T, Mandalia C (2020) Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance. Expert Syst Appl 161:113725. https://doi.org/10.1016/j.eswa.2020.113725
19. Roy PK, Tripathy AK, Das TK, Gao X-Z (2020) A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. IEEE Access 8:204951–204962. https://doi.org/10.1109/ACCESS.2020.3037073
20. Maxime (2019) What is a Transformer?No Title. Medium. https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04
21. Horev R (2021) BERT Explained: State of the art language model for NLP Title. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
22. Mozafari M, Farahbakhsh R, Crespi N (2020) A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. Stud Comput Intell 881 SCI:928–940. https://doi.org/10.1007/978-3-030-36687-2_77
23. Mutanga RT, Naicker N, Olugbara OO (2020) Hate speech detection in twitter using transformer methods. Int J Adv Comput Sci Appl 11(9):614–620. https://doi.org/10.14569/IJACSA.2020.0110972
24. Qian J, Niu Z, Shi C (2018) Sentiment analysis model on weather related tweets with deep neural network. ACM Int Conf Proc Ser 31–35. https://doi.org/10.1145/3195106.3195111
25. Plaza-Del-Arco FM, Molina-González MD, Ureña-López LA, Martín-Valdivia MT (2020) Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. ACM Trans Internet Technol 20(2). https://doi.org/10.1145/3369869
26. Wang G, Wang B, Wang T, Nika A, Zheng H, Zhao BY (2014) Whispers in the dark: Analysis of an anonymous social network. In: Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC, pp 137–149. https://doi.org/10.1145/2663716.2663728
27. Qureshi KA, Sabih M (2021) Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text. IEEE Access 9:109465–109477. https://doi.org/10.1109/ACCESS.2021.3101977
28. Kapil P, Ekbal A (2020) A deep neural network based multi-task learning approach to hate speech detection. Knowl-Based Syst 210. https://doi.org/10.1016/j.knosys.2020.106458
29. Corazza M, Menini S, Cabrio E, Tonelli S, Villata S (2020) A Multilingual Evaluation for Online Hate Speech Detection. ACM Trans Internet Technol 20(2):1–22. https://doi.org/10.1145/3377323
30. Greevy E, Smeaton AF (2004) Classifying racist texts using a support vector machine. In: Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 468–469. https://doi.org/10.1145/1008992.1009074
31. Agarwal S, Sureka A (2017) But i did not mean it! - Intent classification of racist posts on tumblr. In: Proceedings - 2016 European Intelligence and Security Informatics Conference, EISIC 2016, pp 124–127. https://doi.org/10.1109/EISIC.2016.032
32. Davidson T, Bhattacharya D, Weber I (2019) Racial Bias in Hate Speech and Abusive Language Detection Datasets. In Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy. Association for Computational Linguistics, pp 25–35
33. Miró-Llinares F, Moneva A, Esteve M (2018) Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. Crime Sci 7(1):1–12. https://doi.org/10.1186/s40163-018-0089-1
34. Paetzold GH, Malmasi S, Zampieri M (2019) UTFPR at SemEval-2019 Task 5: Hate Speech Identification with Recurrent Neural Networks. International Workshop on Semantic Evaluation
35. Gambäck B, Sikdar UK (2017) Using Convolutional Neural Networks to Classify Hate-Speech. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada. Association for Computational Linguistics, pp 85–90
36. Liu Y et al (2020) Multilingual Denoising Pre-training for Neural Machine Translation. Trans Assoc Comput Linguist 8:726–742. https://doi.org/10.1162/tacl_a_00343
37. Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L (2020) Multilingual Denoising Pre-training for Neural Machine Translation. Trans Assoc Comput Linguistics 8:726–742
38. Bloomberg M (n.d.) Google, "Python programming language (version 3.6)." https://www.python.org/downloads/
39. Pedregosa F et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
40. Bisong E (2019) Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA, pp. 59–64. https://doi.org/10.1007/978-1-4842-4470-8_7
41. Davidson T, Warmsley D, Macy M, Weber I (2017) Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media. https://doi.org/10.1609/icwsm.v11i1.14955
42. Bohra A, Vijay D, Singh V, Akhtar SS, Shrivastava M (2018) A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. In Proceedings of the Second Workshop on Computational

Modeling of People's Opinions, Personality, and Emotions in Social Media, New Orleans, Louisiana, USA. Association for Computational Linguistics, pp 36–41
43. Mathur P, Sawhney R, Ayyar M, Shah R (2018) Did you offend me? Classification of Offensive Tweets in {H}inglish Language. In: Proceedings of the 2nd Workshop on Abusive Language Online ({ALW}2), pp 138–148. https://doi.org/10.18653/v1/W18-5118
44. Omar A, Mahmoud TM, Abd-El-Hafeez T (2020) Comparative Performance of Machine Learning and Deep Learning Algorithms for Arabic Hate Speech Detection in OSNs. Adv Intell Syst Comput 1153 AISC:247–257. https://doi.org/10.1007/978-3-030-44289-7_24
45. Zhou Y, Yang Y, Liu H, Liu X, Savage N (2020) Deep Learning Based Fusion Approach for Hate Speech Detection. IEEE Access 8:128923–128929. https://doi.org/10.1109/ACCESS.2020.3009244
46. Chaudhari A, Parseja A, Patyal A (2020) CNN based Hate-o-Meter: A Hate Speech Detecting Tool. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), pp 940–944. https://doi.org/10.1109/ICSSIT48917.2020.9214247
47. Jain R, Goel D, Sahu P, Kumar A, Singh JP (2021) Profiling Hate Speech Spreaders on Twitter.
48. Mathur P, Shah RR, Sawhney R, Mahata D (2018) Detecting Offensive Tweets in Hindi-English Code-Switched Language. Proc Ann Meet Assoc Comput Linguist 18–26. https://doi.org/10.18653/v1/w18-3504
49. Mazari AC, Boudoukhani N, Djeffal A (2023) BERT-based ensemble learning for multi-aspect hate speech detection. Cluster Comput 0123456789. https://doi.org/10.1007/s10586-022-03956-x
50. Ali R, Farooq U, Arshad U, Shahzad W, Beg MO (2022) Hate speech detection on Twitter using transfer learning. Comput Speech Lang 74:101365. https://doi.org/10.1016/j.csl.2022.101365
51. Madhu H, Satapara S, Modha S, Mandl T, Majumder P (2023) Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. Expert Syst Appl 215:119342. https://doi.org/10.1016/j.eswa.2022.119342
52. Ryzhova A, Devyatkin D, Volkov S, Budzko V (2022) Training Multilingual and Adversarial Attack-Robust Models for Hate Detection on Social Media. Procedia Comput Sci 213:196–202. https://doi.org/10.1016/j.procs.2022.11.056
53. STUFIIT at SemEval-2019 Task 5: Multilingual Hate Speech Detection on Twitter with MUSE and ELMo Embeddings. In Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 464–468, Minneapolis, Minnesota, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/s19-2082

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.

Pattern Analysis and Applications (2023) 26:1771–1791
https://doi.org/10.1007/s10044-023-01203-6

THEORETICAL ADVANCES

Hybrid ABC and black hole algorithm with genetic operators optimized SVM ensemble based diagnosis of breast cancer

Indu Singh[1] · K. G. Srinivasa[2] · Mridul Maurya[1] · Aditya Aggarwal[3] · Himanshu Sheokand[1] · Harsh Gunwant[4] · Mohit Dhalwal[1]

Received: 12 February 2023 / Accepted: 6 September 2023 / Published online: 26 October 2023
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

Abstract
Forever and a day, breast cancer has caused significant negative impacts on the quality of lives of number of women, more often than not turning into a fatal disease. The growth in the number of such cases has constantly been a major concern for the community as well as medical experts. To prevent irreversible damages caused by the disease, early identification of breast cancer is essential. Various researches and techniques have been devised in the past as an attempt to achieve this task with appreciable accuracy. As an advancement to these pre-existing algorithms and methods, we have devised a model by exploiting the techniques of nature-inspired metaheuristics in order to efficiently detect breast cancer at an early stage while maintaining acceptable levels of accuracy. In this paper, we propose a hybrid model, namely "hybrid artificial bee colony and black hole with genetic operators (GBHABC)", for the early detection of breast cancer. In the proposed model, we employed a support vector machine (SVM) ensemble technique, optimized using the proposed GBHABC model. This model combines the techniques of two major algorithms, namely artificial bee colony (ABC) and black hole (BH), guided through crossover and mutation genetic operators. Datasets from the well-known UCI breast cancer repository have been used to train the models and evaluate test result. For a fair and accurate evaluation of the model, a number of metrics have been examined including accuracy, sensitivity, specificity, $F1$-score and precision. An impeccable accuracy of 99.42% was obtained on the UCI dataset, clearly outperforming any literature in the same field.

Keywords Breast cancer diagnosis · Genetic operators · Artificial bee colony · Black hole algorithm · Ensemble learning

1 Introduction

Cancer is the uncontrolled regeneration and reproduction of untypical and unwanted cells in different parts of a body. Cancer really is a comprehensive term for a group of diseases caused when untypical or unusual cells grow up in different parts of the body, thus also known as neoplasm (literally meaning, a new growth). There are hundreds of strains of cancer diseases such as colon cancer, oral cancer, skin cancer, prostate cancer, breast cancer and lung cancer, depending upon the site of reproduction of the cells. Among all these types of cancers, breast cancer

✉ Indu Singh
indusingh@dtu.ac.in

K. G. Srinivasa
srinivasa@iiitnr.edu.in

Mridul Maurya
mridulmaurya99@gmail.com

Aditya Aggarwal
aditya99.iitd@gmail.com

Himanshu Sheokand
sheokandhimanshu@gmail.com

Harsh Gunwant
hgunwant@gmail.com

Mohit Dhalwal
mohitdhalwal12@gmail.com

[1] Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

[2] Department of Data Science and Artificial Intelligence, International Institute of Information Technology, Naya Raipur, Chhattisgarh, India

[3] Department of Mechanical Engineering, Indian Institute Of Technology, Delhi, India

[4] Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, GGSIPU, Delhi, India

(B-Can) is a sort of cancer with contrasting appearances among women. It is the most common type of cancer that affects women throughout the world [3]. A breast lump is the most classic presenting manifestation among women with B-Can and has a relatively outrageous predictive value for it to be malignant [27].

Based on histological and molecular evidences, B-Can is tabulated into three crucial sub-types based on the presence or absence of molecular indicators for progesterone (PR+) or oestrogen receptors (ER+) and human epidermal growth factor 2 (ERBB/HER2+): hormone receptor-positive (ERBB2-), ERBB2+ and triple-negative B-Can (TNBC) (lacking all 3 standard molecular markers). The treatment and the duration of it vary from the type of B-Can [50]. There are numerous ways of diagnosis and treatment of breast cancer. Some of the methods for treatment are radiotherapy, chemotherapy, immunotherapy (using immune cells already present in breast tissue) [17], combining both radiotherapy and immunotherapy [38], breast conservation therapy (BCT) and mastectomy which offer analogous survival and can be regarded as equivalent therapy in early-stage B-Can. The last two therapies, BCT and mastectomy, are the most conflicting ones for women to decide [18]. Digital mammography (DM) is the most recommended method for diagnosing breast tumours [33]. The diagnosis of breast cancer using IOT-enabled classifier systems is also available in the literature [31]. Following these methods and earlier diagnosis, there is an improvement in B-Can treatment leading to an increase in survival rate. However, such therapies and diagnosis may have some side effects, majorly in pregnant women who may have the chance of miscarriage and terminations [16]. Due to these reasons, patients have started refusing to undergo the treatment which increases the risk of mortality to 2.42 times higher according to the study [15].

Typically being the most commonly occurring cancer type among women and second-most periodic cancer overall, B-Can accounted for as much as 19.3 million new cases and 10 million cancer deaths around the world in the year 2020 (estimate by Cancer Journal) [46]. Though there are some studies that focus on male breast cancer, it still exists to be rare and understudied when compared to female breast cancer, which shows that the 5-year survival rate for male counterparts was lower than that for female patients. In the subcontinent of India itself, it accounted for almost 14 of total cancer cases among Indian women. Figures show that breast cancer has now outgrown even the most frequent forms of cancers including cervical cancer. Being on the rise in both rural and urban parts of the country, every four minutes an Indian woman is diagnosed with this disease in India with post-survival rates reported to be as low as 80 among US women and an even worse 60 among Indian women. B-Can is the second leading cause of cancer-associated mortality in the USA and has been reckoned that 1 out of 8 women will experience breast cancer at some point of their life [15].

It is well known to many that as the stages of this disease grow, the chances of survival of the patient diminish. With more than half of Indian women diagnosed with the disease suffering from stage three–four of breast cancer, the chances of survival seem very far-fetched.

It is seen that a lot of factors affect the chances of developing breast cancer in women. Correlation of different factors includes body mass index (BMI) [8], pregnancy obesity [25], alcohol consumption patterns and intensities [49], active–passive smoking [35], age groups [30] among others helping in more accurately evaluating the risks of developing breast cancer in later ages and chances of recovery. As close as 50% of the total cases fall in the age group of 25-50 years. Making the situation worse, more than 70% of the advanced-stage cases have low survival rates and high mortality rates. Various studies have been done in the past discussing the incidence and mortalities of female breast cancer, particularly for the Asia-Pacific regions.

Due to the most recent COVID-19 pandemic, the patients of B-Can are more susceptible to be infected by the virus, leading to more deaths. Also, due to lockdown and shortage of medical supplies, their treatments and therapies are given a pause, increasing the chances of augmentation of raging cancer cells. The aftereffects of this pandemic included high rates of anxiety, insomnia, distress and depression, especially in breast cancer patients [22]. To beat these effects as well as the disease itself, an active and healthy lifestyle is believed to be very helpful. Physical exercise has been shown to lower the risk of breast cancer, reduce the incidence of recurrence and improvise the survival rate of breast cancer patients. It can also aid with mental wellness and the immune system [10]. The low survival rates of breast cancer patients in the world are due to the failure in the detection of the disease in the early phase. As time passes and the stages of the disease progress, the chances of survival become less. Apart from spreading awareness, the only possible ways to change this situation are to develop and deploy some state-of-the-art accurate and efficient algorithms that can detect and raise alarms when the disease is at its early phase itself. Breast cancer can be cured and the survival rates plausible if detection is done in time.

1.1 Past approaches for detection in brief

Various works and researches have been done in the past for the early phase detection of cancer by the tools and algorithms of nature-inspired metaheuristics.

Kiyan et al. [26] proposed an algorithm which used radial basis function (RGF), general regression neural network (GRNN) and probabilistic neural network (PNN) to classify Wisconsin Breast Cancer Dataset (WBCD) and compare

the result with multilayer perceptron (MLP). Yeh et al. [55] devised a data mining method using the concept of discrete PSO. This method created a novel PSO that has a feasible system structure, and every particle is coded in positive integer numbers. Statistical analysis is used to eliminate feature variables that are insignificant.

Nazarian et al. [37] introduced a model for the classification of binary issues through utilizing information mining procedures, based on artificial bee colony algorithms.

Krawczyk et al. [28] presented a work in progress on a computer-aided breast cancer determination. They utilized a multi-objective memetic algorithm to choose the pool of one-class predictors that show simultaneously high diversity and consistency. Dheeba et al. [12] have advocated a PSO optimized WNN model. The model attempted to increase the classification accuracy of WNN by tuning the initial network parameter using PSO. By using PSOWNN, they evolved connected WNN and optimized it by using the best network architecture which was created by optimizing the momentum factor, learning rate and number of neurons in the hidden layer.

Bhardwaj et al. [4] came up with a new algorithm, genetically optimized neural network (GONN), for problem-solving. They modified the neural network to improve its architecture (structure and weight) of planning. They used the GONN algorithm to classify breast cancer tissue as malignant or benign. To demonstrate their results, they took a WBCD database.

Wang et al. [52] have proposed the CMWOAFS-SVM model which is an improved WOA based on the chaos and control engineering theories which are lodged to optimize two critical parameters for SVM with the RBF kernel and simultaneously select features in practical classification tasks. Similarly systems have also been developed for heart disease prediction and Brainstroke Classification using machine learning [21, 41].

1.2 Brief about our model

Machine learning models integrating metaheuristic optimization algorithms have been employed for various applications including detection of breast cancer. However, such models can often face issues like less generalization and overfitting and need better parameter optimization. Though metaheuristic algorithms have been effective in optimizing SVM ensembles and proposed better ability to find global optima for SVM, synchronization of multiple metaheuristic algorithms needs more exploration for optimizing the combination strategies in SVM ensembles. This should be implemented by considering ensemble diversity and performance improvement. The SVM can be made suitable for specific characteristics of the breast cancer dataset, by optimizing the parameters using the ABC algorithm for better model performance. In the model, we have employed the artificial bee colony that is based on the communication and cognitive skills of honeybees [24]. The issue with ABC is that it has an inherent problem of stagnation. This suggests that in the first phase, the location gradually changes, and so, the exploration phase moves at a slow pace. We have implemented an improved black hole [11] algorithm as a means to increase convergence of the model and to expand the exploration phase. The classic black hole algorithm possesses difficulty in maintaining a balance between the exploitation and exploration phases; thus, it was modified to tackle the very same limitation. It has a tendency to get stuck in the local optima. The improved black hole algorithm did not face any such limitation, since the stars orbiting around the focal point were at random angles and move to a better position, and therefore, exploration was done efficiently. This algorithm was incorporated into the first phase of the artificial bee colony's exploration, thus making it more efficient, hence establishing a balance between the exploration and exploitation phases. Once a star enters the event horizon, it is destroyed. Traditional black hole algorithms create points randomly in the search space which can possibly make the search space uneven and very different from the initial distribution. To rectify this drawback, we applied mutation and crossover genetic operators. The crossover operation is one of the well-known evolutionary algorithms, such as differential evolution (DE) [45], that can significantly increase an algorithm's effectiveness. Over the binary coded GA, we applied a two-point crossover. The extreme points were rectified using a two-point crossover. In order to arrive at a wider range of options, we employed the crossover operator. After crossover, we also employed swap mutation which is also one of the most prominent genetic operators. This makes sure that no two points, which are generated in different iterations, are exact replicas of each other. This model is then used to optimize a majority voting ensemble technique which classifies the data points as per the task. All these improvements over respective base algorithms work in harmony with each other and work towards combating their limitations. The proposed model can improve the overall performance of SVM ensemble and increase model accuracy. Better parameter optimization and reduction in errors in the proposed breast detection model can produce excellent results.

The contributions of research work are summarized as follows:

1. ABC algorithm is synchronized with improved black hole algorithm to make it more efficient and overcome their respective drawbacks.
2. Crossover and mutation genetic operators are used to generate new stars instead of randomized generation to counter the limitation of black hole algorithm, thereby

making a diversified population and a better newly generated star.
3. A novel GBHABC model is proposed to optimize the weights of the ensemble model for the best possible solution.
4. Majority voting ensemble is used to set priorities for different kernels available in the model.

2 Related work

In the past, a diverse range of datasets were utilized for the early diagnosis of various diseases such as diabetes, breast cancer, hepatitis and others. Many studies used the Wisconsin Prognostic Breast Cancer Dataset (WBCD) to detect breast cancer. This dataset is still used to create and validate various algorithms in the field of breast cancer detection today. In an earlier study, the use of data mining techniques in conjunction with nature-inspired algorithms was seen to solve diagnosis and detection challenges. Some of the major state-of-the-art techniques are illustrated in Table 1.

Yeh et al. [55] proposed a data mining technique using the discrete PSO on WBCD dataset. The proposed method created a novel PSO with a feasible system structure, with every particle being coded in positive integer numbers. Statistical analysis was used to eliminate insignificant feature variables, and after that, only six feature variables were remaining. The proposed hybrid data mining method reduced the computational complexity and sped up the whole data mining process. The proposed method showed satisfactory results on a range of performance metrics including sensitivity, accuracy and specificity.

Chen et.al. [6] showed the superiority of a hybrid SVM-based model rough set-support vector machine (RS-SVM) over other relevant state-of-the-art algorithms for tackling the problem of breast cancer diagnosis. In the proposed RS-SVM model, the RS feature reduction algorithm was employed for feature selection to eradicate the redundant features of the dataset and add to the diagnostic accuracy of the model by SVM. RS feature reduction algorithm was employed on the whole dataset for optimal subset generations, and only a small fraction was retained for the complete study. The combination subset of features used by the model for categorizing the breast tumours was pinned down to be the most promising feature of the reduction algorithm. Moreover, experiments done on altered slices of the WBCD showed that the RS-SVM model performed satisfactorily in distinguishing between the two classes of breast tumour. The performance of this model greatly depended on the optimal parameters of the SVM. Thus, the need for new algorithms to accurately calculate the optimal parameters arose. Moreover, a single dataset like WBCD was used for evaluating the model, thus undermining the accuracy and real-life application of this model.

Chen et al. [7] proposed the PSO-SVM Classifier for breast cancer diagnosis on the WBCD dataset. They incorporated feature selection and parameter optimization in the suggested technique. The proposed approach is divided into two parts. The first stage involves using the PSO algorithm to firmly conduct SVM parameter optimization with feature selection. The second stage involves conducting a classification job with the help of a selected feature and an ideal value using the tenfold approach and the SVM model. In the proposed model, the particle swarm framework simultaneously solved the issue of feature selection and model selection in SVM. For the designing purposes of the objective function of particle swarm optimization a weighted function was adopted in which simultaneously accounted for a number of support vectors, selected features and average accuracy rates of SVM, but the model was not feasible in practical diagnosis as it costed a lot of CPU time for PSO-based system training.

Nazarian et.al [37] put forward a novel classification method by using artificial bee colony algorithm for classification. This algorithm was another straightforward, new, simple and proficient algorithm dealing with real-world scenarios for optimization. In this paper, analysis of this ailment is conceived through utilizing data related to B-Can samples. The aftereffects of this technique were contrasted with different strategies for breast cancer determination.

Krawczyk et al. [28] proposed a CAD-based breast cancer determination, tested on genuine and proven medical datasets of 675 pictures, gathered from 75 patients out of which 25 were amiable, 25 were dangerous and 25 were fibroadenoma. In the study, each patient was addressed by 9 pictures chosen by a pathologist self-assertively. They proposed a productive medical choice support structure that permitted recognizing amiable, dangerous and fibroadenoma cases from the dataset. The nuclei recognition methodology depends on the firefly calculation algorithms. The methodology produced nuclei markers that were utilized in the marker-controlled watershed division. In the final step, an image recognition step was finished by a novel classifier. The model was built utilizing a one-class classifier trouped for a multi-class decomposition, permitting to capture proficiently the unique attributes of inspected classes and henceforth to precisely classifying them. The outcomes introduced in this paper showed that a modernized medical determination system dependent on the proposed strategy would be successful and could give precise diagnostic data.

Sheikhpour et. al. [44] proposed a PSO-KDE-based model: hybridizing the PSO and classifier based on non-parametric kernel density estimation for breast cancer determination on the WBCD. In this model, they used PSO to simultaneously find a subset of features that could minimize a number of features as well as the classification error in the

Table 1 Related work comparison

Authors	Year	Dataset	Proposed algorithm	Benefits	Limitation
Chang et al. [55]	2009	WBCD	A new hybrid approach of discrete PSO and statistical method for mining breast cancer pattern	Reduced computational complexity	Only patterns and correlations may be generated from the data, and their validity cannot be validated. Inaccurate or skewed data gathering might result in erroneous statistical analysis
Hui-Ling Chen et al. [6]	2011	WBCD	Feature selection using rough set-based SVM classifier for breast cancer diagnosis	A subset of features was the most promising feature of the reduction algorithm	The results and accuracy of this model greatly depend on the parameter tunings of the SVM
Yang et al. [7]	2012	WBCD	SVM parameter optimization with feature selection using PSO and classification using SVM with the help of tenfold technique	Particle swarm framework simultaneously solved the enigma of feature selection and model selection in SVM	For practical diagnosis, problem costs a lot of CPU time for PSO-based system
Nazarian et al. [37]	2013	WBCD	Classification of breast cancer using the ABC algorithm	This strategy is easier than different strategies	The success percentage in this proposed method is comparatively low
Krawczyk & Filipczuk [28]	2014	675 pictures were gathered from 75 patients of the Regional Hospital in Zielona Góra, Poland	Analysis of cytological image using hybrid one-class classification decomposition with firefly nuclei detection	Utilizing one-class classifier troupes for a multi-class decomposition permits us to seize proficiently the unique attributes of inspected classes and in result to precisely separate between them	Use of small dataset could lead to overfitting. Fewer data points can improve training set accuracy, but reduce test set accuracy as the model generalizes. An overfit statistical model fits its training data exactly
Dheeba et al. [12]	2014	Real clinical database of 216 mammograms collected from mammogram screening centres	PSO optimized WNN model for determination of breast cancer on mammograms	Improvement in the classification accuracy of computer-aided analysis of the digital mammogram	Accuracy of the model is 93.671%. It cannot be practically used for the diagnosis because of the low accuracy
Sheikhpour et al. [44]	2016	WBCD, WDBC	PSO for bandwidth determination and feature selection of KDE-based classifiers for diagnosis of breast cancer	The proposed approach KDE-PSO has a better mean performance than the mean performance of GA-KDE model in breast cancer detection	The PSO used for minimizing the classification error and number of features is not multi-objective
Shahbeig et al. [43]	2017	Breast cancer microarray dataset	Selection of the associated genes with breast cancer using fuzzy multi-objective hybrid TLBO-PSO approach	Using this model, they were able to enhance the sensitivity to 90.55%, specificity to 93.33% and accuracy to 91.88% by choosing 195 genes	This method can only be used to do prognosis at very early stages only
Wang & Chen [52]	2020	Wisconsin Breast Cancer Dataset (WBCD), Diabetes Data Set, Dermatology Data Set	Medical diagnosis using chaotic multi-swarm whale optimizer boosted SVM	Tackle feature selection and parameter optimization simultaneously	It has a complex search process because it is based on multi-swarm which has stratified mechanisms

Table 1 (continued)

Authors	Year	Dataset	Proposed algorithm	Benefits	Limitation
Thawkar et al. [48]	2021	Database of 651 mammograms obtained from Digital Database for Screening Mammography (DDSM)	Hybrid feature selection method based on butterfly optimization algorithm and ant lion optimizer	Achieves high level of accuracy while using fewest possible features	It failed to show better accuracy on WBC dataset and result based on ANFIS sometimes provide inaccurate prediction as they converge slowly
Afolayan et al. [2]	2022	Wisconsin breast Cancer dataset (WBCD)	PSO optimized using decision tree machine learning technique for breast cancer diagnosis	Efficient and beneficial for decision making	Accuracy of the model is 92.26%, and hence, it cannot be practically used for the diagnosis because of the low accuracy

classifier based on kernel density estimation. This model contained feature selection using PSO and classification with the selected feature subset and optimal kernel bandwidth determination. KDE-PSO model had better average performances than the average performance of the GA-KDE model in breast cancer diagnosis. Though results could show the supremacy of this model over other proven methods, the PSO used for minimizing the classification error and number of features was not for multi-objective function. On the contrary, some researchers have also proposed a TLBO-PSO [43]-based model in their study having a multi-objective optimization with two primary objectives: the first one for attaining the highest classification accuracy and the second to select the least number of informative genes. They used the breast cancer microarray dataset for implementation of their study. Using the proposed TLBO-PSO model, they were able to enhance various performance metrics in the study including sensitivity, accuracy and specificity.

Another study proposed by Wang et al. [53], based on CNN- and CAD-based methods, was proposed in a study that used a dataset of 219 patients with 614 ABUS—automated breast ultrasound—volumes that included 745 cancer regions, and 144 healthy women with a total of 900 volumes, without abnormal findings. In this study, a 3D CNN architecture was proposed for CAD cancer diagnosis in ABUS volumes. The proposed system used the densely deep supervision (DDS) mechanism to learn more discriminative features for malignant representations of cancer cells and to enhance the sensitivity of detection, and utilize the recommended threshold map with a unaccustomed threshold loss to streamline the cancer probability map for the false-positive reduction while maintaining a high record of sensitivity metric. Although the proposed detection system presented astonishing results, it still posed a few limitations. In the model, cystic lesions and fatty lumps would be misclassified as cancerous lesions and small cancerous lesions may be neglected by considering them a common ultrasound shadow.

Wang et al. [52] proposed the CMWOAFS-SVM model which was an enhanced WOA model based on the control and chaos engineering theories, proposed to optimize two principal parameters for SVM with the RBF kernel and simultaneously select features in practical classification tasks. They tested the prototype on WBCD, diabetes and dermatology datasets. The proposed model consisted of two parts: the first part included feature selection and parameter optimization of SVM that were simultaneously conducted by an improvised WOA algorithm. In the latter part, to evaluate the suitability of SVM for the classification task using a tenfold CV analysis, a subset of the optimal parameter pair and feature was obtained from the initial part. The proposed model could prove supremacy over other models by better performances and feature subset sizes, and the ability to tackle feature selection and parameter optimization simultaneously. Along with these astonishing results, it is not obscured that CMWOA failed to have any advantage in terms of computational cost over the original WOA. Moreover, it also has a complex scavenging process being based on multi-swarm having stratified mechanisms.

Some authors relied more on other datasets than WBCD, seen in studies like one done by Dheeba et. al. [12]. They proposed a PSO optimized WNN model on the real-world clinical database of 216 mammograms collected from mammogram screening centres. In this model, they have attempted to enhance the classification accuracy of WNN by tuning the initial network parameter using PSO. By using PSOWNN, they evolved connected WNN and optimized it by using the best network architecture which was created by optimizing the momentum factor, learning rate and the number of neurons in the hidden layers. The proposed PSOWNN classifier showed improvement in the classification accuracy of CAD analysis of the digital mammogram for diagnosis of breast cancer.

3 Methods and data

This section is divided into four primary subsections: a discussion of the artificial bee colony algorithm, a description of the black hole algorithm, a discussion of genetic operators and a discussion of the WBCD dataset.

3.1 ABC algorithm

Algorithm 1 Pseudo Code for ABC Algorithm

Initialization:
  Population of stars with a random location in the search space using the formula:

$$x_{ij}; \text{ where,}(i = 1, 2, \ldots, SN) \text{ and } (j = 1, 2, \ldots, D)$$

  Evaluate the fitness function fitiof all solutions in the population.
  Set trial = 0
The Employed Bee Phase
for $i = 1 \;\; TO \;\; SN \;\; STEP \; 1$ **do**
  Update a new candidate solution v_i

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj})$$

  Apply a greedy selection process between v_i and x_i to select the better one
  if solution x_i does not improves **then**
    $trial_i = trial_i + 1$
  else
    else $trial_i = 0$
  end if
end for
Calculate the probability p_i

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n}$$

The onlooker bee phase
for $(i = 1 \;\; TO \;\; SN \;\; STEP \;\; 1)$ **do**
  if $rand(0, 1) \leq p_i$ **then**
    Update a new candidate solution v_i for the onlookers bees
    Evaluate the fitness value of the candidate solution v_i
    Apply a greedy selection process between v_i and x_i to select the better one
    if the solution does not improve **then**
      $trial_i = trial_i + 1$
    else
      $trial_i = 0$
    end if
  end if
end for
The scouts bee phase
if max $(trial_i) >$ limit **then**
  Replace x_i with a new randomly produced candidate solution by using:

$$x_i^j = x_{min}^j + rand(0, 1) \times (x_{max}^j - x_{min}^j)$$

end if
if stop criterion is satisfied **then**
  Stop and Output the best solution achieved
else
  return to Step 16
end if

One of many nature-inspired algorithms, "artificial bee colony (ABC)" [24] proposed by Karaboga, is inspired by honeybees' foraging behaviour. The algorithm proposes to see any possible solution to the problem as a food source for the bees, whereas the amount of nectar in the food source corresponds to the solution's quality/fitness obtained. The artificial bee colony consists of three groups, namely employed bees, scouts and onlooker bees. The employed bees exploit local information to extract the locally best possible solution. Once all employed bees are done with the search process, onlooker bees evaluate and choose a food source based on the information supplied by the employed bees. This process is then used until a global best possible solution is found to the optimization problem. The proposed equation, to produce a candidate food (feasible possible solution) source from the information of an old one present in memory, is:

$$v_{\mathrm{ij}} = x_{\mathrm{ij}} + \Phi_{\mathrm{ij}}(x_{\mathrm{ij}} - x_{\mathrm{kj}}), \tag{1}$$

where $k \in \{1, 2, ..., \mathrm{SN}\}$ and $j \in \{1, 2, ..., D\}$ are the randomly chosen indices. Each solution denoted by $x_i (i = 1, 2, \ldots, \mathrm{SN})$ is a D-dimensional vector, where D is the number of optimization parameters. Φ_{ij} is a randomly chosen value lying between [−1, 1], controlling the production of neighbouring food sources around x_{ij} and representing a comparison among two food positions as seen by a bee. A food source is chosen by an artificial onlooker bee based on the probability value attached with that food source, p_i given by Eq. (2).

$$p_i = \mathrm{fit}_i / \sum_{n=1}^{\mathrm{SN}} \mathrm{fit}_n, \tag{2}$$

where fit_i is the fitness value of the solution i and SN is the number of food sources which is equal to the number of employed bees.

In this algorithm, scout bees replace an old food source, of which the nectar is rejected by the bees, with a new food source. This process is simulated by the algorithm by generating a random position in dimension space and producing a position at random and replacing that with the rejected food source. In the ABC algorithm, after a prefixed number of cycles, if the quality of food source cannot be improved any further, then that food source is abandoned by the bees. This prefixed number of cycles, called as limit for abandonment, is a very important and sometimes a deciding control parameter in determining the efficiency and accuracy of the algorithm. Taking the abandoned source as x_i, while $j \in \{1, 2, \ldots, D\}$, the scout bees discover a new food source to be replaced with x_i. This operation can be defined by using the following equation:

$$x_i^j = x_{\mathrm{min}}^j + \mathrm{rand}(0, 1)\left(x_{\mathrm{max}}^j - x_{\mathrm{min}}^j\right), \tag{3}$$

The ABC algorithm is tried to be represented with a pseudocode shown in Algorithm 1.

3.2 Black hole algorithm

Algorithm 2 Pseudo Code Black Hole Algorithm

1: **Initialization:** Population of stars with a random location in the search space using the formula:

$$P(t) = \{x_1^t, x_2^t, x_3^t, \ldots, x_n^t\}$$

2: These randomly generated stars are placed in the search space of function.
3: **for** $i = 1$ to the number of stars **do**
4: **for** each i^{th} star objective function is evaluated according to the following formula **do**

$$f_i = \sum_{i=1}^{pop_size} eval(p(t)$$

$$f_{BH} = \sum_{i=1}^{pop_size} eval(p(t)$$

5: **end for**
6: **end for**
7: Selecting the best star that has the best fitness value as the black hole
8: **while** max iteration or convergence criteria is not met **do**
9: **for** $i = 1$ to numbers of stars **do**

$$X_i(t) = X_i(t) + \text{rand} \times (X_{BH} - X_i(t))$$

10: **if** $f_i > f_{BH}$ **then**
11: $\text{Swap}(X_i), X_{BH})$
12: **end if**
13: Replace the new fitness value of the star (X_i) with the previous value
14: Update and Calculate :

$$R_{BH} = \frac{f_{BH}}{\sum_{i=1}^{N} f_i}$$

15: **if** $\sqrt{(X_{BH} - X_i)} < R_{BH}$ **then**
16: replace X_i with a new star in an optional location in the search scope
17: **end if**
18: **end for**
19: **end while**

The black hole algorithm [29, 19, 20] is a bio-inspired population-based method having similar features with another population-based algorithm. This method was inspired by the black hole phenomenon. A black hole is essentially a region of space with so much mass concentrated inside it that if an entity approaches it, there is no way of escaping its gravitational pull. If anything falls into black hole including light, then that object is forever gone from the universe. In the black hole algorithm, we start with an initial population of candidates that are a feasible solution to optimization problems, for every iteration, the best candidate out of the population is selected, and the remaining becomes the normal star after the initialization process the normal star starts moving towards black hole, i.e. best candidate. Whenever a normal star is swallowed by the black hole then for replacing the candidates that have entered the black hole new candidates are generated in the search space.

3.2.1 The absorption rate of stars by a black hole

After the initialization, absorption of the star around the black hole is started and all the normal stars start moving to the black hole. The absorption rate of stars by a black hole is calculated as follows:

$$X_i(t) = X_i(t) + \text{rand} \times (X_{\text{BH}} - X_i(t)) \tag{4}$$

where i=1,2,3,4…. n, X_i^t and X_i^{t+1}, respectively, are the location of ith star at t and t+1. X_{BH} is the location of the black hole in the search space and rand is a randomly generated number between the interval of [0,1] and N is the number of normal stars.

3.2.2 Probability of crossing the event horizon during moving star

We use the likelihood of crossing the event horizon (E_H) of the black hole while moving the normal star towards the black hole to obtain the most optimal data from the problem's search space. Every star which crosses the event horizon is absorbed by the black hole that means it dies and a new star is generated in the search space. To calculate the radius of the event horizon, the following equation is used:

$$R_{\text{BH}} = \frac{f_{\text{BH}}}{\sum_{i=1}^{N} f_i} \tag{5}$$

where f_{BH} and f_i are the fitness values of black hole and ith star, respectively, is the number of candidate stars. The pseudo-code for black hole algorithm is shown below in algorithm.

3.3 Genetic algorithms

Genetic algorithms (GAs) are adaptive heuristic search algorithms that are a subset of evolutionary algorithms. They are working on a concept based on natural selection and genetics. The random search is given historical data to direct it into the region of greater performance in the solution space. They are frequently utilized to produce high-quality solutions to optimization and search challenges. GAs work on a population of simulated chromosomes. Typically, these are binary coded. Each chromosome represents a solution to a problem and has a fitness, which is a real number that indicates how good a solution it is to the specific problem. It is a population-based search method that employs the survival of the fittest idea. A genetic algorithm uses operators that instruct the algorithm to solve a particular issue. During the search procedure, the GAs employed a number of operators. Encoding scheme, crossover, mutation and selection are among these operators.

3.3.1 Encoding scheme

In many computational issues, the encoding technique, or the method used to change data into a certain form, is critical. The information provided must be transformed into a certain bit string. The encoding strategy is different depending on the issue domain. They can be used in neural networks to identify the best weights, generate programmes or expressions and so forth. Binary, octal, hexadecimal, permutation, value-based and tree encoding techniques are well known.

3.3.2 Selection techniques

In the genetic algorithm, selection plays a vital role in selecting whether or not a certain string will participate in the reproduction process. It basically chooses parents who will mate and recombine to produce offspring for the future generation. Parental selection is critical to the GA's convergence rate because good parents lead individuals to better and fitter solutions. This operator is also known as the reproduction operator. The rate of convergence of GA is determined by the selection pressure.

3.3.3 Crossover operators

Reproduction and biological crossover are analogous to the crossover operator. In crossover, more than one parent is chosen, and one or more offspring are generated using the parents' genetic material. Crossover is typically used with a high probability in a GA. Single-point, two-point, k-point, uniform, partially matched, order, precedence preserving crossover, shuffle, reduced surrogate and cycle are examples of well-known crossover operators. These crossover operators are fairly generic, and they can also be used to create a problem-specific crossover operator.

For our proposed model, we have used a two-point crossover. The concept of two-point crossover is identical to that of one-point crossover. Instead of a single cut point, two are randomly selected at the same places in two parents, and two offspring are formed from them.

Process for a two-point crossover:

(i) Choose two cut locations at random between each parent's genes.
(ii) Copy the sub-string between the two cut places from Parent 2 to Offspring 1, as indicated in Fig. 1.
(iii) The remaining values are copied from the first parent and placed exactly as they are in Offspring 1, excluding duplicate values. (Genes 1 and 8 have already been inserted in the middle of the Offspring 1 in Fig. 1. As a result, by being copied to it, they are omitted.)
(iv) Empty areas in Offspring 1 are filled with genes that did not occur in the first parent, but did appear in the second parent after the second cut point, as illustrated in Fig. 2. (As gene 4 comes after the second cut point of Parent 2, it appears first in Offspring 1, followed by gene 5.)

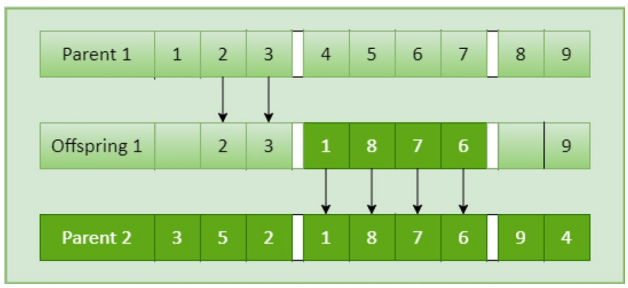


Fig. 1 Diagram representation of two-point crossover

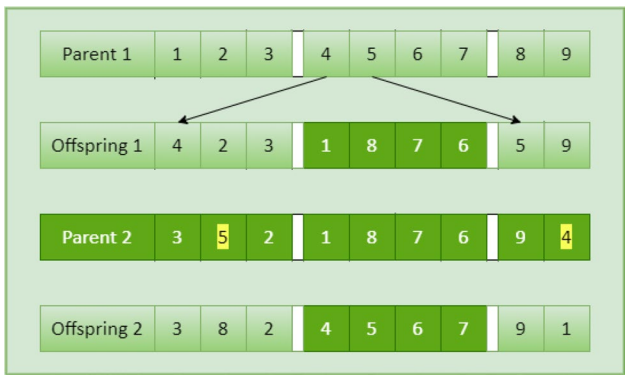


Fig. 2 Diagram representation of two-point crossover

(v) Similarly, Offspring 2 can be formed by swapping the roles of the parents as seen in Fig. 2, which is represented as Offspring 2.

3.3.4 Mutation operators

Mutation is a genetic operator that preserves genetic variation from one population to the next. It can be characterized as a minor random change in the chromosome that results in a new solution. It is typically used with a low probability. The GA is simplified to a random search if the probability is sufficiently high. The "exploration" of the search space is related to mutation. In several circumstances, it has been discovered that mutation is required for GA convergence, whereas crossover is not. Displacement, simple inversion and scramble mutation are well-known mutation operators.

3.4 SVM

SVM is one of the most successful binary classification algorithms due to its excellent generalization capabilities. SVM was developed in the 1990 s by Vapnik et al. [9] with the goal of minimizing structural risk and making the model very resistant to overfitting. If a model classifies testing or unknown data with a lower accuracy than the training data, it is said to be overfit.

By creating a hyperplane or a group of hyperplanes, SVM divides the data into distinct classes. A hyperplane is defined in SVM as a surface that entirely divides data points into two classes so that no data point from one class falls into the other [47].

The hyperplane with the greatest distance to the nearest data points in both classes, out of an unlimited number of potential hyperplanes, is deemed the most optimum [9].

A hyperplane is the set of points $\mathbf{x}$ satisfying the equation:

$$\mathbf{w} \cdot \mathbf{x} + \mathbf{b} = 0 \tag{6}$$

where b is the bias and $\mathbf{w}$ is the weight vector which in linear case can be easily solved using Lagrangian multipliers.

SVM may efficiently classify high-dimensional data by mapping it onto a high-dimensional feature space such that the mapped data can be readily separated using lower-order polynomial functions, despite its origins in linearly separable situations. A kernel function is used to map original data into a high-dimensional feature space [5].

The following common kernels are used in SVM classifier:

1. Linear kernel:

$$f(\alpha_i, \alpha_j) = \alpha_i \cdot \alpha_j \tag{7}$$

2. Polynomial kernel:

$$f(\alpha_i, \alpha_j) = (\beta \alpha^t \alpha_j + r^2)^2 \tag{8}$$

3. RBF kernel:

$$f(\alpha_i, \alpha_j) = e^{[\gamma ||\alpha_i - \alpha_j||^2]} \tag{9}$$

4. Sigmoid kernel:

$$f(\alpha_i, \alpha_j) = \tan h(\beta \alpha_i^t \alpha_j + r) \tag{10}$$

3.5 WBCD dataset

Breast cancer is one of the major reasons for the death of women around the world. It is an absolute urgency of finding treatment solutions and some accurate diagnosing. To deal with this problem, the raw breast cancer data are taken from Wisconsin Diagnosis Breast Cancer dataset from UCI machine learning repository. Features in the dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image [14].

These data are preprocessed by replacing the missing values with 0 instead of removing them as it should not introduce a noticeable impact on statistical data patterns of the dataset. The dataset is too small for our use after dividing the dataset into training and testing subsets. Another common issue with the dataset is skewness. So, in order to resolve the small dataset size issues and the data unbalance new data are generated of the same dimensions by introducing normal distributed random noise. Almost 75% of data are considered for training and 25% for testing. A piece of detailed information about the WBCD dataset is presented in Table 2.

4 Proposed model

ABC faces a notable challenge in the form of stagnation, where the exploration phase experiences a gradual progression in the initial stages due to the evolving nature of the location. This indicates that the exploration process encounters difficulties in advancing at a desirable pace during

Table 2 Description of the WBCD dataset

Attribute number	Attribute description	Minimum value	Maximum value	Standard deviation	Mean value	Total count
1	Clump thickness	1	10	2.82	4.42	699
2	Uniformity of cell size	1	10	3.05	3.13	699
3	Uniformity of cell shape	1	10	2.97	3.21	699
4	Marginal adhesion	1	10	2.85	2.81	699
5	Single epithelial cell size	1	10	2.21	3.22	699
6	Bare nuclei	1	10	3.64	3.46	699
7	Bland chromatin	1	10	2.44	3.44	699
8	Normal nucleoli	1	10	3.05	2.87	699
9	Mitoses	1	10	1.71	1.59	699

the early phases of ABC. Finding an appropriate balance between the exploitation and exploration stages is one of the main difficulties addressed by the basic black hole algorithm. The algorithm frequently runs into problems because it tends to become stuck in local optima, which makes it difficult for it to adequately traverse the search space, which subsequently results in sub-optimal solutions. Another property of black hole algorithm is that when a star crosses the event horizon it gets destroyed. To keep the number of data points same as initial it generates new points randomly. Some areas of the search space may become densely inhabited as a result of the algorithm's random creation of points, while other areas may remain mostly unexplored. As a result, the algorithm can find it difficult to fully explore the whole search area, which could result in biased findings or a failure to find the best answers. The algorithm's efficiency and efficacy may be hampered by the uneven distribution produced by the random point creation, making it more difficult for it to successfully navigate the search environment. All these shortcomings led to the development of the proposed algorithm.

The architecture of the proposed GBHABC (hybrid artificial bee colony and black hole with genetic operators) model is given in Fig. 3. The pseudo-code for GBHABC is illustrated in Algorithm 3. In the first step, a support vector machine (SVM) is applied. SVM gives the best outcome when we use the RBF kernel function for classification. Different kernel functions can be specified for the decision function. Common kernels are provided, but we can also specify custom kernels. We are training six different kinds of SVM experts with varied values of RBF parameters. Values designated to the RBF variable for different SVM classifiers are [0.1, 0.2, 0.5, 1, 2, 5]. Later, in the result section, it has been shown that based on the selected RBF value, the performance measures of each SVM classifier differ. In the second step, the outcomes of the SVM expert system are ensemble using weighted majority voting. Littlestone and Warmuth [34] showed the minimization of errors made by the ensemble system, by providing weights to the majority voting procedure. Based on their classification accuracy, all the experts are provided with some weights.

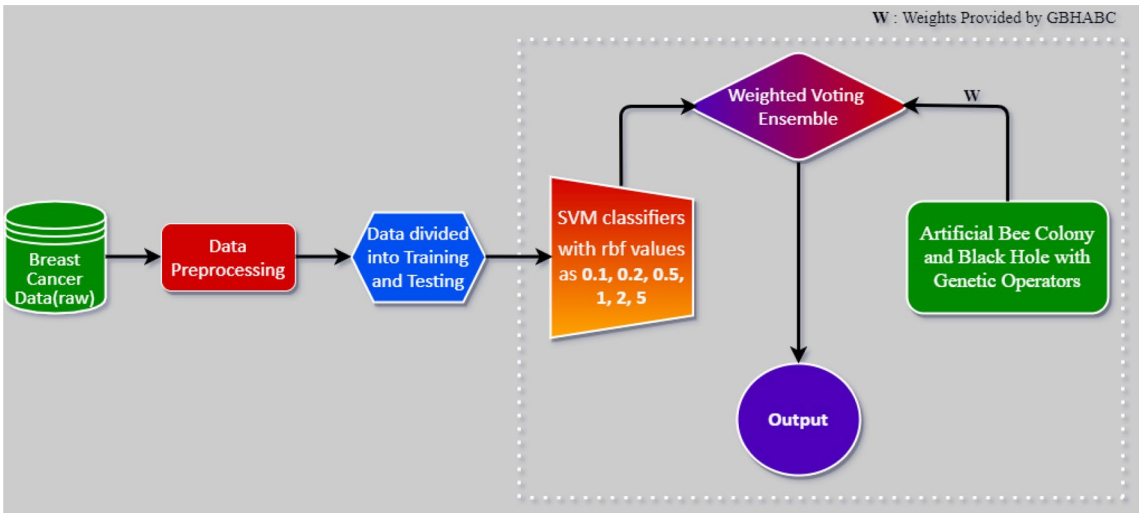


Fig. 3 Architecture of proposed GBHABC

Algorithm 3 Pseudo Code For Proposed Model

Initializing Parameters: m = Maximum Number of Cycles, d = Dimension of the Problem, α = Swarm Size, l = Lower Bound, μ = Upper Bound, n, a;
Initialise the swarm having food sources/stars, x_i where $(i = 1, 2, \ldots, \alpha)$ by using $Eq.(1)$;
t=1;
while $t < m$ **do**
    Employed Bee Segment:
    Evaluate fit_i for each x_i
    Select the best star from the population as BH(black hole)
    for each x_i **do**
        Move the star toward the BH using $Eq.(2)$
        Re-evaluate fit_i
        if $(fit_i < fit_{BH})$ **then**
            Set x_i as BH
        end if
        Calculate the RBH(event horizon radius) using $Eq.(3)$
        if $(t/m * 10\%2) = 0$ **then**
            $e_r = 0.65$
        else
            $e_r = 0.35$
        end if
    end for
    for each x_i **do**
        if $(x_i - x_{BH}) < R_{BH}$ **then**
            Remove x_i
            if $r < e_r$ **then**

                r: random number in the uniform range of $[0, 1]$
                Perform two-point crossover to generate new star
            else
                Generate the new star randomly using $Eq.(1)$
                Add the generated star to the population
            end if
        end if
        if $fit_i < fit_{BH}$ **then**
            Set x_i as BH
        end if
    end for
    Onlooker Bee Segment:
    Scout Bee Segment:
    $t = t + 1$
end while
return x_{BH}

We have proposed GBHABC model to generate weights for the six (SVM) experts for weighted majority voting. ABC has the limitation of being stagnant due to which the ABC algorithm gets stuck to the local optima or converges prematurely. So, black hole is introduced in order to make ABC more explorative. To create a simulation of the black hole phenomena, all candidate solutions are chosen as stars, and the solution with the highest fitness among the candidates is chosen as the black hole. The flowchart of our proposed algorithm is represented in Fig. 4.

4.1 Initialization and updation of positions

All the solutions are initialized in a random manner in the search space as shown in Eq. (11).

$$x_{ij} = x_{lb} + \text{rand}[0, 1](x_{ub} - x_{lb}) \tag{11}$$

where x_{ij} represents the ith food source in the populous in $j$thdimension. x_{ub} and x_{lb} are the upper bound and the lower bound, respectively, for x_{ij} and rand[0, 1] is a uniformly scattered arbitrary number in the bounds [0, 1].

After initializing the BH (black hole) and the stars, the updation of positions by the solutions (stars) is done using the direction and the distance of the BH. The position of the stars is updated by the equation expressed by Eq. (12).

$$x_i(t + 1) = x_i(t) + \text{rand}(x_{BH} - x_i(t)) \tag{12}$$

where $x_i(t + 1)$ and $x_i(t)$ represent the position of the ith solution during iteration $t + 1$ and t, respectively, x_{BH} is the position of the black hole (best solution) in the search space and rand[0, 1] is a uniformly scattered arbitrary number in the bounds [0, 1].

If a solution (star) achieves a better position than the black hole, their positions are swapped, and a new star is added as the BH of the search space. Furthermore, if a star advances towards the BH, it has the potential to cross the BH's event horizon (EH). Crossing the horizon signifies that the star will be absorbed in the BH, i.e. eliminated, and a new star will be generated, i.e. a new solution in the search space will be produced.

4.2 Generation of a new star

When a star is absorbed by the black hole, a new star is created using one of the following strategies:

1. A new star is created at random with a probability of 0.65 (by using uniform distribution).

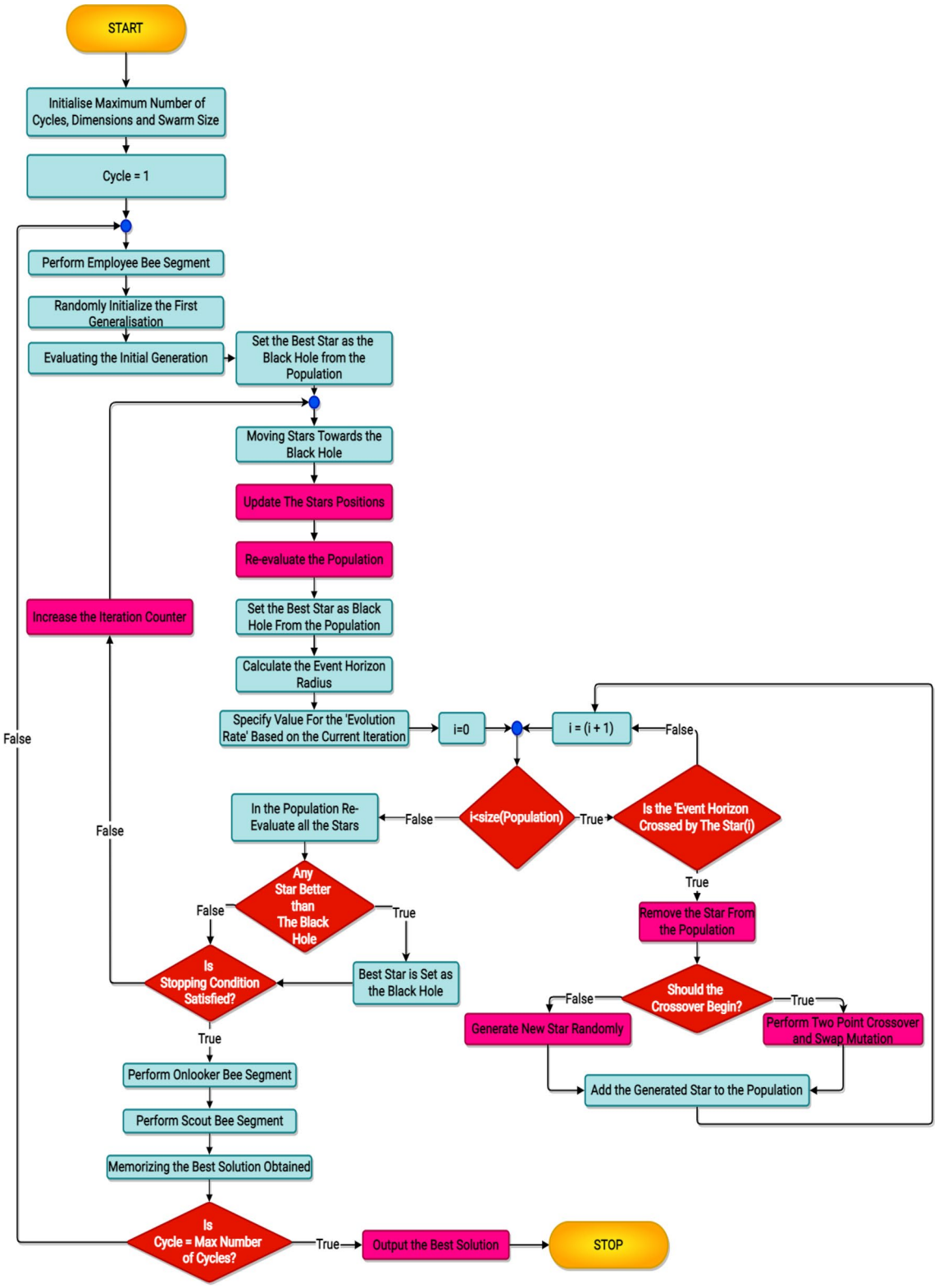


Fig. 4 Flowchart of the proposed GBHABC

2. A new star is created by selecting two stars from the population and recombining their attributes using a recombination procedure. The best of them will be added to the population later on. This approach is used with a probability of 0.35 at first.

The probabilities for regeneration [0.65, 0.35] change during the progress of the model. For instance, using the 1st strategy the model starts with a 0.65 probability of generating stars. When the ratio of the iterations to the total number of iterations reaches 0.2, BH changes the probability for generating stars using the 2nd strategy to be 0.65 and sets the probability to be 0.35 for the 1st strategy. It will alternate again when the ratio reaches 0.4. The alternation of probabilities will continue for each 20% progress of the iterations until the stop condition is satisfied. By implementing the above strategies, we can alternate the focus on exploitation and exploration during model progression and achieve the best balance between them in order to find the optimal solution.

4.3 Recombination method for generation of a new star

According to the fitness value, the current population is sorted and divided into 2 subpopulations, i.e. p_{male} and p_{female}. Randomly one star is selected from both, and a two-point crossover is applied over the selected stars. Then, among the two resulting stars created by crossover, the best one is selected. Finally, swap mutation is performed on the newly generated star to have diversity and even distribution in the search space. This star is added to search space which makes the search space better and proposed model efficient.

Then the subsequent phases of the model follow the same steps as ABC. Similar to ABC, GBHABC is also divided into three segments which are improved employed bee segments where the BH phenomenon with the genetic operators is applied, the onlooker bee segment and the scout bee segment. All other phases except the first one are kept similar as in the original ABC algorithm.

5 Experimentations and results

This section compares and contrasts the findings achieved using the suggested model and those obtained without it. First, there will be a quick review of the assessment measures. As previously stated, our research consists of two key experiments. The outcomes of traditional approaches will be given and debated in the first experiment. The outcomes of the suggested model are then presented and discussed in the second experiment.

5.1 Evaluation metrics

One of the most crucial steps is the comparison of a model by evaluation metrics. By utilizing and studying these metrics, we can analyse the performance of the proposed model and compare it with the existing methods. For this regard, 8 metrics have been considered, which are sensitivity, precision, specificity, $F1$ score, accuracy, FPR, FNR, MCC computed using the true positive (TP), false negative (FN), false positive (FP) and true negative (TN) rates. The confusion matrix was created for different RBF values.

As mentioned above, 8 metrics have been used to compute the performance of various RBF kernels of SVM. The different values taken for RBF in SVM were 0.1,0.2,0.5,1.0,2.0 and 5.0. They were calculated using Eqs. (13)–(20):

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{13}$$

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{(\text{TP} + \text{TN})} \tag{14}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{15}$$

$$F1 = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})} \tag{16}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{17}$$

$$\text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})} \tag{18}$$

$$\text{FNR} = \frac{\text{FN}}{(\text{FN} + \text{TP})} \tag{19}$$

$$\text{MCC} = \frac{(\text{TP} \cdot \text{TN}) - (\text{FP} \cdot \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{20}$$

5.2 First experiment

This experiment discusses the results obtained by using different RBF values for SVM algorithm. As previously mentioned in our first experiment, 699 records of whole WBCD dataset

Table 3 Confusion matrix for various SVM RBF values

Algorithm	TP	FP	FN	TN
SVM 0.1	61	0	5	109
SVM 0.2	61	0	8	106
SVM 0.5	61	0	15	99
SVM 1.0	61	0	24	90
SVM 2.0	61	0	30	84
SVM 5.0	61	0	52	62

were used. The confusion matrices are shown in Table 3 that shows the results for different RBF values of SVM. The table shows the confusion matrix of the testing set used for this experiment. 25 per cent of the dataset is used for this step to verify the feasibility of SVMs in the current study. It can be observed that in the best case, as shown in Table 3, 61 malignant records were detected as malignant and were classified as TP, and in the best case, 109 benign records were detected as benign and were classified as TN. Moreover, in the best case, 0 benign records were classified as malignant and 52 malignant records were classified in the benign class. The calculated $F1$ score, specificity, sensitivity, precision and accuracy values are illustrated in Table 4, for training steps.

According to Table 4, the evaluation metrics with the greatest values were produced in SVM with RBF value 0.1. It was able to obtain a 97.14 per cent accuracy rate. It can also be observed that the different RBF values tested were 0.1, 0.2, 0.5, 1.0, 2.0 and 5.0. There are two important factors in RBF SVM that determine the final results: gamma(γ) and C. The C parameter strikes a compromise between the proper classification of training examples and the maximization of the decision function's margin. If the decision function is better at reliably categorizing all training points, a narrower margin will be acceptable for higher values of C. A lower C encourages a larger margin and, as a result, a simpler decision function; however, this comes at the cost of training accuracy. This means that the C parameter functions as a model regularizer. The gamma(γ) parameter controls the impact of individual training samples on the overall model, which influences the final output. The gamma(γ) parameter has a big influence on how the model behaves. If gamma is set too high, the radius of the support vectors' sphere of action is reduced to the support vector itself, and no amount of regularization using C can avoid overfitting. To make the procedure efficient and less difficult, just the gamma(γ) parameter was adjusted for multiple SVM kernels based on these data. The parameters used during the experiment are listed with their values in Table 5.

Table 5 Parameter settings of the proposed GBHABC algorithm

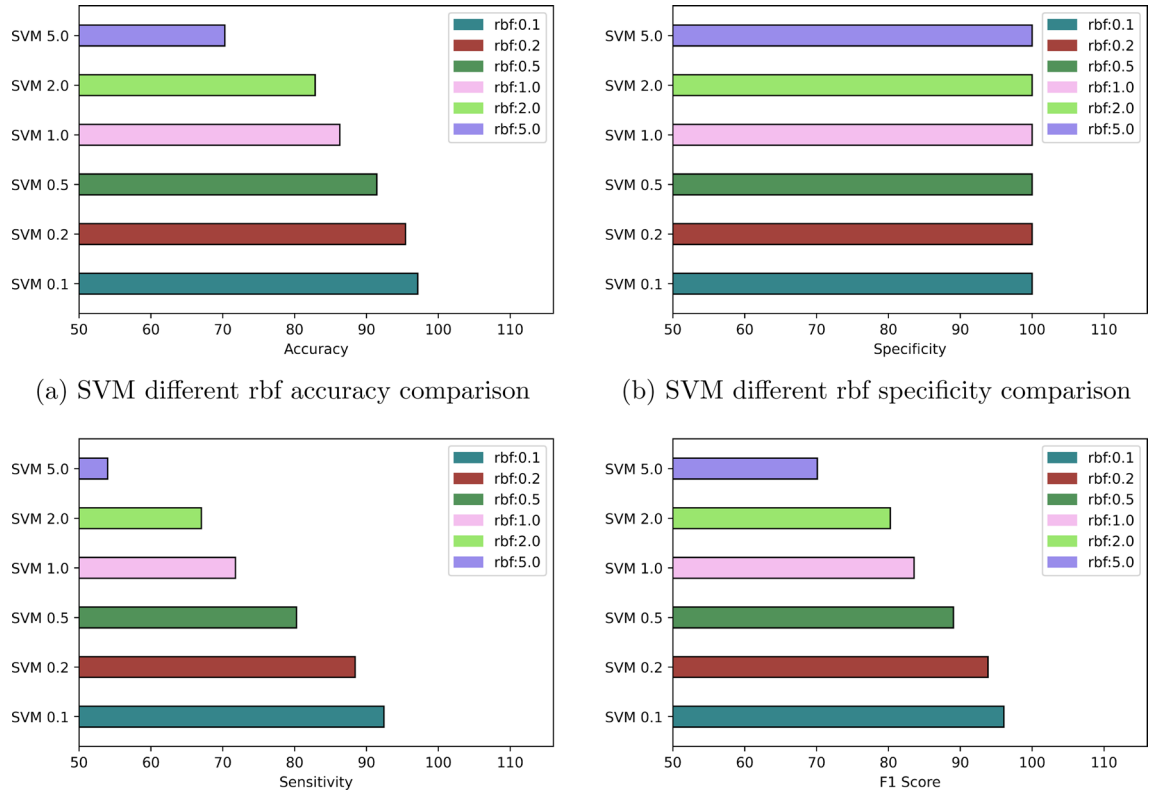
Parameters	Settings
Maximum iterations (Max_iter)	50
Lower bound (lb)	− 1000
Upper bound (ub)	1000
Dimensions (dim)	12
SearchAgents_no	5
npop	10
A	1

Figure 5a–d shows the comparison between above-discussed SVMs' according to different evaluation metrics. Figure 5a compares all the kernels according to their accuracy, data are given in Table 4. Similarly, Fig. 5b–d compares specificity, sensitivity and $F1$ score, respectively. These graphs confirm that changing gamma from 0.1 to 5.0 affects the outcomes of the models. Table 4 shows that when the value of gamma(γ) is increased from 0.1 to 5.0, the accuracy reduces from 97.14 per cent to 70.29 per cent. Similarly, sensitivity and f1 score also decrease with increasing the gamma(γ) value. On a fundamental level higher value of gamma(γ) makes the model more forgiving, but trades off the accuracy and other metrics. On the other hand a lower value of gamma(γ) is more stern on the search space selection and has higher accuracy. Both higher value and lower value of gamma(γ) has their benefits and limitations. So in the next experiment we utilize the maximum potential of these different kernels by using them in cohesion with each other with the help of reliable ensemble learning and our proposed hybrid model to improve the results further.

A little discussion regarding the relevance of the features of the dataset is furthered. According to Table 6, the most essential aspects for identifying BC were bare nuclei, uniformity of cell shape, normal nucleoli, uniformity of cell size and bland chromatin, whereas the least important criteria for identifying BC were single epithelial cell size and

Table 4 Evaluation metrics of SVM

Algorithm	Accuracy	Specificity	Sensitivity	Precision	$F1$ Score	FPR	FNR	MCC
SVM 0.1	97.14	100.0	92.42	100.0	96.06	0.00	0.08	0.94
SVM 0.2	95.43	100.0	88.41	100.0	93.85	0.00	0.12	0.91
SVM 0.5	91.43	100.0	80.26	100.0	89.05	0.00	0.20	0.83
SVM 1.0	86.29	100.0	71.76	100.0	83.56	0.00	0.28	0.75
SVM 2.0	82.86	100.0	67.03	100.0	80.26	0.00	0.33	0.70
SVM 5.0	70.29	100.0	53.98	100.0	70.11	0.00	0.46	0.54



(a) SVM different rbf accuracy comparison

(b) SVM different rbf specificity comparison

(c) SVM different rbf sensitivity comparison

(d) SVM different rbf F-1 score comparison

Fig. 5 SVM performance metrics

Table 6 Coefficient for each attribute

No	Features name	Coefficient for attributes
1	Bare nuclei	0.28
2	Uniformity of cell shape	0.23
3	Normal nucleoli	0.14
4	Uniformity of cell size	0.12
5	Bland chromatin	0.10
6	Mitoses	0.07
7	Clump thickness	0.05
8	Single epithelial cell size	0.02
9	Marginal adhesion	0.00

marginal adhesion. Table 6 shows the relative relevance of all aspects using the coefficient characteristics. For example, the most crucial trait for identifying breast cancer is the coefficient value of Bare nuclei, which is 0.28. With a value of 0.00, the coefficient value of marginal adhesion has the lowest relevance of all the attributes. These discoveries might be extremely beneficial to doctors who treat patients with breast cancer. In other words, taking into account the most essential characteristics can increase the quality of breast cancer diagnosis while lowering the mistake rate.

5.3 Second experiment

The next phase of our study was to implement our proposed model to optimize the weights of an ensemble model which uses different SVM kernels discussed in experiment 1. As discussed before, SVM kernels with varying gamma (γ) values have different results and benefits. So, we propose that we use all these different kernels in a weighted voting ensemble model with weights of the model being optimized by our proposed GBHABC model.

Finally we evaluated our model on the same metrics on which individual SVM kernels were evaluated. For even more clarity, results of this experiment are on a larger pool of dataset than the previous experiment, but the general distribution of the dataset and the dataset itself were same. Our model achieved an accuracy of 99.42%, sensitivity of 98.41%, specificity of 100.00%, precision of 100.00% and F1 score of 99.20%. These results show that our proposed

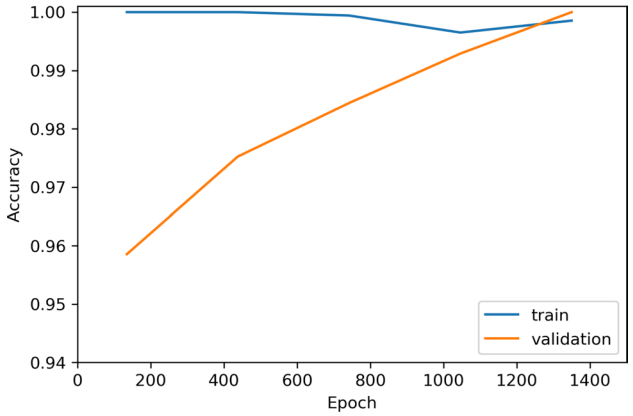


Fig. 6 Accuracy vs epochs for training and validation steps

approach has improved the overall results of the model on the Wisconsin dataset.

Figure 6 shows the accuracy vs epochs comparison for training and validation steps. The y-axis represents accuracy in the range from [0, 1] which on multiplication with 100 can easily be converted to percentage measure. The x-axis represents the number of epochs. Training step is denoted with a blue line and the validation step is denoted with an orange line. As observed, the training accuracy remains almost constant throughout all the epochs, only dipping by a mere 0.57% during the entire process. Whereas the validation accuracy improves consistently from a value around 95% to 99% and during the process, it becomes equal to the training accuracy and finally settles down. This graph proves that our model handles overfitting really well and is generalized for more data points rather than just being able to classify only the training examples.

We performed our experiment on different dataset distributions. The distributions were 50–50, 60–40, 75–25 as training and test sets. We calculated accuracy, sensitivity, specificity, precision and *F*1 score for the above-mentioned dataset distributions. The results of the experiment with different dataset distributions are given in the Table 7. As observed from the table, the accuracy and other evaluation metrics were very much close to each other, and thus, it can be inferred that the model is stable for different splits of the data.

Equations (13)–(20) were used for final evaluation of our model. The final results for the proposed hybrid GBHABC on the basis of accuracy, specificity, sensitivity, precision, *F*-1 score, FPR, FNR and MCC are 99.42%, 100.00%, 98.41%, 100.00%, 99.20%, 0.00, 0.02 and 0.99, respectively, and the same is listed in Table 8. These results signify that the proposed model is superior to the classical SVM models. The confusion matrix for the given results is shown in Table 9. Figure 7 shows the visual representation of the obtained results for better and easy understanding. All these results imply that the proposed GBHABC model had a positive impact on the detection of breast cancer by improving the ability to correctly classify all the cases.

Table 7 Dataset distribution

Dataset distribution	Accuracy	Specificity	Sensitivity	Precision	*F*1 score
50–50	98.78	98.58	98.04	97.39	98.26
60–40	98.93	98.80	98.09	97.78	98.46
75–25	99.42	100.00	98.41	100.00	99.20

Table 8 Evaluation of our model

Metric	Value (%)
Accuracy	99.42
Specificity	100.00
Sensitivity	98.41
Precision	100.00
*F*1 score	99.20
FPR	0.00
FNR	0.02
MCC	0.99

Table 9 Confusion matrix of our proposed GBHABC

		Predicted class	
		Malignant	Benign
Actual class	Malignant	124	0
	Benign	2	224

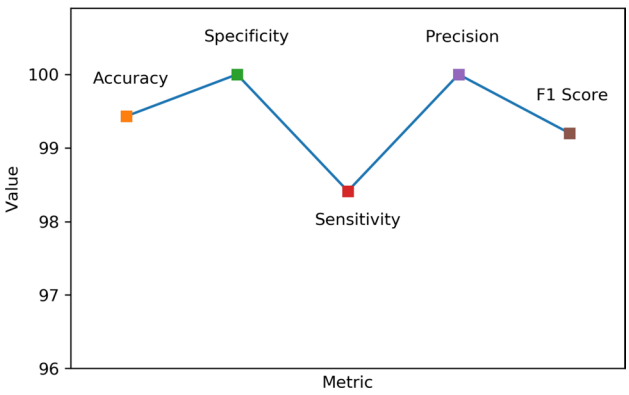


Fig. 7 Metrics representation

Table 10 Comparison of different methods applied for breast cancer prediction

S. No	Study	Year	Algorithm	Mean accuracy
1	Liu and Li [32]	2010	SVM-CPBK	93.26
2	Senapati and Dash [42]	2013	LLWNN	98.14
3	Zheng et al. [57]	2014	K-means with SVM	97.38
4	Nguyen et al. [39]	2015	GSAM wavelets	97.40
5	Murat Karabatak [23]	2015	Weighted NB	98.54
6	Sheikhpour et al. [44]	2016	PSO-KDE	98.45
7	Dora et al. [13]	2017	GNRBA	98.46
8	Zarbakhsh et al. [56]	2017	ANFIS with AR	99.26
8	Wang et. al. [51]	2018	WAUCE	97.10
9	Wuniri et al. [54]	2019	BIG-F	97.10
10	Punitha et al. [40]	2019	IABC-EMBOT	97.53
11	Abdar et al. [1]	2020	NE	98.07
12	Thawkar et al. [48]	2021	BOAALO	98.16
13	Afolayan et al. [2]	2022	PSO-DT	92.26
14	**Proposed model**	**2023**	**GBHABC**	**99.42**

Bold indicates our proposed model

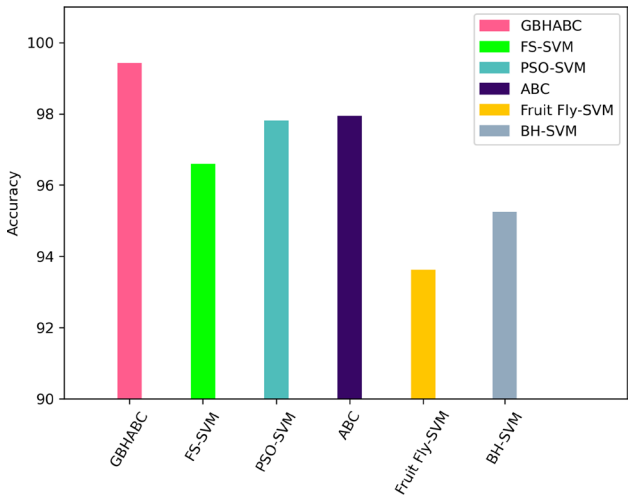


Fig. 8 Proposed vs old accuracy comparison

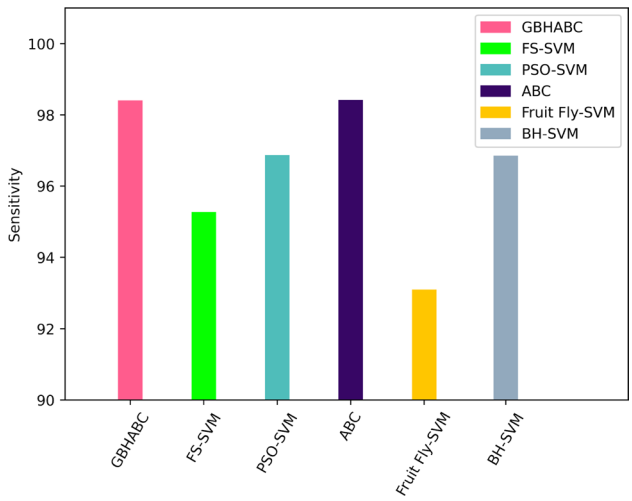


Fig. 9 Proposed vs old sensitivity comparison

5.4 Comparison with previous studies

We are going to expand on matters a bit by mentioning that we compared the findings obtained by utilizing the proposed methodology, with those that were already reported in the prior study. These studies had differing results, but all focused on the prediction of breast cancer incidence. Table 10 given in our study is used to represent the present levels of accuracy in our model's development, compared to those of other models built in the past.

Some of the models were able to achieve significantly higher accuracy on the breast cancer detection study. Sheikhpour et al. [44] proposed an approach by applying particle swarm optimization (PSO) with nonparametric KDE (kernel density estimation) and achieved an accuracy of 98.45%. Abdar et al. [1] achieved a analogous accuracy of 98.07% by implementing a branched ensemble approach that utilized the stacking and voting techniques among the classifiers. Similarly, Nahato et al. [36], Zarbakhsh et al [56], Punitha et al [40] achieved 98.6%, 99.26% and 97.53% accuracy, respectively.

Figures 8, 9, 10 shows the comparison between the proposed model and some of the previous studies with the help of graphs that display a cross-sectional comparison in the form of a bar graph with different hybrid algorithms that

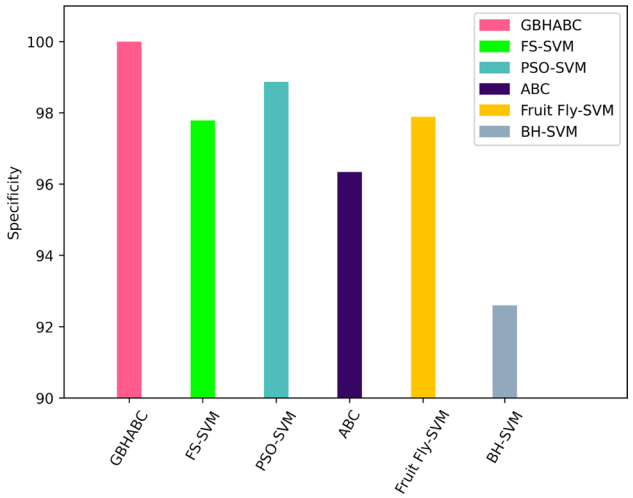


Fig. 10 Proposed vs old specificity comparison

existed in the past. A tabular representation of the comparison on the evaluation metrics like accuracy, sensitivity and specificity of the proposed model GBHABC has been done with previously existing hybrid algorithms on the WBCD breast cancer dataset as shown in Table 11. Various hybrid algorithms like FS-SVM, PSO-SVM, ABC, BH-SVM and fruit fly-SVM were taken into account for the comparison with GBHABC. All of these algorithms were run on our preprocessed WBCD dataset for comparison with the proposed research, and the results were analysed. Among the aforementioned algorithms, FS-SVM had an accuracy of 96.60%, specificity of 95.27% and sensitivity of 97.79%. PSO-SVM had an accuracy of 97.81%, specificity of 98.87% and sensitivity of 96.87%. ABC had an accuracy of 97.95%, specificity of 96.34% and sensitivity of 98.42%. BH-SVM had an accuracy of 95.25%, specificity of 92.60% and sensitivity of 96.85%. Similarly fruit fly-SVM had an accuracy of 93.62%, specificity of 96.53% and sensitivity of 93.10%. Our model GBHABC has an accuracy of 99.42%, a sensitivity of 100.00% and a specificity of 98.41%. GBHABC definitely appears to be superior when compared to the other tests, according to its accuracy, sensitivity and specificity. Thus, we can conclude that the proposed model GBHABC is quite effective for an early diagnosis of breast cancer and better than already present methodologies.

6 Conclusion and future scope

According to the findings of the present study, we have suggested the GBHABC (artificial bee colony with black hole and genetic operators) model for the detection of breast cancer by distinguishing tumours as malignant or benign at an early stage of development. To do this, we used the well-known WBCD dataset from the University of California, Irvine dataset repository to get raw breast cancer data. The proposed GBHABC-SVM model consists of six SVM classifiers (differentiated by the RBF parameter), which are combined into a weighted voting ensemble to improve performance.

It consisted of two primary experiments. The SVM method with three major parameters, C, ϵ and γ , and was used in the first experiment. Here, a basic SVM model, a polynomial SVM (with default parameters) and multiple modified SVMs with different C, ϵ, γ and values were compared. Our findings show that optimizing these parameters improves SVM performance. We utilized the first experiment's optimal parameters of C, ϵ, γ and in subsequent studies. As part of the second experiment, we combine all of these distinct kernels into a weighted voting ensemble model, with the weights of the model being optimized using our suggested model. A new model was introduced under the name of GBHABC. Our final findings show that the suggested technique allows us to achieve an accuracy, precision, sensitivity and f1-score of 99.42%, 100%, 98.41% and 99.20% and, as a consequence, to be highly successful in terms of predicting breast cancer. In conclusion, we believe the GBHABC model should be used in medical research for breast cancer prediction.

It would be intriguing to do a preliminary investigation of the model on various kinds of medical data in the future. In future, the integration of this hybrid model into multi-class classification is extendable to other datasets and various types of cancer detection. Furthermore, alternative machine learning models can leverage diverse forms of data, such as images and X-ray signals, for cancer detection. Moreover, future research endeavours may also encompass the examination of early cancer detection and the anticipation of cancer staging and can be extended to other diseases like brain tumour, glaucoma and diabetes detection.

Table 11 Model comparison

	FS-SVM	PSO-SVM	ABC	BH-SVM	Fruit fly-SVM	GBHABC (our model)
Accuracy (%)	96.60	97.81	97.95	95.25	93.62	99.42
Sensitivity (%)	95.27	96.87	98.42	96.85	93.10	98.41
Specificity (%)	97.79	98.87	96.34	92.60	96.53	100.00

Funding The authors declare that no funds, grants or other support were received during the preparation of this manuscript.

Data availability The dataset generated and/or analysed during the current study is WBCD (machine learning repository).

Declarations

Conflict of interest The authors declare that we have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Informed consent This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Abdar M, Zomorodi-Moghadam M, Zhou X, Gururajan R, Tao X, Barua PD, Gururajan R (2020) A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recogn Lett 132:123–131. https://doi.org/10.1016/j.patrec.2018.11.004
2. Afolayan J, Adebiyi M, Arowolo M, Chakraborty A (2022, 06). Breast cancer detection using particle swarm optimization and decision tree machine learning technique. 61-83. https://doi.org/10.1007/978-981-16-8150-94
3. Barzaman K, Karami J, Zarei Z, Hosseinzadeh A, Kazemi MH, Moradi- Kalbolandi S, Farahmand L (2020) Breast cancer: biology, biomarkers, and treatments. Int Immunopharmacol 84:106535. https://doi.org/10.1016/j.intimp.2020.106535
4. Bhardwaj A, Tiwari A (2015) Breast cancer diagnosis using genetically optimized neural network model. Expert Syst Appl 42(10):4611–4620. https://doi.org/10.1016/j.eswa.2015.01.065
5. Boser BE, Guyon IM, Vapnik VN (1992) Training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual acm workshop on computational learning theory, pp 144–152. Publ by ACM. https://doi.org/10.1145/130385.130401
6. Chen HL, Yang B, Liu J, Liu DY (2011) A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst Appl 38(7):9014–9022. https://doi.org/10.1016/j.eswa.2011.01.120
7. Chen HL, Yang B, Wang G, Wang SJ, Liu J, Liu DY (2012) Support vector machine based diagnostic system for breast cancer using swarm intelligence. J Med Syst 36(4):2505–2519. https://doi.org/10.1007/s10916-011-9723-0
8. Chen MJ, Wu WY-Y, Yen AM-F, Fann JC-Y, Chen SL, Chiu SY-H, Chiou ST (2016) Body mass index and breast cancer: analysis of a nation-wide population-based prospective cohort study on 1 393 985 Taiwanese women. Int J Obes 40(3):524–530. https://doi.org/10.1038/ijo.2015.205
9. Cortes C, Vapnik V (1995) Supportvector networks. Mach Learn 20(3):273–297. https://doi.org/10.1007/bf00994018
10. de Boer MC, Wörner EA, Verlaan D, van Leeuwen PA (2017) The mechanisms and effects of physical activity on breast cancer. Clin Breast Cancer 17(4):272–278. https://doi.org/10.1016/j.clbc.2017.01.006
11. Deeb H, Sarangi A, Mishra D, Sarangi SK (2021) Improved Black Hole optimization algorithm for data clustering. J King Saud Univ Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2020.12.013
12. Dheeba J, Albert Singh N, Tamil Selvi S (2014) Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach. J Biomed Inform 49:45–52. https://doi.org/10.1016/j.jbi.2014.01.010
13. Dora L, Agrawal S, Panda R, Abraham A (2017) Optimal breast cancer classification using Gauss-Newton representation based algorithm. Expert Syst Appl 85:134–145. https://doi.org/10.1016/j.eswa.2017.05.035
14. Dr. WIlliam H Wolberg (1992) UCI machine learning repository: breast cancer wisconsin (Original) Data Set
15. Gaitanidis A, Alevizakos M, Tsalikidis C, Tsaroucha A, Simopoulos C, Pitiakoudis M (2018) Refusal of cancer-directed surgery by breast cancer patients: risk factors and survival outcomes. Clin Breast Cancer 18(4):e469–e476. https://doi.org/10.1016/j.clbc.2017.07.010
16. Gerstl B, Sullivan E, Ives A, Saunders C, Wand H, Anazodo A (2018) Pregnancy outcomes after a breast cancer diagnosis: a systematic review and meta analysis. Clin Breast Cancer 18(1):e79–e88. https://doi.org/10.1016/j.clbc.2017.06.016
17. Goff SL, Danforth DN (2021) The role of immune cells in breast tissue and immunotherapy for the treatment of breast cancer. Clin Breast Cancer 21(1):e63–e73. https://doi.org/10.1016/j.clbc.2020.06.011
18. Gu J, Groot G, Boden C, Busch A, Holtslander L, Lim H (2018) Review of factors influencing women's choice of mastectomy versus breast conserving therapy in early stage breast cancer: a systematic review. Clin Breast Cancer 18(4):e539–e554. https://doi.org/10.1016/j.clbc.2017.12.013
19. Hatamlou A (2013) Black hole: a new heuristic optimization approach for data clustering. Inf Sci 222:175–184. https://doi.org/10.1016/j.ins.2012.08.023
20. Hatamlou A (2018) Solving travelling salesman problem using black hole algorithm. Soft Comput 22(24):8167–8175. https://doi.org/10.1007/s00500-017-2760-y
21. Jayachitra S, Prasanth A (2020) Multi-feature analysis for automated brain stroke classification using weighted Gaussian Naive Baye's classifier. J Circuits Syst Comput 30:2150178. https://doi.org/10.1142/S0218126621501784
22. Juanjuan L, Santa-Maria CA, Hongfang F, Lingcheng W, Pengcheng Z, Yuanbing X, Chuang C (2020) Patientreported outcomes of patients with breast cancer during the COVID-19 Outbreak in the epicenter of China: a cross-sectional survey study. Clin Breast Cancer 20(5):e651–e662. https://doi.org/10.1016/j.clbc.2020.06.003
23. Karabatak M (2015) A new classifier for breast cancer detection based on Naïve Bayesian. Measure J Int Measure Confed 72:32–36. https://doi.org/10.1016/j.measurement.2015.04.028
24. Karaboga D, Basturk B (2007) Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics, Vol 4529 LNAI, pp 789–798. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72950-1-7
25. Kessous R, Davidson E, Meirovitz M, Sergienko R, Sheiner E (2017) Prepregnancy obesity: a risk factor for future development of ovarian and breast cancer. Eur J Cancer Prev 26(2):151–155. https://doi.org/10.1097/CEJ.0000000000000228
26. Kiyan T, Yildirim T (2004) Breast cancer diagnosis using statistical neural networks. Istanb Univ J Electr Electron Eng 4(2):1149–1153
27. Koo MM, von Wagner C, Abel GA, McPhail S, Rubin GP, Lyratzopoulos G (2017) Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: evidence from a national audit of cancer diagnosis. Cancer Epidemiol 48:140–146. https://doi.org/10.1016/j.canep.2017.04.010
28. Krawczyk B, Filipczuk P (2014) Cytological image analysis with firefly nuclei detection and hybrid one-class classification

decomposition. Eng Appl Artif Intell 31:126–135. https://doi.org/10.1016/j.engappai.2013.09.017
29. Kumar S, Datta D, Kumar Singh S, Azar AT, Vaidyanathan S (2015) Black hole algorithm and its applications. Stud Comput Intell 575:147–170. https://doi.org/10.1007/978-3-319-11017-2-7
30. Laamiri FZ, Bouayad A, Hasswane N, Ahid S, Mrabet M, Amina B (2015) Risk factors for breast cancer of different age groups: Moroccan data? Open J Obstet Gynecol 05(02):79–87. https://doi.org/10.4236/ojog.2015.52011
31. Lavanya S, Prasanth A, Jayachitra S, Shenbagarajan A (2021) A tuned classification approach for efficient heterogeneous fault diagnosis in IoT-enabled WSN applications. Measurement 183:109771
32. Li DC, Liu CW (2010) A class possibility based kernel to increase classification accuracy for small data sets using support vector machines. Expert Syst Appl 37(4):3104–3110. https://doi.org/10.1016/j.eswa.2009.09.019
33. Li Z, Yu L, Wang X, Yu H, Gao Y, Ren Y, Zhou X (2018) Diagnostic performance of mammographic texture analysis in the differential diagnosis of benign and malignant breast tumors. Clin Breast Cancer 18(4):e621–e627. https://doi.org/10.1016/j.clbc.2017.11.004
34. Littlestone N, Warmuth MK (1994) The weighted majority algorithm. Inf Comput 108(2):212–261. https://doi.org/10.1006/INCO.1994.1009
35. Luo J, Margolis KL, Wactawski-Wende J, Horn K, Messina C, Stefanick ML, Rohan TE (2011) Association of active and passivesmoking with risk of breast cancer among postmenopausal women: a prospective cohort study. BMJ 342(7796):536. https://doi.org/10.1136/bmj.d1016
36. Nahato KB, Harichandran KN, Arputharaj K (2015) Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Comput Math Methods Med. https://doi.org/10.1155/2015/460189
37. Nazarian M, Dezfouli MA, Haronabadi A (2013) Classification of breast cancer samples through using the artificial bee colony algorithm. Int J Comput Appl Technol Res 2(5):522–525. https://doi.org/10.7753/ijcatr0205.1004
38. Nguyen AT, Shiao SL, McArthur HL (2021) Advances in combining radiation and immunotherapy in breast cancer. Clin Breast Cancer 21(2):143–152. https://doi.org/10.1016/j.clbc.2021.03.007
39. Nguyen T, Khosravi A, Creighton D, Nahavandi S (2015) Classification of healthcare data using genetic fuzzy logic system and wavelets. Expert Syst Appl 42(4):2184–2197. https://doi.org/10.1016/j.eswa.2014.10.027
40. Punitha S, Amuthan A, Joseph KS (2019) Enhanced monarchy butterfly optimization technique for effective breast cancer diagnosis. J Med Syst 43(7):1–14. https://doi.org/10.1007/s10916-019-1348-8
41. Sekar J, Aruchamy P, Sulaima Lebbe Abdul H, Mohammed AS, Khamuruddeen S (2022) An efficient clinical support system for heart disease prediction using TANFIS classifier. Comput Intell 38(2):610–640. https://doi.org/10.1111/coin.12487
42. Senapati MR, Mohanty AK, Dash S, Dash PK (2013) Local linear wavelet neural network for breast cancer recognition. Neural Comput Appl 22(1):125–131. https://doi.org/10.1007/s00521-011-0670-y
43. Shahbeig S, Helfroush MS, Rahideh A (2017) A fuzzy multi-objective hybrid TLBO-PSO approach to select the associated genes with breast cancer. Signal Process 131:58–65. https://doi.org/10.1016/j.sigpro.2016.07.035
44. Sheikhpour R, Sarram MA, Sheikhpour R (2016) Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. Appl Soft Comput J 40:113–131. https://doi.org/10.1016/j.asoc.2015.10.005
45. Storn R, Price K (1997) Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 11(4):341–359. https://doi.org/10.1023/A:1008202821328
46. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 71(3):209–249. https://doi.org/10.3322/caac.21660
47. Tao D, Tang X, Wu X (2006) Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans Pattern Anal Mach Intell 28(7):1088–1099. https://doi.org/10.1109/TPAMI.2006.134
48. Thawkar S, Sharma S, Khanna M, Kumar Singh L (2021) Breast cancer prediction using a hybrid method based on butterfly optimization algorithm and ant lion optimizer. Comput Biol Med 139:104968. https://doi.org/10.1016/j.compbiomed.2021.104968
49. Vieira R, Sánchez Tobar JS, Dardes R, Thuler LCS (2018) Alcohol consumption as a risk factor for breast cancer development: a case-control study in Brazil. Asian Pac J Cancer Prev 19(3):703–707. https://doi.org/10.22034/APJCP.2018.19.3.703
50. Waks AG, Winer EP (2019) Breast cancer treatment: a review. JAMA 321(3):288–300. https://doi.org/10.1001/jama.2018.19323
51. Wang H, Zheng B, Yoon SW, Ko HS (2018) A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur J Oper Res 267(2):687–699. https://doi.org/10.1016/j.ejor.2017.12.001
52. Wang M, Chen H (2020) Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. Appl Comput J 88:105946. https://doi.org/10.1016/j.asoc.2019.105946
53. Wang Y, Wang N, Xu M, Yu J, Qin C, Luo X, Ni D (2020) Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound. IEEE Trans Med Imag 39(4):866–876. https://doi.org/10.1109/TMI.2019.2936500
54. Wuniri Q, Huangfu W, Liu Y, Lin X, Liu L, Yu Z (2019) A generic-driven wrapper embedded with feature-type-aware hybrid bayesian classifier for breast cancer classification. IEEE Access 7:119931–119942. https://doi.org/10.1109/ACCESS.2019.2932505
55. Yeh WC, Chang WW, Chung YY (2009) A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. Expert Syst Appl 36(4):8204–8211. https://doi.org/10.1016/j.eswa.2008.10.004
56. Zarbakhsh P, Addeh A, Demirel H (2018) Early detection of breast cancer using optimized ANFIS and features selection. In: Proceedings-9th international conference on computational intelligence and communication networks, cicn 2017, Vol 2018-Janua, pp 39–42. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CICN.2017.8319352
57. Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl 41(1):1476–1482. https://doi.org/10.1016/j.eswa.2013.08.044

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.

Signal, Image and Video Processing
https://doi.org/10.1007/s11760-023-02845-z

ORIGINAL PAPER

Improving speech command recognition through decision-level fusion of deep filtered speech cues

Sunakshi Mehra[1] · **Virender Ranga**[1] · **Ritu Agarwal**[1]

Received: 10 September 2023 / Revised: 29 September 2023 / Accepted: 13 October 2023
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

Abstract
Living beings communicate through speech, which can be analysed to identify words and sentences by recognizing the flow of spoken utterances. However, background noise will always have an impact on the speech recognition process. The detection rate in the presence of background noise is still unsatisfactory, necessitating further research and potential remedies in the speech recognition process. To improve the noisy speech information, this research suggests speech recognition based on a combination of median filtering and adaptive filtering. In this study, speech command recognition is achieved by employing popular noise reduction techniques and utilizing two parallel channels of filtered speech independently. The procedure involves five steps: firstly, enhancing signals using two parallel independent speech enhancement models (median and adaptive filtering); secondly, extracting 2D Mel spectrogram images from the enhanced signals; and thirdly, passing the 2-dimensional Mel spectrogram images to the tiny Swin Transformer for classification. The classification is performed among the large-scale ImageNet dataset, which consists of 14 million images and is approximately 150 GB in size. Fourth, the posterior probabilities extracted from the tiny Swin Transformer modelling are then fed into our proposed 3-layered feed-forward network for classification among our 10-speech command categories. Lastly, decision-level fusion is applied to the two parallel, independent channels obtained from the 3-layered feed-forward network. For experimentation, the Google Speech Command dataset version 2 is used. We obtained a test accuracy of 99.85% when compared with other state-of-the-art methods, demonstrating satisfactory results that can be reported.

Keywords Speech filtering techniques · Swin-tiny transformer · Feed-forward neural network (FNN) · Speech command recognition

1 Introduction

A wide range of human–computer interaction (HCI) systems, such as those used in media, contact centres, sports, repellents, audio monitoring, and various other applications, utilize audio data. However, existing speech emotion recognition (SER) techniques have several shortcomings when it comes to an efficient phone system, including the need for wise feature selection and capable deep learning algorithms.

✉ Sunakshi Mehra
mehra.sunakshi623@gmail.com

Virender Ranga
drvirender.ranga@gmail.com

Ritu Agarwal
ritu.jeea@gmail.com

[1] Department of Information Technology, Delhi Technological University, Delhi, India

To address the ongoing issues in the SER domain effectively, a significant and equitable approach has been developed, which recognizes both temporal and sequential emotional cues [1]. RNNs (recurrent neural networks) have shown success in processing speech in various ways, including voice recognition, speech amplification, speech separation, and speech activity detection. However, training RNNs can be challenging due to the problems of vanishing and exploding gradients, which make it difficult to capture long-term correlations [2]. While collapsing gradients can be managed with straightforward truncation methods, the issue of vanishing gradients requires a more comprehensive approach to be addressed. The term "speech modulation spectrum" has historically been used to refer to the magnitude spectrum of the temporal envelope [3]. Speech recognition performance is affected when these temporal dynamics are distorted. Modern Automatic Speech Recognition (ASR) systems typically generate speech features at a fixed sample rate of 100 Hz. To

Published online: 11 November 2023

incorporate speech dynamics into ASR features, delta and double-delta features are commonly used on the frontend, or raw/smoothed speech temporal patterns (TRAPS) [4] are directly fed into the ASR system over a wide context. Acoustic frontends for ASR often utilize time–frequency masking techniques [5]. In a single-channel scenario, clean speech is estimated by applying a time–frequency mask and then passed to the backend ASR model, either in the complex feature domain [5] or the spectral domain [6]. Alternatively, frontend processing can be performed directly in the time domain [7]. In ASR systems, frontends can cause speech amplification due to alignment discrepancies, further heightened by streaming constraints that impose one-way, causal model limitations.

This paper offers two key contributions: firstly, it introduces a late fusion method for combining posterior scores from transformed filtered speech-to-image data, and secondly, it employs a three-layered feed-forward neural network to classify embeddings generated by the Swin-Tiny Transformer [8]. Experimental evaluations are conducted on the Google Speech Command dataset, comprising 38,546 audio recordings across ten-word categories in .WAV format. Techniques such as median filtering [9] and adaptive filtering [10] are applied to enhance audio quality. In this paper, the use of decision-level fusion with an audio–visual pair has been proven to be effective for speech recognition by incorporating background noise. Section 2 provides a literature review on speech recognition, Sect. 3 presents the proposed dense architecture for merging posterior scores, Sect. 4 outlines the experimental plan and summarizes the findings, and Sect. 5 concludes the paper while suggesting future research directions. The drive to engage in speech recognition research or development arises from the aspiration to enhance communication, broaden accessibility, boost operational efficiency, and foster innovation across diverse industries and applications.

2 Related work

In recent times, there has been a growing interest in the application of neural networks for voice processing [11], noise processing, particularly in the field of voice enhancement. A notable approach for noise filtering is using a DNN with a multilayer architecture based on regression-based speech enhancement [12]. Nonparametric and parametric augmentation algorithms are commonly used, with the former lacking statistical information and the latter not fully exploiting the speech model [13]. Data augmentation techniques, such as speed perturbation [14] and tempo adjustment [15], have demonstrated their usefulness and provide some limited benefits.

ASR transcribes spoken language, commonly used in daily applications, where target speeches, after voice activity detection, are input to early hybrid ASR systems with manually designed and separately optimized acoustic, language, and pronunciation models, incurring high training costs [16]. To assess ASR quality for different groups of speakers, such as gender, age, speakers of sociolects or regional language variants, or those with speech impairments, it is essential to understand demographic data [17]. Voice onset time (VOT) values of stop consonants in a language can also carry indications regarding the speaker's gender, despite the fact that numerous ways and approaches have been suggested to detect a speaker's gender [18]. Acoustic features that are relevant and distinctive play a vital role in distinguishing between speech classes, mitigating the impact of external noise, and effectively handling speaker variability [19].

Thus, ensuring noise resilience has become a critical challenge in speech processing. Previous studies have demonstrated the widespread use of Mel frequency cepstral coefficients (MFCC) as a speech parameterization technique for tasks like speaker identification and speech recognition. These features are designed to emulate the characteristics of the human auditory system and have proven to be effective in achieving high performance in various speech applications. Gammatone frequency cepstral coefficients (GFCC) demonstrate superior performance to MFCC, especially in noisy environments [20, 21]. However, under high SNR conditions, MFCC outperforms GFCC [21]. Power-normalized cepstral coefficients (PNCC) show significantly improved accuracy in noisy conditions compared to MFCC [22]. The effectiveness of PNCC can be attributed to key components such as power-law nonlinearity, asymmetric noise reduction, and temporal masking, which contribute to its enhanced accuracy. For feature extraction in speech recognition using the PNCC algorithm, a bank of Gammatone filters [23] is employed to replicate the nonlinearly increasing bandwidths of human auditory filters. Convolutional neural networks (CNNs) successfully simulate structural localization and reduce translational variance in the feature space [24]. For the Uzbek language and its dialects [25], an end-to-end deep neural network-hidden Markov model (DNN-HMM) speech recognition model and a hybrid connectionist temporal classification (CTC)-attention network are developed. The proposed method effectively utilizes the CTC objective function in attention model training to reduce training time and improve speech recognition accuracy [25]. Feature transformation techniques, such as f-MLLR and VTLN, address variability in DNN-based ASR systems [26, 27]. Speaker-level feature-space maximum likelihood linear regression (f-MLLR) transforms [26], derived from GMM-HMM-based ASR systems [28], are commonly used for this purpose.

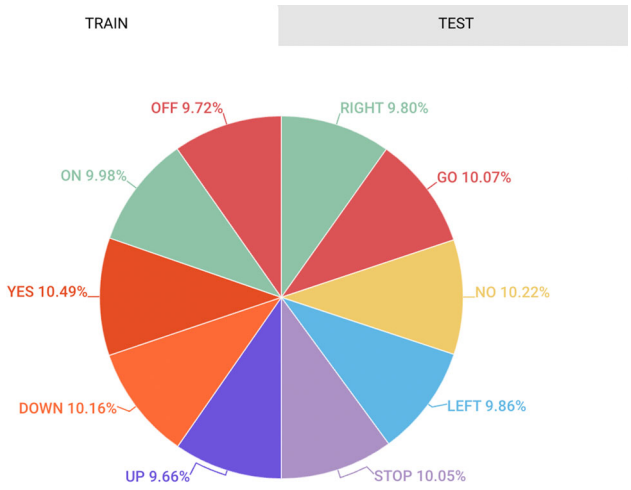


Fig. 1 Train–test measures in percentages for Google Speech Command dataset

3 Proposed methodology

In this section, we explore decision-level fusion [29] by employing the average weighting score to combine the results obtained from two parallel channels. Previous studies [29, 30] have demonstrated that the late fusion of features extracted from different parts of the enhanced input signal leads to excellent categorization. Following the learning phase, many researchers have adopted late fusion to integrate multiple modalities [31] into a unified representation [32, 33]. By concatenating probabilistic scores, late fusion is expected to optimize the performance of the speech recognition system [34]. In Sect. 3.1, we presented details about the dataset employed in our proposed model. In Sect. 3.2, we elaborated on the speech denoising process, while in Sect. 3.3, we described the utilization of the Tiny Swin Transformer for speech classification via Mel spectrograms.

3.1 Dataset

For the purpose of a speech recognition challenge, the TensorFlow and AIY teams provided the Google Speech Command dataset (GSCD) [35], which is used in our studies. Due to the inclusion of varied accents and speakers, the GSCD dataset, which comprises audio files in the .WAV format sampled at 16 kHz, presents a number of difficulties. The experiment consequently gets difficult and applies to real-world situations. Ten target speech command categories—"**YES**", "**NO**", "**UP**", "**DOWN**", "**LEFT**", "**RIGHT**", "**ON**", "**OFF**", "**STOP**", and "**GO**"—are the focus of our experiments. The split for each category in the training and testing sets can be seen in Fig. 1.

3.2 Speech denoising

Our proposed approach incorporates two channels: the median filter and the adaptive filter [9, 10] as shown in Fig. 2. The speech signals processed through these filters are then used to extract two-dimensional Mel spectrograms. The median and adaptive filters adjust their parameters based on the input signal, allowing them to handle different types of noise and signal variations. They selectively remove noise while preserving signal features, effectively suppressing impulsive noise without distorting the underlying signal. This adaptability makes them suitable for varying noise levels. In speech processing, the median filter reduces various types of noise while preserving signal integrity and edge information.

It is computationally efficient, easy to implement, and suitable for real-time applications. For the median and adaptive filtering approaches, the code utilizes the SciPy Python library. After extracting the filtered speech and converting it into 2-dimensional Mel spectrogram images using the librosa library, the subsequent step involves feeding these images into a pre-trained Tiny Swin Transformer network for image classification. Further details on this process will be explained in the next section.

3.3 Pre-trained Swin-Tiny Transformer model

The filtered speech is processed through two channels: a median filter and an adaptive filter. Subsequently, these filtered audio signals are transformed into 2-dimensional Mel spectrogram images using the librosa library. These images undergo further pre-processing, including translation, rotation, and resizing. Following these pre-processing steps, the images are resized to achieve a uniform dimension of 256 $\times$ 256. Once the resizing is complete, these standardized images are input into the Tiny Swin-T version 2 model. The Swin-T V2 model is composed of sequential layers, Swin Transformer blocks, and patch merging, which are iterated. Specifically, within the model's core, there are 6 Swin Transformer blocks. The Swin Transformer [8], pre-trained on vast image datasets like ImageNet (comprising 14 million images and 10,000 classes), stands as a benchmark in visual object recognition research. It includes 1,281,167 training images, 50,000 validation images, and 100,000 test images. Renowned for its exceptional performance in visual tasks, the Swin Transformer surpasses its predecessor, the Vision Transformer (ViT) [36], in terms of efficiency and accuracy. Consequently, Swin Transformers have gained widespread adoption as the backbone of various vision-based model architectures. The Mel spectrograms are transformed into posterior scores through the Swin-Tiny Transformer. The original audio input, after noise removal, is converted into 10-dimensional vectors using a pre-trained tiny Swin-T model

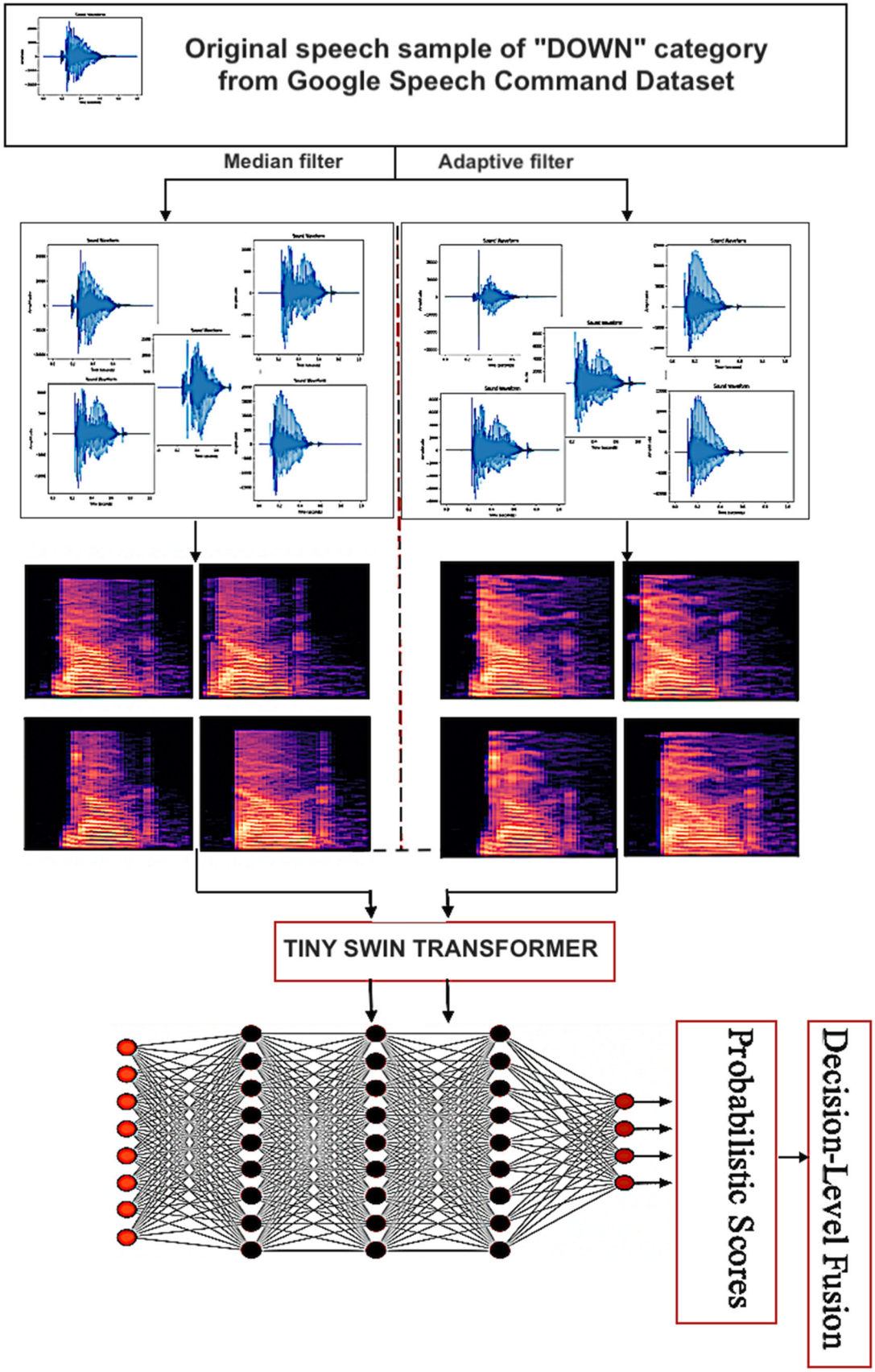


Fig. 2 Workflow of proposed approach

Table 1 Hyperparameters of proposed approach

Parameters	Values
Dataset size	1,21,30,79,830 bytes, 1.25 GB
Total samples	38, 546
Testing samples	4074
Epochs	100
Fully connected layer activation function	ReLU
Number of neural network layers	3

for each word category. This Swin-Tiny Transformer model, pre-trained on ImageNet, is capable of classifying images and generating 10 posterior probabilistic scores for each category. The input features of the Swin-Tiny Transformer model possess a dimension of 768. Our study explores a late fusion technique that combines the posterior probabilistic scores acquired from two distinct channels for speech command recognition, as illustrated in Fig. 2.

4 Experimental results

The experiments are conducted using Google Colab Pro++ GPU and Python. The optimization process employed the Adam optimizer [37] with a learning rate of 0.001. The model is trained for 100 epochs, using a batch size of 32, and the ReLU activation function is applied to introduce nonlinearity. In the speech-to-image process, embeddings are generated and fed into a three-layered feed-forward network described in Table 1. This network consisted of three dense layers with unit sizes of 1024, 256, and 64, respectively, followed by a flattened layer. The ReLU activation function is utilized for these dense layers. For optimization, the widely recognized Adam optimizer [37] is chosen to handle stochastic gradient descent due to its consistent performance across different workloads. The model is trained using a loss function called sparse categorical cross-entropy, which is suitable for multi-class classification problems where the classes are mutually exclusive. After obtaining classification scores from the 3-layered dense model, a decision-level fusion technique is applied, which involves considering the average score value of the obtained probabilistic scores from two independent channels as shown in Fig. 2. Table 1 includes a general description of hyperparameters of proposed methodology.

Table 2 presents a comprehensive comparison of our multimodal fusion approach with state-of-the-art methods. Our technique's superiority is evident in well-resourced data environments, achieving a remarkable test accuracy of 99.85%. In contrast to alternative methods, our approach outperformed a convolutional neural network (CNN) utilizing MFCCs as input by a substantial margin of 6.57% [38]. Furthermore, despite GFCC being acknowledged for their efficacy in emotion perception, our method showcased a superior performance compared to the CNN-Gammatone frequency cepstral coefficients (GFCC) approach [39] by 6.76%. Similarly, compared to using Mel spectrogram with LSTM, our method demonstrated a higher accuracy by 4.41% [40]. In the domain of Urdu acoustic modelling, our approach exhibited a superiority of 4.71% over the LSTM architecture [41].

Table 2 Performance evaluation against the state-of-the-art for GSCD's 10-word categories

Methods	Accuracy (%)
MFCC + CNN [38]	93.28
GFCC + CNN [39]	93.09
MFCC + LSTM-RNN [40]	95.44
MFCC + LSTM-RNN [41]	95.14
MelSpec + LSTM [42]	95.07
DenseNet + BiLSTM [43]	94.88
RNN neural attention [44]	94.11
EdgeCRNN [45]	98.20
Semi supervised audio tagging [46]	95.58
Attention based sequence to sequence model [47]	97.50
Our proposed method	99.85

In addition, when compared to DenseNet-BiLSTM [43], a model designed for keyword spotting, their method achieved an accuracy of 94.88%. In comparison with EdgeCRNN [45], a feature-enhanced method leveraging residual structures and depth-wise separable convolution, our approach exceeded it by 1.65%. Furthermore, we outperformed the Deep CO-Training algorithm (DCT) [46]. Additionally, even against attention-based encoder–decoder models [47] known for their high performance, our method achieved a substantial improvement of 2.35% on the GSCD dataset.

In summary, our proposed method achieved an outstanding accuracy of 99.85% on the Google Speech Command dataset, surpassing the state-of-the-art approaches for categorizing the 10-speech command categories. The test accuracy achieved when using the adaptive filtered channel is 88.75%. However, the test accuracy significantly improved to 99.80% when using the median filtered channel. By averaging the outputs of these two channels through late fusion, the accuracy further improves to 99.85%. The experiment possesses 99.85% accuracy after late fusion using the proposed method. The satisfactory outcome can be attributed to the characteristics of the dataset, which included speaker-dependent samples with various accents. Furthermore, the number of data samples per category is considered optimal compared to other available datasets.

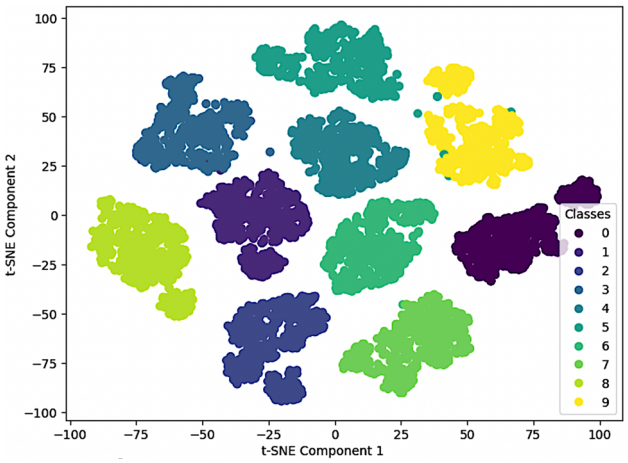


Fig. 3 Multiclass classification evaluation with t-SNE for Speech Command Classification using median filtering technique

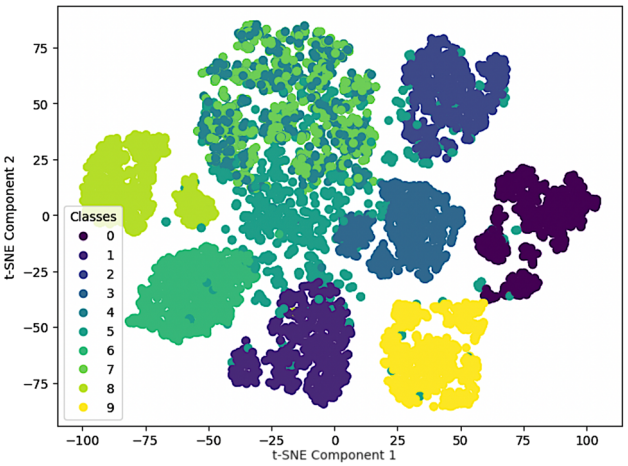


Fig. 4 Multiclass classification evaluation with t-SNE for Speech Command Classification using adaptive filtering technique

Figure 3 illustrates clear differentiation of 10-speech command categories using t-SNE for median filters, while adaptive filters pose visualization challenges. Figures 3 and 4 demonstrate the superior performance of the median filter compared to the adaptive filter. The fusion of both techniques yields the best outcome, as depicted in the confusion matrix in Fig. 5. Combining the strengths of two already successful approaches can result in even greater advancements in classification performance, surpassing the state-of-the-art benchmark.

Table 3 showcases the precision, recall, and F1-scores attributed to each speech command category. Precision, a crucial metric, is defined as the ratio of true positives to the total count of positives identified by the model, as detailed in Eq. (1). This metric gauges the model's capability to accurately recognize positive instances while minimizing the occurrence of false positives.

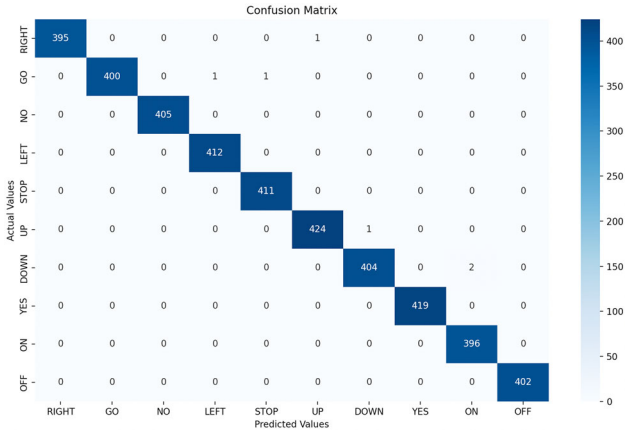


Fig. 5 The confusion matrix is constructed to evaluate the fusion of deep frameworks of adaptive and median filtering approaches

Table 3 Mathematical analysis per speech command categories

Categories	Precision	Recall	F1-score
Right	1.00	0.99	0.99
Go	0.99	1.00	0.99
No	1.00	0.99	1.00
Left	1.00	0.99	1.00
Stop	1.00	0.99	1.00
Up	1.00	0.99	0.99
Down	0.99	1.00	0.99
Yes	1.00	0.99	1.00
On	1.00	0.99	1.00
Off	1.00	0.99	1.00

$$\text{Precision} = \frac{tp}{[tp + fp]} \tag{1}$$

The recall, often referred to as the true positive rate or sensitivity, is a key metric. It is quantified as the proportion of true positives in relation to the total actual positives, as illustrated in Eq. (2). This metric gauges the model's effectiveness in capturing all relevant positive instances, aiming to minimize false negatives.

$$\text{Recall} = \frac{tp}{[tp + fn]} \tag{2}$$

The values for true positives, true negatives, false negatives, and false positives, respectively, are denoted in this context by the symbols tp, tn, fn, and fp. The harmonic mean of precision and recall is used to produce the F1-score, which provides a fair assessment of the model's performance. Equation (3) gives the F1-score's mathematical formulation. This

Table 4 List of misclassified speech command categories

Categories	Number of misclassification/errors
Right	1
Go	2
No	0
Left	0
Stop	0
Up	1
Down	2
Yes	0
On	0
Off	0

metric is crucial because it provides a thorough evaluation of the model's overall accuracy by simultaneously taking precision and recall into account.

$$\text{F1 - score} = \frac{2 \times \text{ precision} \times \text{ recall}}{\left[\text{precision} + \text{recall}\right]} \tag{3}$$

In our experimental setup, we observed notably high values for precision, recall, and F1-score metrics. These elevated values collectively point towards a robust classification performance. Aligning with the defined mathematical formulas, the statistical analysis is presented in Table 3. This table underscores that a substantial portion of the classes have been accurately classified, while only a handful of classes exhibit minor deviations. This underscores the efficacy of our approach in achieving accurate classification results across a majority of categories.

Table 4 elucidates the instances where misclassifications occurred, depicting the words that are erroneously associated with different categories. Notably, specific word categories including "**NO**", "**LEFT**", "**STOP**", "**YES**", "**ON**", and "**OFF**" have demonstrated zero misclassifications, showcasing the robustness of our method. In comparison with other cutting-edge approaches, our method stands out by achieving optimal classification performance, attaining the highest level of accuracy across the spectrum.

Figure 6 unmistakably illustrates the tangible outcomes of our approach, showcasing a test accuracy of 99.80%. However, the integration of adaptive and median filtering-based probabilistic results through fusion propels this accuracy to an even higher threshold, culminating at an impressive 99.85%. This enhancement underscores the efficacy of our approach in achieving superior classification performance through the synergistic utilization of filtering techniques.

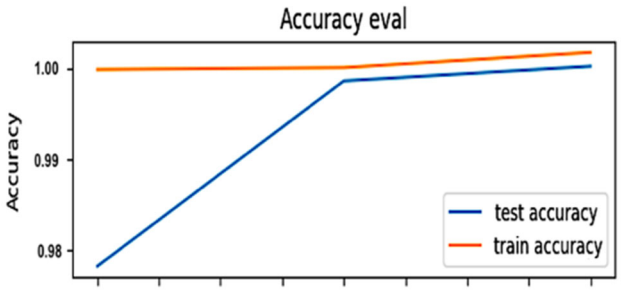


Fig. 6 The train–test accuracy for 10-speech command categories in median filtered speech passed through our proposed neural network

5 Conclusion

This paper presents an innovative approach that enhances speech command recognition by combining audio and image modalities through a late fusion technique. The method incorporates a feed-forward neural network model along with median and adaptive filtering methods to enhance audio signals. The research employs the Google Speech Command dataset, comprising 10-word categories, and attains an impressive test accuracy of 99.85% through the integration of audio and image modalities. The fusion technique involves soft fusion, utilizing posterior class probabilities from two filtered channels extracted from each audio file. For feature extraction, the Swin-Tiny Transformer is employed, followed by a 3-layered feed-forward neural network. In comparison with existing state-of-the-art methods, the proposed fusion technique excels in terms of classification accuracy. This achievement is attributed to its effective capturing and utilization of information from both audio and image modalities, resulting in fine-grained speech command classification. The integration of a pre-trained Swin-Tiny Transformer model, trained on an expansive image dataset, significantly contributes to the notable accuracy achieved in this study. The fusion of audio and visual modalities yields an exceptionally high classification score.

Remarkably, within this context, it is worth noting that median filtering has proven to be a superior pre-processing technique in delivering more favourable results compared to the adaptive filtering method. Furthermore, the fusion of adaptive and median filtering-based probabilistic outcomes significantly enhances accuracy, resulting in an exceptional 99.85% success rate. One limitation of our proposed fusion approach is its increased memory demand, especially when it is applied to extensive datasets. This heightened memory requirement is a result of the substantial computational load imposed by large-scale dataset processing. This limitation is not unique to the proposed approach, but rather is a shared challenge encountered in all other research endeavours conducted on the same dataset.

In the future, we aim to delve deeper into the linguistic aspects of speech command categories, which regrettably are not addressed in this article.

Authors' contributions SM contributed to implementation of proposed approach, wrote the manuscript text, and was involved in writing—original draft and visualization. VR was involved in conceptualization, methodology, writing—review and editing, and supervision. RA contributed to conceptualization, methodology, writing—review and editing, and supervision.

Funding Not applicable.

Data availability The authors verify that the data underpinning the results and discoveries of this research can be found within the confines of the article.

Declarations

Conflict of interest The authors affirm that they do not possess any identifiable financial or personal affiliations that might have influenced the conclusions presented in this study.

Ethical approval The article does not include any studies or investigations that involve human or animal participants.

References

1. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
2. Iqbal, A., Aftab, S.: A classification framework for software defect prediction using multi-filter feature selection technique and MLP. Int. J. Mod. Educ. Comput. Sci. **12**(1), 18 (2020)
3. Hermansky, H.: The modulation spectrum in the automatic recognition of speech. In: 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, pp. 140–147. IEEE (1997)
4. Sadhu, S., Hermansky, H.: Importance of different temporal modulations of speech: a tale of two perspectives. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
5. Narayanan, A., Wang, D.: Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7092–7096. IEEE (2013)
6. Wang, Z.-Q., Wang, P., Wang, DeLiang: Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. IEEE/ACM Trans. Audio, Speech, Lang. Process. **28**, 1778–1787 (2020)
7. Luo, Y., Mesgarani, N.: Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 696–700. IEEE (2018)
8. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022. 2021.
9. Rabiner, L.R., Schafer, R.W.: Introduction to digital speech processing. Found. Trendsin® Signal Process. **1**(1–2), 1–194 (2007)
10. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice hall, Upper Saddle River (2008)
11. Korkmaz, Y., Boyacı, A.: A comprehensive Turkish accent/dialect recognition system using acoustic perceptual formants. Appl. Acoust. **193**, 108761 (2022)
12. Xu, Y., Jun, Du., Dai, L.-R., Lee, C.-H.: A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio, Speech, Lang. Process. **23**(1), 7–19 (2014)
13. Chen, Z.: Noise reduction of bird calls based on a combination of spectral subtraction, Wiener filtering, and Kalman filtering. In: 2022 IEEE 5th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), pp. 512–517. IEEE (2022)
14. Liu, S., Geng, M., Shoukang, Hu., Xie, X., Cui, M., Jianwei, Yu., Liu, X., Meng, H.: Recent progress in the CUHK dysarthric speech recognition system. IEEE/ACM Trans. Audio, Speech, Lang. Process. **29**, 2267–2281 (2021)
15. Xiong, F., Barker, J., Christensen, H.: Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5836–5840. IEEE (2019)
16. Korkmaz, Y., Boyacı, A.: Hybrid voice activity detection system based on LSTM and auditory speech features. Biomed. Signal Process. Control **80**, 104408 (2023)
17. Aksënova, A., Chen, Z., Chiu, C.C., van Esch, D., Golik, G., Han, W., King , L. et al.: Accented speech recognition: benchmarking, pre-training, and diverse data. *arXiv preprint* arXiv:2205.08014 (2022)
18. Korkmaz, Y., Boyacı, A.: Analysis of speaker's gender effects in voice onset time of Turkish stop consonants. In: 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1–5. IEEE (2018)
19. Sadhu, S., Li, R., Hermansky, H.: M-vectors: sub-band based energy modulation features for multi-stream automatic speech recognition. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6545–6549. IEEE (2019)
20. Das, P., Bhattacharjee, U.: Robust speaker verification using GFCC and joint factor analysis. In: Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1–4. IEEE (2014)
21. Shi, X., Yang, H., Zhou, P.: Robust speaker recognition based on improved GFCC. In: 2016 2nd IEEE international conference on computer and communications (ICCC), pp. 1927–1931. IEEE (2016)
22. Kim, C., Stern, R.M.: Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**(7), 1315–1329 (2016)
23. Badi, A., Ko, K., Ko, H.: Bird sounds classification by combining PNCC and robust Mel-log filter bank features. J. Acoust. Soc. Korea **38**(1), 39–46 (2019)
24. Passricha, V., Aggarwal, R.K.: A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. J. Intell. Syst. **29**(1), 1261–1274 (2019)
25. Mukhamadiyev, A., Khujayarov, I., Djuraev, O., Cho, J.: Automatic speech recognition method based on deep learning approaches for Uzbek language. Sensors **22**(10), 3683 (2022)
26. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, pp. 24–29. IEEE (2011)
27. Uebel, L.F., Woodland, P.C.: An investigation into vocal tract length normalisation. In: Sixth European Conference on Speech Communication and Technology (1999)

28. Gales, M.J.F.: Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**(2), 75–98 (1998)
29. Mehra, S., Susan, S.: Improving word recognition in speech transcriptions by decision-level fusion of stemming and two-way phoneme pruning. In: Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10, pp. 256–266. Springer, Singapore (2021)
30. Zhang, Q., Zhang, H., Zhou, K., Zhang, Le.: Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (PMD) for emotion recognition. Tsinghua Sci. Technol. **28**(4), 673–685 (2023)
31. Mehra, S., Susan, S.: Deep fusion framework for speech command recognition using acoustic and linguistic features. Multimed. Tools Appl., 1–25 (2023)
32. Das, R., Singh, T.D.: Image-text multimodal sentiment analysis framework of assamese news articles using late fusion. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **22**, 1–30 (2023)
33. Zhu, J., Huang, C., De Meo, P.: DFMKE: a dual fusion multi-modal knowledge graph embedding framework for entity alignment. Inf. Fusion **90**, 111–119 (2023)
34. Mehra, S., Susan, S.: Early fusion of phone embeddings for recognition of low-resourced accented speech. In: 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), pp. 1–5. IEEE (2022)
35. Warden, P.: Speech commands: a dataset for limited-vocabulary speech recognition. *arXiv preprint* arXiv:1804.03209 (2018)
36. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M. et al.: An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929 (2020)
37. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(7), 2121–2159 (2011)
38. Haque, M.A., Verma, A., Alex, J.S.R., Venkatesan, N.: Experimental evaluation of CNN architecture for speech recognition. In: First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019, pp. 507–514. Springer, Singapore (2020)
39. Abdelmaksoud, E.R., Hassen, A., Hassan, N., Hesham, M.: Convolutional neural network for Arabic speech recognition. Egypt. J. Lang. Eng. **8**(1), 27–38 (2021)
40. McDermott, E., Sak, H., Variani, E.: A density ratio approach to language model fusion in end-to-end automatic speech recognition. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 434–441. IEEE (2019)
41. Zia, T., Zahid, U.: Long short-term memory recurrent neural network architectures for Urdu acoustic modeling. Int. J. Speech Technol. **22**, 21–30 (2019)
42. Lezhenin, I., Bogach, N., Pyshkin, E.: Urban sound classification using long short-term memory neural network. In: 2019 federated conference on computer science and information systems (FedCSIS), pp. 57–60 (2019)
43. Zeng, M., Xiao, N.: Effective combination of DenseNet and BiLSTM for keyword spotting. IEEE Access **7**, 10767–10775 (2019)
44. De Andrade, D.C., Leo, S., Viana, M.L.D.S., Bernkopf, C.: A neural attention model for speech command recognition. *arXiv preprint* arXiv:1808.08929 (2018)
45. Wei, Y., Gong, Z., Yang, S., Ye, K., Wen, Y.: EdgeCRNN: an edge-computing oriented model of acoustic feature enhancement for keyword spotting. J. Ambient Intell. Humaniz. Comput., 1–11 (2022)
46. Cances, L., Pellegrini, T.: Comparison of deep co-training and mean-teacher approaches for semi-supervised audio tagging. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 361–365. IEEE (2021)
47. Higy, B., Bell, P.: Few-shot learning with attention-based sequence-to-sequence models. *arXiv preprint* arXiv:1811.03519 (2018).

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.

arXiv:2310.20704v1 [cs.CV] 31 Oct 2023

Limited Data, Unlimited Potential: A Study on ViTs Augmented by Masked Autoencoders

Srijan Das[1], Tanmay Jain[2], Dominick Reilly[1], Pranav Balaji[3], Soumyajit Karmakar[4], Shyam Marjit[4], Xiang Li[5], Abhijit Das[3], and Michael Ryoo[5]
[1]UNC Charlotte, [2] Delhi Technological University, [3] BITS Pilani Hyderabad
[4] Indian Institute of Information Technology Guwahati, [5] Stony Brook University
sdas24@charlotte.edu

Abstract

Vision Transformers (ViTs) have become ubiquitous in computer vision. Despite their success, ViTs lack inductive biases, which can make it difficult to train them with limited data. To address this challenge, prior studies suggest training ViTs with self-supervised learning (SSL) and fine-tuning sequentially. However, we observe that jointly optimizing ViTs for the primary task and a Self-Supervised Auxiliary Task (SSAT) is surprisingly beneficial when the amount of training data is limited. We explore the appropriate SSL tasks that can be optimized alongside the primary task, the training schemes for these tasks, and the data scale at which they can be most effective. Our findings reveal that SSAT is a powerful technique that enables ViTs to leverage the unique characteristics of both the self-supervised and primary tasks, achieving better performance than typical ViTs pre-training with SSL and fine-tuning sequentially. Our experiments, conducted on 10 datasets, demonstrate that SSAT significantly improves ViT performance while reducing carbon footprint. We also confirm the effectiveness of SSAT in the video domain for deepfake detection, showcasing its generalizability. Our code is available at https://github.com/dominickrei/Limited-data-vits.

1. Introduction

Vision Transformers (ViTs) have become a common sight in computer vision owing to their success across various visual tasks, and are now considered a viable alternative to Convolutional Neural Networks (CNNs). Despite this, ViTs are structurally deficient in inductive bias compared to CNNs, which necessitates training them with large-scale datasets to achieve acceptable visual representation, as noted by Dosovitskiy et al. [13]. As a result, when dealing with small-scale datasets, it is essential to utilize a ViT pre-trained on a large-scale dataset such as ImageNet [12] or JFT-300M [50]. However, in domains such as medical datasets, pre-training ViTs on ImageNet or JFT-300M may not result in an optimal model for fine-tuning on those datasets due to a significant domain gap. Thus, the aim of this research is to address the following question: *how can ViTs be trained effectively in domains with limited data*?

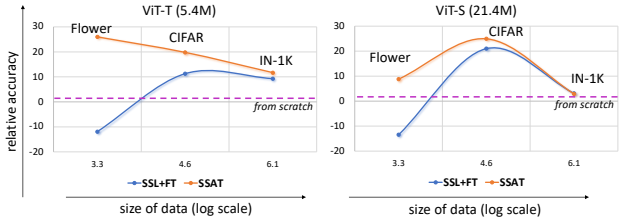


Figure 1. Relative classification accuracy on three datasets with different sizes: (i) Oxford Flower [37] (2K samples), (ii) CIFAR [25] (50K samples), and (iii) ImageNet-1K [12] (IN-1K, 1.2M samples). SSAT consistently outperforms others on all three datasets with two backbones. On the other hand, given the same SSL method, SSL+FT achieves a compromised performance than SSAT, especially on the tiny Oxford Flower dataset (even worse than training from scratch).

Following the introduction of ViTs, second-generation vision transformers have emerged with two different approaches. The first approach is to use a hierarchical structure to introduce inductive bias in ViTs [18, 34, 54]. The second approach involves using hybrid architectures, such as introducing convolutional blocks within ViTs [39, 57]. However, both approaches primarily benefit medium-sized datasets and not small-scale datasets. Several efforts have been made to enhance their locality inductive bias, as reported in literature [15, 28, 29, 33]. Among these methods, SSL has demonstrated exceptional efficacy in training transformers from scratch on small datasets [5, 6, 15, 24, 33, 49, 56]. These methods typically involve sequentially conducting SSL and fine-tuning on the same small dataset to enhance ViT performance.

Meanwhile, another straightforward approach that takes

advantage of SSL is to jointly optimize the self-supervised task along with the primary task like classification or segmentation. We name such SSL tasks as **S**elf-**S**upervised **A**uxiliary **T**ask (**SSAT**). Although SSAT has been explored in the vision community [29, 33, 42] and robotics community [27, 30], there are still many open questions, especially when the size of the dataset is limited.

This paper empirically analyzes the aforementioned joint learning approach with SSAT, as an alternative to sequentially performing SSL and fine-tuning (SSL+FT) on the same dataset. Through an extensive amount of experiments on *ten* image classification datasets of various sizes as well as *two* video classification datasets, surprisingly, we observe that SSAT works significantly better than other baselines like SSL+FT and training from scratch, especially for ViT on small datasets (see Figure 1). Further experiments empirically show that it is most effective when the auxiliary task is image reconstruction from missing pixels among the well known SSL methods we tested. Finally, we perform a detailed model and feature analysis to highlight the unique properties of SSAT-driven models in comparison to other representative baselines. This distinction is particularly notable when comparing with the SSL+FT models which are trained with similar loss functions. We reveal that the advantages of SSAT in a limited-data regime come from better semantic richness, a distinct attention distribution, and an increased capability for feature transformation, which results in higher feature variance.

2. Related Work

Vision Transformers. Several vision transformers [1, 2, 8, 13, 35, 43, 45, 54, 58, 62, 64] have been introduced in recent times for a wide range of tasks. However, these models require large-scale pre-training to be effective on different datasets. In an effort to reduce their reliance on extensive training, DeiT [52] introduced extensive data augmentation, regularization, and distillation tokens from convolutions in ViTs. T2T [61], in a similar vein, employed a tokenization technique that flattened overlapping patches and applied a transformer to allow for learning local structural information around a token. Meanwhile, some ViT models [10, 23, 57] have introduced inductive bias into the transformers through the use of convolutional filters. Hierarchical transformers [14, 32, 34, 55] have introduced inductive bias by reducing the number of tokens through patch merging and thus operating at different scales. However, these architectures do not overcome the limitation of ViTs, which require at least a medium-sized dataset for pre-training [39].

Self-supervised Learning. Self-Supervised Learning (SSL) aims to learn visual representations through pretext tasks. Contrastive methods, such as SimCLR [7] and MoCo [21], minimize the distance between differently augmented views of the same image (positive pairs) while maximizing it for dissimilar images (negative pairs). On the other hand, non-contrastive methods like BYOL [17] and DINO [4] only impose minimization between the positive pairs. In contrast, reconstruction based methods [16, 20, 51, 59] have shown to be effective self-supervised learners for various downstream computer vision tasks. In these methods, an encoder operates on a small portion of an image to learn a latent representation, and a decoder decodes the latent representation to reconstruct the original image in the pixel space. These SSL methods are commonly used for large-scale pre-training of ViTs to enhance their effectiveness in various downstream tasks.

ViTs for small datasets. Liu et al. [33] proposed an auxiliary self-supervised task that improves the robustness of ViT training on smaller datasets. The task involves predicting relative distances among tokens and is jointly trained with primary tasks. On the other hand, Li et al. [29] conducted distillation in the hidden layers of ViT from a lightweight CNN-trained model. To address the lack of locality inductive bias, Lee et al. [28] introduced a ViT architecture with shifted patch tokenization and locality self-attention. Gani et al. [15] proposed an SSL+Fine-tuning methodology where the SSL is similar to the pretext task in DINO [4]. These methods eliminate the need for large-scale pre-training and allow ViTs to learn meaningful representations with limited data. In contrast to these methods, we propose SSAT akin to [33], but with an approach that combines the functionality of self-attention and MLPs through image reconstruction.

3. Preliminaries

ViT utilizes a non-overlapping grid of image patches to process a given image X, where each patch is linearly projected into a set of input tokens. ViT consists of a stack of multi-head attention and linear layers as in [13]. The transformer attention layers model the pairwise relationship between the input tokens [53]. For generalizability, We denote the transformer encoder as f. For brevity, we have omitted the parameters of the encoder. In practice, f operates on an augmented version of the input image X to output a discriminative representation $f(A(X))$ where A is the set of image augmentation. This representation is subsequently classified into class labels using a classifier h. A class-wise cross-entropy loss L_{cls} is used to train the transformer encoder.

4. Self-supervised Auxiliary Task (SSAT)

Our objective is to improve the ViT training on the dataset with limited samples. Consequently, we propose to jointly train the primary classification task of ViT alongside a self-supervised auxiliary task (SSAT). The joint optimization of the SSAT and classification task allows the ViT to capture inductive biases from the data without requiring any

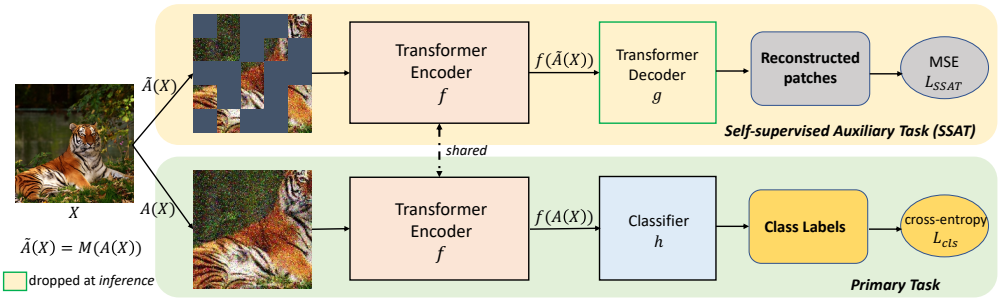


Figure 2. An overview of **ViT training with SSAT**. The input X to the ViT undergoes data augmentations, $A(X)$ and $\tilde{A}(X) = M(A(X))$, using a mask operation M. These augmented inputs are then fed to a Transformer Encoder f, resulting in two latent representations: $f(A(X))$ and $f(\tilde{A}(X))$. These correspond to the masked and full image, respectively. The latent representation of the full image is utilized for the image classification task, while the masked image's representation is used for the image reconstruction task. ViT training involves joint optimization of losses from both these tasks.

additional labels. An overview of our framework is depicted in Figure 2.

In our joint optimization framework for SSAT, we have utilized the widely adopted Masked Autoencoder (MAE) approach [19] for reconstructing the missing pixels. Nonetheless, it is worth noting that any SSL method can be integrated into our framework, given its generic nature. Our decision to use MAE was based on its superior performance, as evidenced by our experimental analysis (Table 4).

To the existing ViT frameworks, where the Transformer encoder f and Classifier h process the full image patches $A(X)$ to compute the classification loss L_{cls}. we introduce an augmentation set $\tilde{A} = M(A(X))$, where operation M randomly masks out patches in the input image X. The transformer encoder f also operates on the unmasked tokens, generating latent representation $f(\tilde{A}(x))$ for these tokens. In parallel to the classifier h , SSAT employs a shallow decoder g to reconstruct back the unseen image pixels from the latent representation of the seen tokens $f(\tilde{A}(x))$. Following [19], the decoder takes as input the latent representation of the seen tokens $f(\tilde{A}(x))$ and a learnable masked token. Each token representation at the decoder's output is linearly projected to a vector of pixel values representing a patch. The output $g(f(\tilde{A}(X)))$ is reshaped to form the reconstructed image, thereafter computing the normalized Mean Square Error (MSE) loss L_{SSAT} between the original and reconstructed image. In practice, the MSE is computed only for the masked patches as in [20].

Thus, the entire framework performs a primary task, i.e. *classification* and a self-supervised auxiliary task, i.e. *reconstruction*. This framework can be jointly optimized using a convex combination of the losses from the primary task and SSAT. Thus, the total loss is computed by

$$L = \lambda * L_{cls} + (1 - \lambda) * L_{SSAT} \quad (1)$$

λ is the loss scaling factor. During inference, the decoder is discarded and the encoder f processes all input patches to generate the classification output only. Our framework supports training of any ViT model and SSAT variants.

5. Experimental Analysis

In this section, we present the superiority of using SSAT while training any vision transformer. Our experiments are based on image and video classification tasks. We use 12 different datasets: (i) 4 small sized datasets: CIFAR-10 [25], CIFAR-100 [25], Oxford Flowers102 [37] (Flowers) and SVHN [36], (ii) 1 medium sized dataset: ImageNet-1K [12] (IN-1K), (iii) 2 medical datasets: Chaoyang [63] and PMNIST [60], (iv) 3 datasets of DomainNet [41]: ClipArt, Infograph, and Sketch, and (v) 2 video datasets for deepfake detection: DFDC [46] and FaceForensics++ [44].

Our experiments for image reconstruction in the context of SSAT generally follow the procedure outlined in [20], unless otherwise stated. In particular, we employ the decoder design from [20] for ViT and utilize the decoder design from ConvMAE [16] and SimMIM [59] for hierarchical encoders such as CVT and swin, respectively. To optimize hyper-parameters for the decoder, we conduct our experiments with the ViT encoder. For augmentation $\tilde{A}$, we use a random masking with 75% masking ratio. Our decoder has a depth of 2 (i.e. 2 transformer layers) and an embedding dimension of 128. We provide ablations on the choice of these hyper-parameters in Appendix C. It is worth noting that our decoder is shallower than that in MAE [20]. The loss scaling factor λ is set to 0.1 for all the datasets.

Our ViT encoders (f) are trained using the training recipe of DeiT [52], unless otherwise specified. The configuration of ViT-T, ViT-S, and ViT-B is identical to the configuration described in [52]. We borrow the network architecture for CVT-13, and Swin from the official code of [57], and [34], respectively. Training is conducted for 100 epochs, unless otherwise specified, using 8 A5000 24GB GPUs for IN-1K and one A5000 24 GB GPU for all other

Table 1. Top-1 classification accuracy (%) of different ViT variants with and without SSAT on CIFAR-10, CIFAR-100, Flowers102, and SVHN datasets. All models were trained for 100 epochs.

Method	# params. (M)	CIFAR-10	CIFAR-100	Flowers102	SVHN
ViT-T [52]	5.4	79.47	55.11	45.41	92.04
+SSAT	5.8	**91.65** (+12.18)	**69.64** (+14.53)	**57.2** (+11.79)	**97.52** (+5.48)
ViT-S [52]	21.4	79.93	54.08	56.17	94.45
+SSAT	21.8	**94.05** (+14.12)	**73.37** (+19.29)	**61.15** (+4.98)	**97.87** (+3.42)
CVT-13 [57]	20	89.02	73.50	54.29	91.47
+SSAT	20.3	**95.93** (+6.91)	**75.16** (+1.66)	**68.82** (+14.53)	**97** (+5.53)
Swin-T [34]	29	59.47	53.28	34.51	71.60
+SSAT	29.3	**83.12** (+23.65)	**60.68** (+7.4)	**54.72** (+20.21)	**85.83** (+14.23)
ResNet-50 [22]	25.6	91.78	72.80	46.92	96.45

datasets. Additional training details for each dataset can be found in Appendix B.

5.1. Main Results

SSAT on small-sized dataset: In Table 1, we present the classification accuracy on the small-sized datasets with different variants of vision transformers: ViT-T, ViT-S, CVT-13, and Swin-T. In this table, we demonstrate the impact of using SSAT while training the transformers for learning the class labels. All the models have been trained for 100 epochs from scratch. Although the models with SSAT have more training parameters, they have identical operations during inference. SSAT improves the classification accuracy on all the datasets for all the transformer encoders. It is worth noting that ViT-T with 5.4M parameters when trained with SSAT outperforms ViT-S with 21.4M parameters. The highest classification accuracy is achieved with CVT-13 (20M parameters) due to the introduction of convolutions that infuse inductive bias into the transformers. Although convolutions are generally more effective than transformers on small datasets, our experiments demonstrate that the most effective convolutional network (ResNet-50 [22]) for these datasets underperforms most of the transformers when trained with SSAT, except for ViT-T on CIFAR-10 and CIFAR-100 datasets.

SSAT on medium-sized dataset: In Table 2, we present the impact of SSAT on ViTs that were trained on a medium-sized dataset, such as IN-1K [12]. Our results demonstrate that SSAT consistently enhances the classification accuracy of ViTs, even as the number of training samples increases. Notably, this improvement is more pronounced for smaller models, which have 5.4M parameters, than for larger ones. Specifically, we observed a relative performance improvement of 11.8% for ViT-T with SSAT, as compared to only 2.9% for ViT-S+SSAT. These findings suggest that SSAT can be effectively utilized to train lighter transformers that can be deployed on edge devices.

Does SSAT promotes overfitting? In Table 2, we also analyse the robustness of ViTs to natural corruptions. Given that we recommend the use of SSAT to enhance representation learning in transformer training, it is reasonable to question whether this approach can lead to overfitting on small training samples. To address this concern, we evaluate the performance of our trained models on perturbed versions of the data, specifically, CIFAR-100-p and IN-1K-p, which are obtained by applying random perspective transformations to images following [48]. Our results demonstrate that ViTs trained with SSAT exhibit greater robustness to these natural corruptions compared to the baseline ViTs. We observe notable improvements in performance for tiny ViTs, as evidenced by the results for ViT-T+SSAT in Table 2, as well as for smaller datasets such as CIFAR-100-p.

Comparison of SSAT with SSL+FT: In Table 3, we present the superiority of joint training of the SSL loss with the classification loss over the two-step sequential training approach, where the model is first trained with SSL and then fine-tuned (FT) for classification. Our empirical analysis is conducted on ViT-T, where we compare the performance of ViT trained from scratch and ViT + SSAT, which are trained for 100 epochs. To establish baselines for our SSL+FT model, we conducted experiments using four different training protocols: (1) 50 epochs of SSL training followed by 50 epochs of fine-tuning, (2) 50 epochs of SSL training followed by 100 epochs of fine-tuning, (2) 100 epochs of SSL training followed by 50 epochs of fine-tuning, and (4) 100 epochs of SSL training followed by 100 epochs of fine-tuning. Additionally, we quantify the carbon emission of the models trained using different methods with the help of a tool provided by [26]. Note that the GFLOPs, training time, and Kg CO_2 eq. are specified for the model trained on IN-1K for better generalizability. Our empirical results show that all models incorporating SSL outperform those trained from scratch, highlighting the importance of self-supervised learning when training transformers on small datasets. Moreover, even when requiring an additional 4 hours of training time and resulting in approximately 0.6 Kg CO_2 equivalent of additional carbon emissions, our SSAT models demonstrate superior performance compared to the SSL+FT model (50 epoch

Table 2. Top-1 classification accuracy (%) on ImageNet-1K (IN-1K), perturbed CIFAR-100 (CIFAR-100-*p*), and perturbed ImageNet-1K (IN1K-*p*)

Method	IN-1K	CIFAR-100-*p*	IN-1K-*p*
ViT-T	65.0	25.1	48.3
+SSAT	**72.7**	**37.6**	**59.6**
ViT-S	74.2	22.5	62.7
+SSAT	**76.4**	**43.9**	**64.5**

Table 3. Top-1 accuracy and efficiency of ViT-T trained from scratch, with SSL+FT, and with SSAT. We provide the GFLOPs, training time (GPU hours), and CO_2 emissions (kg eq) for IN-1K.

Method	GFLOPs	CIFAR-10	CIFAR-100	IN-1K	Train time	Kg CO_2 eq.
Scratch	1.26	79.47	55.11	65.0	60	5.96
(1) SSL+FT	0.43+1.26	85.33	60.43	70.09	55	5.46
(2) SSL+FT	0.43+1.26	86.48	63.28	71.1	82	8.15
(3) SSL+FT	0.43+1.26	85.3	60.3	70.5	74	7.35
(4) SSL+FT	0.43+1.26	88.72	67.53	**74.07**	104	10.33
Ours	1.67	**91.65**	**69.64**	72.69	78	7.55

Table 4. Top-1 accuracy of existing SSL strategies used as SSAT. MAE as the SSAT achieves the best result on both CIFAR-10 and CIFAR-100.

SSAT (SSL)	CIFAR-10	CIFAR-100
SimCLR [7]	55.21	36.49
DINO [4]	80.07	60.6
MAE [20]	**91.65**	**69.64**

SSL + 100 epoch FT). Although accuracy improves when SSL+FT models are trained on CIFAR-10 and CIFAR-100 for 104 GPU hours, our SSAT approach remains superior, requiring 26 GPU hours less training time and burning approximately 2.8 Kg CO_2 equivalent. However, the SSL+FT model outperforms SSAT when a large amount of training data is available.

Appropriate SSL for joint training: Table 4 presents a comparison of the performance of the SSAT approach, implemented with different SSL strategies, namely, contrastive (SimCLR [7]), non-contrastive (DINO [4]), and reconstruction based (MAE [20]), on the ViT model. Our analysis reveals that the use of SimCLR results in a decrease in the ViT's performance, which can be attributed to the conflicting losses that arise while optimizing the cross-entropy loss to learn class labels and the contrastive loss. However, DINO and MAE both enhance the ViT's performance when jointly trained with cross-entropy. Notably, the improvement observed with MAE is more significant than that with DINO. The superior performance of MAE can be attributed to the centering and sharpening technique employed in DINO, which impedes the learning of class labels while only facilitating the SSL. On the other hand, as mentioned in [40], MAE encourages MLPs in ViTs to be more representative. While the cross-entropy loss primarily contributes more to the self-attention blocks. Thus, SSAT implemented with reconstruction based SSL harmonizes the impact of both tasks, thus improving the ViT's learning capabilities.

Superiority of SSAT over Large-scale pre-training: In situations where training samples are limited and the data distribution differs from that of natural images, large-scale pretraining can be challenging. The main obstacle is the lack of data that accurately represents the downstream data distribution. Consequently, we conducted experiments using ViTs on medical and domain adaptation datasets (Tables 5 and 6) where data is scarce. In Table 5, we demonstrate how SSAT significantly enhances the classification performance of ViT-T on the Chaoyang and PMNIST datasets. The resulting model not only surpasses a comparable ViT model that was pre-trained on ImageNet [12], but also outperforms its larger ViT-S model when trained without SSAT. We observed similar trends of improvement on three datasets from DomainNet [41] in Table 6. It is worth mentioning that our CVT model, when trained using SSAT, outperforms $\mathcal{L}_{drloc}$ [33], which is another state-of-the-art self-supervised loss designed to enhance transformer performance on small datasets.

Table 5. Top-1 accuracy on medical image datasets. All models are trained for 100 epochs.

	Method	Chaoyang	PMNIST
ViT-T	Scratch	77.37	90.22
	IN-1K pretrained + FT	78.78	91.99
	Scratch + **SSAT**	**82.52**	**93.11**
ViT-S	Scratch	80.04	91.19
	IN-1K pretrained + FT	80.18	92.63
	Scratch + **SSAT**	**81.25**	**93.27**

Table 6. Top-1 accuracy on DomainNet datasets. All models are trained for 100 epochs

Method	ClipArt	Infograph	Sketch
ViT-T	29.66	11.77	18.95
+SSAT	47.95	16.37	46.22
CVT-13	60.34	19.39	56.98
+$\mathcal{L}_{drloc}$ [33]	60.64	20.05	57.56
+SSAT	**60.66**	**21.27**	**57.71**

Loss scaling factor: In Figure 3 we perform an empirical analysis to determine the optimal value for the loss scaling factor λ. Our analysis focused on CIFAR datasets show that the choice of $\lambda = 0.1$ is an optimal choice when SSAT positively impacts the primary classification task.

Extended training: In this experiment, we extend the training schedules of both the scratch and SSAT model as illustrated in Figure 4. Our findings indicate that the performance enhancement of our SSAT model, relative to the ViT

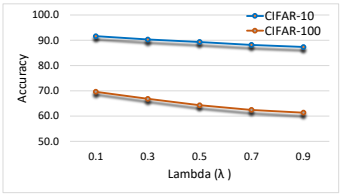


Figure 3. Ablation for loss scaling factor λ.

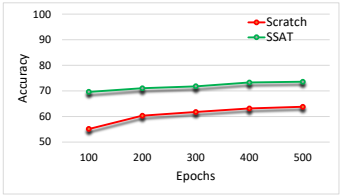


Figure 4. ViT (scratch) vs. ViT+SSAT for longer epochs on CIFAR-100.

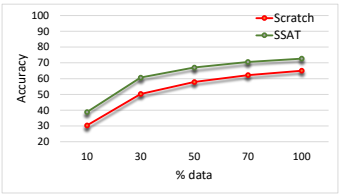


Figure 5. ViT (scratch) vs. ViT+ SSAT for different subset of IN-1K.

baseline, remains consistent throughout the entire training period. These results suggest that the improvement in the SSAT model's performance is not due to a faster convergence rate, but rather to superior optimization capabilities.

Training for different subsets of IN-1K: Figure 5 presents our analysis of the performance of the ViT baseline and SSAT model for varying training sample sizes, specifically on subsets of IN-1K. Our results demonstrate that the performance enhancement of the SSAT model, relative to the baseline model, is consistent across all subsets (i.e., different sizes of the training data). These findings substantiate that models with low training parameters, such as ViT-T, can benefit from SSAT at all scales of training data.

5.2. Diagnosis of features learned by SSAT

In this section, we differentiate the properties of ViTs learned from scratch, SSL+FT, and our SSAT method. We investigate the learned ViT properties by analyzing their attention weights, token representation, feature transformation, and loss landscape. We answer the following key questions:

How are the attention weights distributed? The objective of this experiment is to examine the mean attention weights received from other tokens in a sample in the data distribution. As outlined in [53], the sum of all values in a column of an $n \times n$ self-attention matrix, where n denotes the number of tokens, represents the aggregated attention associated with a token. Figure 6 displays the attention weight distribution across the n tokens for various ViT blocks on both Flower (top row) and CIFAR-100 (bottom row) datasets. The attention weights are uniformly distributed in the first transformer block of the scratch model on both datasets, implying an equal focus on all image regions. However, this distribution changes slightly in the deeper layers. Intriguingly, SSL+FT and SSAT models display sharply peaked attention distributions in the initial and middle transformer layers, but the distributions do not necessarily align with each other. Specifically, in the first transformer block, the attention weight distributions of SSL+FT and SSAT models complement each other, indicating that lower-level features learned by these models are complementary. Moreover, the SSL+FT models exhibit sharp peaks in the final layers, whereas the peaks in SSAT models have a lower magnitude, possibly because the latter model has a better inductive bias. Therefore, although both models are trained on the same set of losses, they use different mechanisms to learn attention weights that differ in the initial layers, and the attention weights learned by the SSAT model are smoother in the final layers, indicating a better inductive bias of the model.

What is the quality of the learned tokens? In this study, we investigated the average distance between tokens within a sample across different transformer blocks. Our analysis involves plotting the average Euclidean distance between tokens in images from the Flower and CIFAR-100 datasets at the output of the transformer layers, as shown in Figure 7. Our results indicate that the scratch model yields a lower inter-token distance than the other models, implying homogeneous token representation. We also observe that SSL+FT models yield higher inter-token distances than SSAT models at the middle transformer layers, but this distance diminishes as we go deeper into the ViTs. Consequently, the SSL+FT models suffer from homogeneous token representation, which adversely affects the ViT training, leading to sub-optimal classification accuracy. In contrast, the inter-token distance of SSAT models increases with ViT depth, indicating that the token representations are discriminative and are semantically rich.

How are representations transformed? The aim of our experiment is to showcase the variation in feature map evolution between ViTs that are trained using different mechanisms. We conducted feature variance measurements across the ViT layers on Flower and CIFAR-100 datasets, and the results are presented in Figure 8. Our analysis confirms the findings of previous studies that the feature variance across ViTs trained from scratch remains constant. However, we observed that the SSL+FT models exhibit an increase in feature variance until a certain layer, after which the rate of an increase either decreases (in Flower dataset) or begins to fall (in CIFAR-100 dataset). Conversely, the feature variance in our SSAT models accumulates with each ViT layer and tends to increase as the depth increases. Consequently, as we go deeper in the SSAT models, the feature map uncertainty decreases, which facilitates optimization through ensembling and stabilizing the transformed feature maps [38].

Why is SSAT better than SSL+FT? In this study, we investigated the loss landscapes of ViT models trained using different training mechanisms. We follow [39] to display the Eigenvalue Spectral Density of Hessian for the different ViT models trained (see Figure 9). Our results indicate

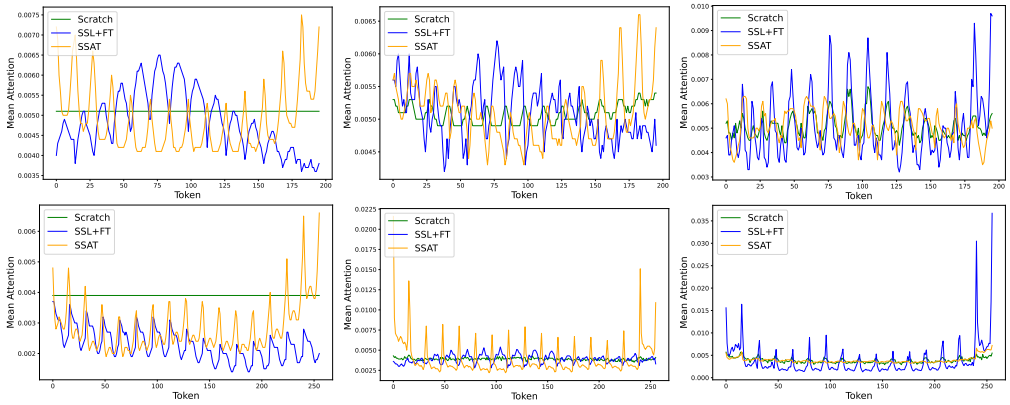


Figure 6. The **distribution of attention weights** across the n tokens for different ViT-T blocks on two datasets: Oxford Flower (top row) and CIFAR-100 (bottom row). The first, second, and third columns correspond to the attention distributions of the first, sixth, and twelfth ViT-T blocks, respectively.

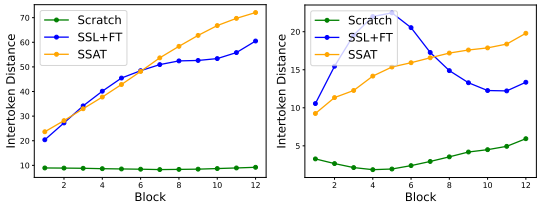


Figure 7. **Average Euclidean Inter-token Distance** of ViTs trained from scratch, using SSL+FT and using SSAT, for two different datasets: Oxford Flowers (on the left) and CIFAR-100 (on the right).

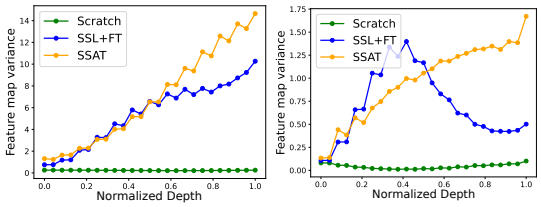


Figure 8. **Feature Map Variance** of ViTs trained from scratch, using SSL+FT and using SSAT, for two different datasets: Oxford Flowers (on the left) and CIFAR-100 (on the right).

that the scratch ViT model exhibits a wide range of negative Hessian eigenvalues, implying non-convex loss landscapes. Interestingly, the number of negative Hessian eigenvalues is slightly higher in the SSL+FT ViT model than in the scratch model (9622 vs 9667). However, the lower magnitude of some of the negative Hessian eigenvalues in the SSL+FT model makes their qualitative visualization difficult. In contrast, SSAT reduces the number of negative Hessian eigenvalues by 12% in comparison to the SSL+FT model. This finding suggests that the SSL approach convexifies losses and suppresses negative eigenvalues in the small data regime. Additionally, the SSAT ViT model reduces the average magnitude of negative Hessian eigenvalues by 70% compared to the SSL+FT models. Therefore, SSAT effectively reduces the magnitude of large Hessian eigenvalues and enhances the ViTs' ability to learn better representations.

Table 7. Comparison of our SSAT to existing state-of-the-art approaches on small datasets. † indicates that [15] is replicated with 300 epochs. Results of [29] is not reported on CIFAR-10.

Method	# enc. params.	epochs	CIFAR-10	CIFAR-100
CVT-13+$\mathcal{L}_{drloc}$ [33]	20M	100	90.30	74.51
CVT-13+ SSAT			**95.93**	**75.16**
ViT (scratch)	2.8M	300	93.58	73.81
SL-ViT [28]			94.53	76.92
ViT† (SSL+FT) [15]			94.2	76.08
ViT + SSAT			**95.1**	**77.8**
DeiT-Ti+$\mathcal{L}_{guidance}$ [29]	6M	300	-	78.15
DeiT-Ti+$\mathcal{L}_{guidance}$ + SSAT			-	**79.46**

5.3. Comparison with the state-of-the-art

Table 7 presents a comparison of SSAT with state-of-the-art (SOTA) methods. To ensure a fair evaluation, we implemented SSAT with the ViT encoder used in the respective methods. We find that MAE as SSAT outperforms Drloc [33] which takes predicting relative distance between patches as SSAT. This shows that the choice of SSAT plays a crucial role in the effective training of ViTs. Moreover, we find that SSAT outperforms SL-ViT [28] and [15] when trained for an equal number of epochs. This indicates that SSAT, without any architectural modifications, can surpass SOTA methods through its joint training strategy. Additionally, we trained a ViT with SSAT and feature-level distillation from a light-weight CNN as described in [29]. The improvement over the baseline [29], which involves training ViT with feature-level distillation only, demonstrates the complementary nature of the representations learned by ViT when trained with SSAT.

In Figure 10, we present the Grad-CAM visualizations [47]. SSL+FT (3rd col) focuses on few specific pixelwise regions, while our method (4th col) focuses on areas corresponding to the entire primary object. We also provide the attention visualization of ViTs trained using different strategies in Appendix E (see Figure 12).

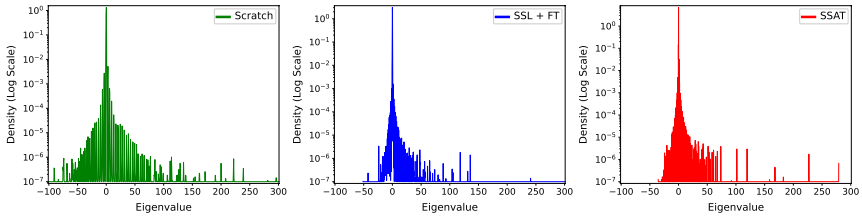


Figure 9. **Hessian max eigenvalues spectra** of ViTs trained from scratch (on the left), SSL + FT (in the middle), and SSAT (on the right).

Table 8. Cross training evaluation and zero-shot transfer results of DeepFake detection on FaceForensics++ with SSAT. [9] is trained on both DFDC and FaceForensics++, thus zero-shot transfer results have not been provided.

Method	cross-training evaluation				zero-shot transfer			
	Deepfakes	Face2Face	FaceSwap	NeuralTextures	Deepfakes	Face2Face	FaceSwap	NeuralTextures
Scratch	84.48	79.21	56.63	**82.08**	-	-	-	-
Cross-efficient-vit [9]	82.67	69.89	79.93	64.87	-	-	-	-
DFDC winner [46]	96.43	73.93	86.07	58.57	88.57	57.50	80.36	54.64
VideoMAE SSL (0.95)	82.67	64.16	58.42	63.44	86.28	49.82	69.18	51.97
VideoMAE SSL (0.75)	78.34	65.59	57.35	61.65	82.67	48.39	65.23	51.97
VideoMAE (0.95) + SSAT	92.42	79.21	89.61	81.36	**92.42**	**61.65**	**92.83**	**62.37**
VideoMAE (0.75) + SSAT	**96.75**	**80.65**	**91.40**	72.76	87.73	60.57	88.17	61.65

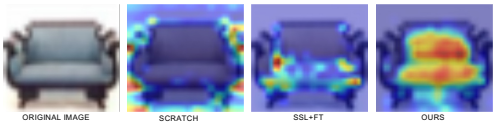


Figure 10. **GradCAM visualizations** of our SSAT model and the representative baselines.

5.4. Performance of SSAT in video domain

We have also assessed the efficacy of the SSAT within the video domain for the task of deepfake detection. In this experiment, the model's generalization capabilities for deepfake detection is validated, as presented in Table 8. For video encoding, we employ ViT as in [51]. Our VideoMAE + SSAT model is a direct extension of the MAE+SSAT model designed for image data; the only modification lies in the choice of encoder. The primary task involves binary classification to distinguish between real and manipulated videos. Notably, we experimented with two masking ratios, 0.75 and 0.95, during the training of VideoMAE + SSAT.

To assess model generalizability, we conducted cross-manipulation training based on the FaceForensics++ dataset [11]. We trained the model using videos generated by all possible combinations of three manipulation techniques (Deepfakes, Face2Face, FaceSwap, and NeuralTextures) plus original videos, and then evaluated its performance on the videos generated by the remaining technique. This approach simulates real-world scenarios where multiple manipulation techniques might be encountered post-training. All models, except the scratch model, are pre-trained on the DFDC dataset [46] before being evaluated on the FaceForensics++ dataset. To enable a fair comparison with VideoMAE + SSAT, the baseline VideoMAE SSL models are first pretrained and fine-tuned on DFDC, and are subsequently employed for deepfake classification task. The evaluation involved both (1) cross-dataset fine-tuning on FaceForensics++ and (2) zero-shot transfer assessment where pre-trained models are evaluated on FaceForensics++ without additional training.

Our findings, as detailed in Table 8, reveal that the VideoMAE+SSAT models demonstrate a superior generalized capability than the other baselines to distinguish between real and manipulated videos. Note that the scratch model outperforms all models on detecting videos generated using NeuralTextures without any pretraining but it is not suitable for zeros-shot transfer. Interestingly, the VideoMAE models exhibit complementary behavior when subjected to different masking ratios, which warrants a future investigation. More details including implementation and training details of these experiments can be found in Appendix D.

6. Conclusion

The main focus of this paper was on the use of self-supervised learning (SSL) to effectively train ViTs on domains with limited data. We demonstrate that by jointly optimizing the primary task of a ViT encoder with SSL as an auxiliary task, we can achieve discriminative representations for the primary task. This simple and easy-to-implement method called SSAT outperforms the traditional approach of sequentially training with SSL followed by fine-tuning on the same data. Our joint training framework learns features that are different from those learned by the dissociated framework, even when using the same losses. These results highlight the potential of SSAT as an effective training strategy with a lower carbon footprint. We anticipate that SSAT will become a standard norm for training vision transformers on small datasets.

Acknowledgments

We thank the members of the Charlotte Machine Learning Lab at UNC Charlotte for valuable discussion. This work is supported by the National Science Foundation (IIS-2245652).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. VIVIT: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 2

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Int. Conf. on Mach. Learn.*, July 2021. 2

[3] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 14

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 5

[5] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1

[6] Richard J Chen and Rahul G Krishnan. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*, 2022. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 5

[8] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv preprint arXiv:2107.06278*, 2021. 2

[9] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pages 219–229. Springer, 2022. 8

[10] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael Ryoo, and Francois Bremond. MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection. In *CVPR*, 2022. 2

[11] Abhijit Das, Srijan Das, and Antitza Dantcheva. Demystifying attention mechanisms for deepfake detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7. IEEE, 2021. 8

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1, 3, 4, 5, 12

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1, 2

[14] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2

[15] Hanan Gani, Muzammal Naseer, and Mohammad Yaqub. How to train vision transformer on small-scale datasets? *arXiv preprint arXiv:2210.07240*, 2022. 1, 2, 7

[16] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 2, 3, 12

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[18] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer, 2021. 1

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 3

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3, 5, 12

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[23] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[24] Saarthak Kapse, Srijan Das, and Prateek Prasanna. Cdnet: Histopathology representation learning using pyramidal context-detail network, 2022. 1

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 3, 12

[26] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019. 4

[27] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *Int. Conf. on Mach. Learn.*, pages 5639–5650. PMLR, 2020. 2

[28] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *CoRR*, abs/2112.13492, 2021. 1, 2, 7

[29] Kehan Li, Runyi Yu, Zhennan Wang, Li Yuan, Guoli Song, and Jie Chen. Locality guidance for improving vision transformers on tiny datasets. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 110–127. Springer, 2022. 1, 2, 7

[30] Xiang Li, Jinghuan Shang, Srijan Das, and Michael S Ryoo. Does self-supervised learning really improve reinforcement learning from pixels? In *Advances in Neural Information Processing Systems*, 2022. 2

[31] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *arXiv:2205.10063*, 2022. 12

[32] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 2

[33] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco De Nadai. Efficient training of visual transformers with small datasets. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 5, 7

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, pages 10012–10022, 2021. 1, 2, 3, 4, 12

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 2

[36] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 3, 12

[37] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 1, 3, 12

[38] Namuk Park and Songkuk Kim. Blurs behave like ensembles: Spatial smoothings to improve accuracy, uncertainty, and robustness. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17390–17419. PMLR, 17–23 Jul 2022. 6

[39] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 1, 2, 6

[40] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *International Conference on Learning Representations*. 5

[41] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 3, 5, 12

[42] AJ Piergiovanni, Anelia Angelova, and Michael S. Ryoo. Evolving losses for unsupervised video representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[43] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2874–2884, 2022. 2

[44] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. 3, 13

[45] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive Space-Time Tokenization for Videos. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[46] Selim Seferbekov. Dfdc 1st place solution, 2020. 3, 8, 13, 14

[47] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 7

[48] Jinghuan Shang, Srijan Das, and Michael S Ryoo. Learning viewpoint-agnostic visual representations by recovering tokens in 3d space. In *Advances in Neural Information Processing Systems*, 2022. 4

[49] Thomas Stegmüller, Antoine Spahr, Behzad Bozorgtabar, and Jean-Philippe Thiran. Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification. *arXiv preprint arXiv:2202.07570*, 2022. 1

[50] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pages 843–852. IEEE Computer Society, 2017. 1

[51] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 2, 8, 13

[52] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers: Distillation through attention. In *Int. Conf. on Mach. Learn.*, volume 139, pages 10347–10357, July 2021. 2, 3, 4

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2, 6

[54] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVTv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021. 1, 2, 12

[55] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 2

[56] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on*

Medical Image Computing and Computer-Assisted Intervention, pages 186–195. Springer, 2021. 1

[57] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 1, 2, 3, 4

[58] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 2

[59] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2, 3, 12

[60] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 3, 12

[61] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 2

[62] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021. 2

[63] Chuang Zhu, Wenkai Chen, Ting Peng, Ying Wang, and Mulan Jin. Hard sample aware noise robust learning for histopathology image classification. *IEEE Transactions on Medical Imaging*, 41(4):881–894, 2021. 3, 12

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2

Appendix

A. Dataset Description

This paper presents a comprehensive evaluation of our ViT models on 10 different image datasets, comprising prominent computer vision benchmarks such as ImageNet-1K [12] (IN-1K), CIFAR-10 and CIFAR-100 [25], Oxford Flowers102 [37], and SVHN [36]. In addition, we include three datasets namely ClipArt, Infograph, and Sketch from DomainNet [41], a widely adopted benchmark for domain adaptation tasks. Moreover, we explore the performance of our approach on two medical image domain datasets: Chaoyang [63] and PneumoniaMNIST [60]. The dataset size, sample resolution, and the number of classes are further elaborated in Table 9. Note that the accuracies reported for CIFAR in Figure 1 of the main paper is an average of the classification accuracy of CIFAR-10 and CIFAR-100.

Table 9. Details of image classification datasets (sample size, resolution, and number of classes) evaluated in our experiments.

Dataset	Train Size	Test Size	Dimensions	# Classes
CIFAR-10	50,000	10,000	32×32	10
CIFAR-100	50,000	10,000	32×32	100
Flowers102	2,040	6,149	224×224	102
SVHN	73,257	26,032	32×32	10
ImageNet-1K	1,281,167	100,000	224×224	1000
ClipArt	33,525	14,604		
Infograph	36,023	15,582	224×224	345
Sketch	50,416	21,850		
Chaoyang	4,021	2,139	512×512	4
PMNIST	5,232	624	28×28	2

Table 10. Ablation of decoder depth.

Decoder Depth	Accuracy	
	CIFAR-10	CIFAR-100
1	91.59	68.41
2	**91.65**	**69.64**
4	90.88	67.46
8	90.59	67.78
12	91.08	66.94

B. Training Configurations

We follow the configurations introduced in MAE [20]. A comprehensive set of training configurations for all datasets used in this study is provided in Table 14 for reference. During training two parameters, **image and patch sizes** vary depending upon the datasets and the rest of the parameters are the same across all the datasets.

Table 11. Ablation of decoder embedding dimension.

Decoder Dimension	Accuracy	
	CIFAR-10	CIFAR-100
64	89.20	67.11
128	**91.65**	**69.64**
256	91.64	66.98
512	90.53	66.19

Table 12. Ablation of Decoder Heads.

Decoder Heads	Accuracy	
	CIFAR-10	CIFAR-100
1	91.54	69.44
2	92.44	69.28
4	**92.49**	68.59
8	92.09	69.52
16	91.65	**69.64**

Table 13. Statistics of video datasets generated by different manipulating techniques available in Faceforensics++

Split	DeepFake	Face2Face	FaceSwap	NeuralTextures	Original	Total
Train	720	720	720	720	720	3600
Val	140	140	140	140	140	700
Test	140	140	140	140	140	700
Total	1000	1000	1000	1000	1000	5000

Swin and ConvMAE training configurations: We have adopted the training pipeline from [31] and [16] for Swin [34] and ConvMAE [16] respectively. For each of them, we have combined their reconstruction based self-supervised learning (SSL) and fine-tuning in a joint learning framework, keeping the training configurations same. Note that UM-MAE [31] with its secondary masking strategy, is an efficient version of SimMIM [59] allowing the reconstruction based SSL in hierarchical transformers like Swin [34] and PVT [54].

C. ViT Decoder for Reconstruction based SSL

In contrast to MAE [20], this paper employs a reconstruction-based SSL approach with class-wise supervision. Consequently, we explore the effect of different design choices of the ViT decoder, which can impact the SSL training while simultaneously optimizing cross-entropy in Self-supervised Auxiliary Task (SSAT). To this end, we conduct experiments that involve modifying three decoder attributes: **depth**, **dimension**, and **attention heads**. We evaluate the resulting impact on the model's top-1 accuracy using two datasets: CIFAR-10 and CIFAR-100.

Decoder Depth: In this study, we investigated the im-

pact of decoder depth on model performance, as shown in Table 10. During the experiments, we maintained a fixed decoder dimension of 128, decoder heads of 16, and a value of λ equal to 0.1. Our findings demonstrate that the optimal results for both datasets were obtained at a decoder depth of 2.

Decoder Embedding Dimension: This section investigates the influence of the decoder embedding dimensions on model performance, as presented in Table 11. Throughout these experiments, we maintained a constant value of λ at 0.1, a decoder depth of 2, and 16 decoder heads. Our results indicate that the optimal performance was achieved with a decoder dimension of 128.

Decoder Heads: Table 12 presents the outcomes of the ablation study performed to evaluate the impact of the number of heads on the ViT's performance. The hyperparameters, namely $\lambda = 0.1$, decoder_depth $= 2$, and decoder_dimension $= 128$, are fixed to their optimal values from the prior experiments. The experimental findings indicate that retaining 4 heads for CIFAR-10 and 8 heads for CIFAR-100 resulted in the highest performance levels. To ensure generalizability across our experiments, we fixed the number of decoder heads to 16.

D. Details of deepfake detection experiments

In this section, we elaborate the cross-training manipulation and zero-shot transfer experimental details for deepfake detection.

D.1. Datasets

We employ two publicly available popular dataset on Deepfakes.

FaceForensics++: The FaceForensics++ dataset [44] is a large-scale benchmark dataset for face manipulation detection, which is created to help develop automated tools that can detect deepfakes and other forms of facial manipulation. The dataset consists of more than 1,000 high-quality videos with a total of over 500,000 frames, which were generated using various manipulation techniques such as facial reenactment, face swapping, and deepfake generation.

The videos in the dataset are divided into four categories, each corresponding to a different manipulation technique: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. Deepfakes use machine learning algorithms to generate realistic-looking fake videos, while Face2Face and FaceSwap involve manipulating the facial expressions and identity of a person in a video. NeuralTextures uses a different approach by altering the texture of a face to make it appear different. The dataset includes both real and manipulated videos, with each manipulation technique applied to multiple individuals. The statistics of different manipulating techniques available in faceforensics++ is provided in Table 13.

DFDC: The Deepfake Detection Challenge (DFDC) dataset [46] is a large-scale benchmark dataset for deepfake detection. The dataset consists of more than 100,000 videos generated using various facial modification algorithms. The DFDC dataset consists of two versions: a preview dataset with 5k videos featuring two facial modification algorithms and a full dataset with 124k videos featuring eight facial modification algorithms. The DFDC dataset is the largest currently and publicly available face swap video dataset, with around 120,000 total clips sourced from 3,426 paid actors. The videos are produced using several Deepfake, GAN-based, and non-learning methods. The official DFDC train, validation and test splits are also designed to simulate real-world performance, with the validation set consisting of a manipulation technique not present in the train set, and the test set containing much more challenging augmentations and perturbations.

D.2. Methodology

VideoMAE [51]: VideoMAE is a self-supervised video pre-training method that extends masked autoencoders (MAE) to videos. VideoMAE performs the task of masked video modelling for video pre-training. It employs an extremely high masking ratio (90%-95%) and tube masking strategy to create a challenging task for self-supervised video pre-training. The temporally redundant video content enables a higher masking ratio than that of images. This is partially ascribed to the challenging task of video reconstruction to enforce high-level structure learning.

SSAT: In this experiment, we use the same backbone as in the original work [51] and we use rigorous augmentations as used by the winners of the DFDC Challenge [46] in our experimental setting. For training VideoMAE along with SSAT on DFDC, we extend our image based framework to videos (as illustrated in Figure 11) and jointly optimize the primary deepfake classification loss L_{cls} and the auxiliary video reconstruction loss L_{SSAT} as

$$L = \lambda * L_{cls} + (1 - \lambda) * L_{SSAT} \tag{2}$$

where $\lambda = 0.1$ is the loss scaling factor.

D.3. Implementation details

While training VideoMAE+SSAT models follow the training recipe of [51], we have incorporated specific modifications tailored for deepfake detection.

Fake class weight: Assigns weight w to the class representing *fake* in the weighted cross entropy loss. This was used since the training set is very imbalanced (82% fake - 18% real).

$$L_{CE} = -(wt_{real} \log p_{real} + (1 - w)t_{fake} \log p_{fake}) \tag{3}$$

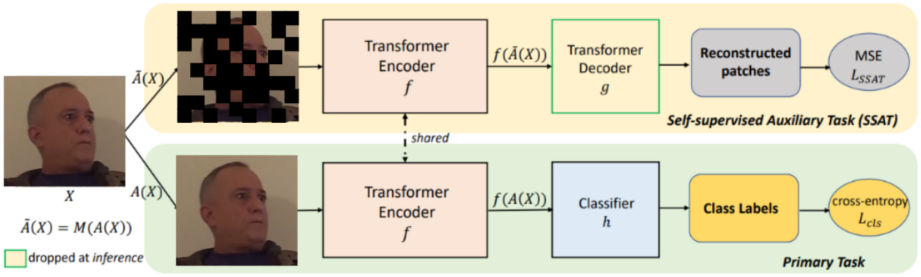


Figure 11. Mask Autoencoder as a Self Supervised Auxiliary Task for deepfake detection.

Equation 3: Weighted Cross-Entropy Loss. w is the weight of the *real* class while p_{real} and p_{fake} are the predicted probabilities, and t_{real} and t_{fake} are the ground truth indicator variables.

Augmentations: The choice of augmentations has a profound impact on validation performance. The set of augmentations that work best are Image Compression, Gaussian Noise, Gaussian Blur, Horizontal Flip, Brightness Contrast, FancyPCA, Hue Saturation, Greyscale and shift-scale-rotate, all available in the Albumentations library [3] and used in the DFDC challenge's winning solution by Selim Seferbekov [46]. Other augmentations like Reversal, Random up / down sampling and heavy Gaussian Noise seem to have a detrimental effect, possibly because they do not generalize to the validation set. Meanwhile, having no augmentations also decreases the generalizability.

Testing: During testing, predictions are obtained by averaging the results from all 16-frame segments across the entire video.

E. Attention Visualization

In Figure 12, we illustrate attention visualization for a few sample images drawn from the Flower and ImageNet datasets. Our analysis of the visualization highlights that the ViT trained with SSAT generates attention maps that emphasize the primary object class to a greater extent than the attention maps computed by ViTs trained from scratch and trained with SSL+FT. These findings indicate that the ViT trained with SSAT exhibits higher efficacy in image classification.

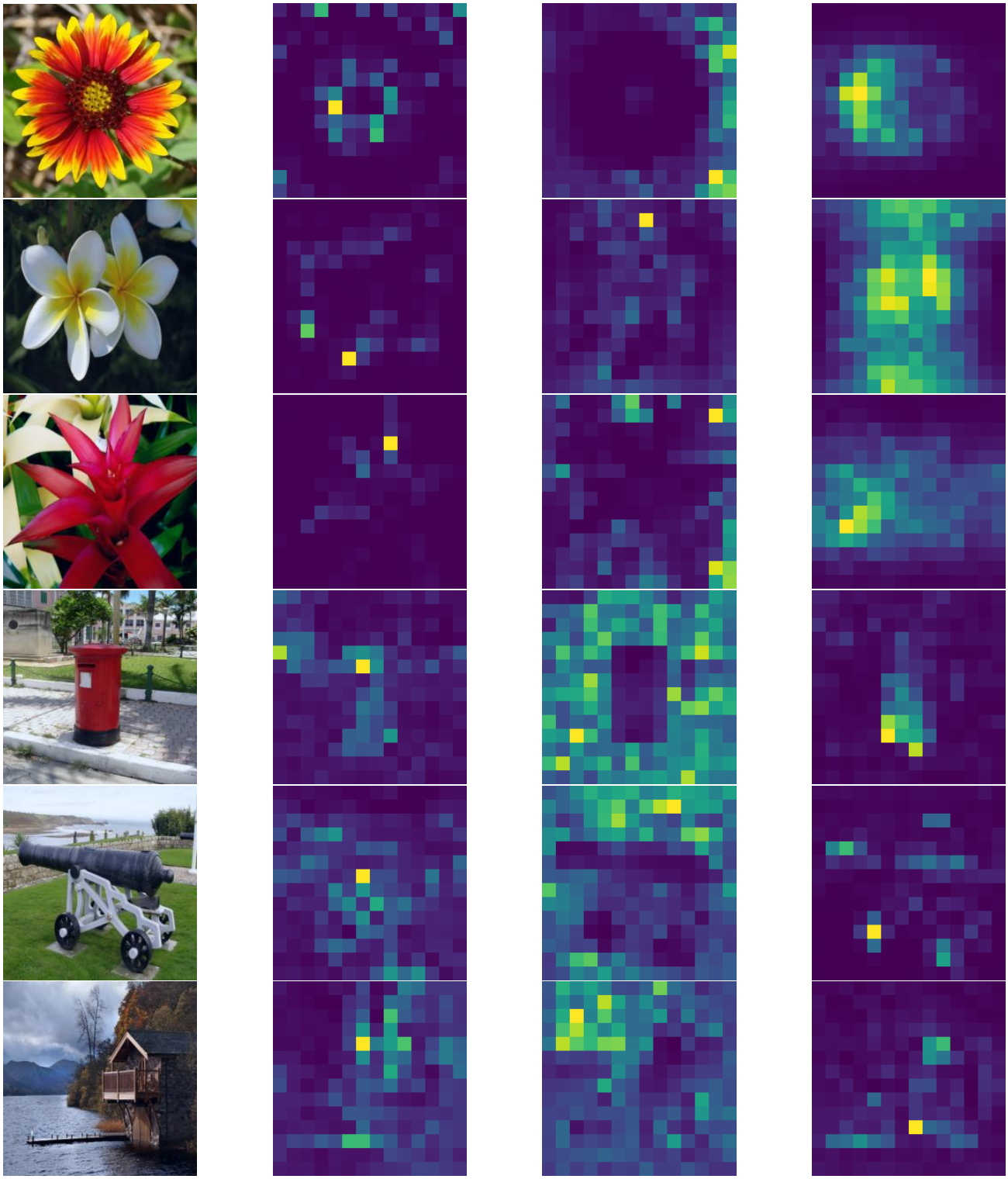


Figure 12. Attention visualization of six images, three from the Oxford Flowers-102 dataset (top 3 rows) and three from the ImageNet dataset (bottom 3 rows). The attention heatmaps in the second, third, and fourth columns correspond to models trained from scratch using ViT, models trained using SSL+FT, and models trained using SSAT, respectively.

Table 14. Our ViT training settings across different datasets.

Input Size	PMNIST	28×28
	CIFAR10, CIFAR100, SVHN	32×32
	Flowers, ImageNet-1K ClipArt, Infograph, Sketch	224×224
	Chaoyang	512×512
Patch Size	PMNIST, CIFAR10, CIFAR100, SVHN	2×2
	Flowers, ImageNet-1K ClipArt, Infograph, Sketch	16×16
	Chaoyang	32×32
Batch Size	64	
Optimizer	AdamW	
Optimizer Epsilon	1e-08	
Momentum	$\beta_1 = 0.9,\ \beta_2 = 0.999$	
layer-wise lr decay	0.75	
Weight Decay	0.05	
Gradient Clip	None	
Learning Rate Schedule	Cosine	
Learning Rate	1e-03	
Warmup LR	1e-06	
Min LR	1e-6	
Epochs	100	
Warmup Epochs	5	
Decay Rate	0.05	
Drop Path	0.01	
Lambda (λ)	0.1	
Masking Ratio	0.75	
Random Resized Crop Scale, Ratio	(0.08, 1.0), (0.75, 1.3333)	
Interpolation	bicubic	
Random Horizontal Flip Probability	0.5	
Rand Augment	n = 2	
Random Erasing Probability, Mode and Count	0.25, Pixel, (1, 1)	
Color Jittering	None	
Auto-augmentation	rand-m9-mstd0.5-inc1	
Mixup	True	
Cutmix	False	
Mixup, Cutmix Probability	1, 0	
Mixup Switch Probability	0.5	
Mixup Mode	Batch	
Label Smoothing	0.1	

Symbiosis
https://doi.org/10.1007/s13199-023-00945-5

Microbial inoculants improve growth in Zea *mays* L. under drought stress by up-regulating antioxidant, mineral acquisition, and ultrastructure modulations

Madhulika Singh[1,2] · Jai Gopal Sharma[1] · Bhoopander Giri[2]

Received: 5 July 2023 / Accepted: 28 September 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract
Drought-induced stress poses a formidable impediment to agricultural productivity, compromising crop growth and yield. While the utilization of arbuscular micorrhizal fungi (AMF) consortium, plant growth promoting bacteria (PGPB) consortium, and *Piriformospora indica* inoculum shows promise in mitigating drought effects, a rigorous scientific assessment of their relative efficacy is lacking. This study thus investigates the comparative effects of *P. indica*, AMF consortium, and PGPB consortium inoculation on nutrient assimilation, ultrastructural adjustments, metabolic responses, and antioxidant defenses in drought-exposed maize plants. A completely randomized block design was employed, subjecting plants to varying drought stress levels (75%, 55%, 45%, and 35% of soil field capacity (FC)) in a greenhouse. The study analyzed multiple aspects, including growth parameters, nutrient levels, physiological responses, osmolyte content, antioxidant defense mechanisms, and leaf ultrastructure. Inoculation with AMF, *P. indica*, and PGPB yielded substantial enhancements in growth metrics like plant height, biomass accrual, leaf area expansion, stomatal conductance, photosynthetic activity, internal CO_2 dynamics, and transpiration rate, particularly evident across all drought stress intensities. Notably, microbial treatments counteracted nutrient deficiencies induced by drought, with augmented levels of osmolytes (total sugars, trehalose, glycine betaine), proteins, attenuated lipoxygenase enzyme (LOX) activity, ion leakage, and lipid peroxidation relative to non-inoculated plants. Elevated enzymatic activity of antioxidants, including ascorbate peroxidase (APX), catalase (CAT), and superoxide dismutase (SOD), in addition to elevated non-enzymatic antioxidants (proline, α-tocopherol, polyamines), was evident in microbial-inoculated plants. Furthermore, microbial inoculation ameliorated drought-induced ultrastructural alterations. Noteworthy disparities in drought resilience were observed among inoculants, with PGPB displaying the highest robustness, followed by AMF and *P. indica*. This research underscores the substantial potential of microbial inoculants in fostering sustainable agriculture practices in drought-challenged contexts.

Keywords Antioxidants · Cell ultrastructure · Minerals · Osmolyte · Photosynthesis · Polyamines

1 Introduction

Drought, being a prominent abiotic stressor, presents a substantial peril to worldwide food security and agricultural production (Mirzabaev et al. 2023). Decreased rainfall and prolonged dry periods contribute to drought stress, affecting nearly 45% of agricultural fields (Yadav et al. 2020; Jarrett et al. 2023). These environmental conditions adversely impact plant growth by reducing nutrient accumulation (Shah et al. 2023), osmolyte content (Haghighi et al. 2023), photosynthesis, and transpiration rates (Naservafaei et al. 2023). Drought-stressed plants face challenges such as ion leakage, reactive oxygen species (ROS) generation, and lipid peroxidation (Yu et al. 2023) leading to irreversible damage and decreased crop yields. However, plants can employ strategies like antioxidant defense mechanisms, stomatal closure, and osmotic adjustment to mitigate these effects (López-Galiano et al. 2019). Addressing drought stress is crucial for sustaining agricultural productivity and ensuring food availability.

✉ Bhoopander Giri
bhoopg@yahoo.com

[1] Department of Biotechnology, Delhi Technological University, Delhi, India

[2] Department of Botany, Swami Shraddhanand College, University of Delhi, Delhi, India

Published online: 27 October 2023

Maize (*Zea mays* L.), a pivotal cereal crop belonging to the Poaceae family, holds significant agricultural importance. The Maharaja (BIO-9637) hybrid maize variety exhibits a growth cycle of 110–115 days. This variety displays resistance and tolerance to prevalent diseases and pests. With a moderate plant height, it offers a high-yielding capacity. Notably, its kernels boast a productive profile characterized by an orange-yellow coloration and a semi-flint grain texture. It has potential tolerance to drought stress (Srinivasa et al. 2016). It is cultivated during the Kharif season, providing a staple food for humans and serving as livestock feed. Maize is also a key raw material for starch, glucose, and cornflake production (Sinha et al. 2019). Nevertheless, abiotic stress exerts a pronounced impact on maize productivity (Ali et al. 2022).

Drought stress poses significant challenges to agricultural productivity, and various mitigation approaches have been explored. These include film farming (Mori 2015), cultivation of drought-tolerant crops (Diatta et al. 2021), utilization of nanoparticles (Silva et al. 2023), incorporation of super-absorbent hydrogels, and application of biochar (Waqar et al. 2022). However, these strategies often involve high costs and complex technical procedures. In recent years, the utilization of plant growth-promoting microorganisms has emerged as a compelling and natural strategy for augmenting plant drought tolerance (Ramakrishna et al. 2020). Among these microorganisms, certain PGPB such as *Rhizobium* (Komatsu et al. 2023), *Bacillus* (Mahreen et al. 2023), *Azotobacter* (Hashemi et al. 2022), and *Enterobacter* (Jofre et al. 2023), and as well as endophytic fungi like *P. indica* (Azizi et al. 2021), AMF (Wang et al. 2023; Bisht and Gupta 2023), have exhibited the capacity to stimulate plant growth in abiotic stress. These microbes synthesize bioactive compounds and exhibit adaptive responses to drought-affected environments (Tyagi et al. 2023). *P. indica* functions as a plant endophyte, augmenting nutrient assimilation, bolstering antioxidant defense mechanisms, and fortifying systemic resistance (Rajput et al. 2022). AMF is known for establishing mutualistic associations with more than 80% of plant species found in terrestrial ecosystems (van der Heijden et al. 2015; Chourasiya et al. 2021), improving nutrient uptake (Tyagi et al. 2021; Agnihotri et al. 2022), and modulating water-holding capacity, antioxidant defense mechanisms, and ionic balance (Fresno et al. 2023). PGPB enhances relative water content, osmotic adjustment, leaf area, and photosynthetic rate (Fadiji et al. 2022) through solubilizing phosphate, producing phytohormones, exopolysaccharides, and siderophores (Mahreen et al. 2023). In a study by Inbaraj (2021), inoculation with PGPB has been shown to improve plant biomass, chlorophyll pigments, membrane stability, proline level, and the activities of antioxidant enzymes in maize, rice, and wheat subjected to drought-induced stress conditions. Additionally, PGPB enhances plant nutrient assimilation and water uptake capacity during stressful conditions (Nemeskeri et al. 2022).

To date, there is a lack of comparative assessments evaluating the efficacy of the AMF consortium (*Scutellospora* sp, *Glomus mosseae*, and *Glomus intraradices*), PGPB consortium (*Lactococcus lactis*, *Enterobacter asburiae*, and *Azotobacter chroococcum*), and *P. indica* in alleviating drought stress in crop plants. While prior studies reveal the potential of AMF, *P. indica*, and PGPB to enhance drought tolerance in plants through osmolytes and membrane integrity, the aspect concerning their impact on preventing ultrastructural damage remains unexplored. While the role of macronutrients (K^+, Ca^{2+}, Na^+, Mg^{2+}) and micronutrients (Cu, Mn^{2+}, Fe, Zn^{2+}) in plants under stress has been explored (Rahmani et al. 2023), there remains a need for further investigation into the impacts of AMF consortium, *P. indica*, and PGPB consortium colonization on the interplay between membrane permeability and nutritional balance. This study aims to investigate the comparative impact of inoculation with *P. indica*, an AMF consortium, and a PGPB consortium on nutrient assimilation, ultrastructural modifications, metabolic responses, and antioxidant defense mechanisms in drought-stressed maize plants.

2 Materials and methods

The maize hybrid Maharaja (BIO-9637), acquired from the Pusa Complex in New Delhi, India, was developed as a single-cross hybrid by Bioseed Research Ltd. in the year 1999 (Hari et al. 2008). It displays distinctive traits viz. tall height, yellow-colored grains, wide leaves, moderate-sized tassel, and thick ears. Designated as a medium-maturing hybrid (A Saga of Progress 2012), "Maharaja" is distinguished by its resilience to common rust (Akhilesh et al. 2000), its adaptability to waterlogged conditions (Jyoti et al. 2013), and its susceptibility to drought (Kumar et al. 2015). After sterilization using a 4% NaOCl solution, the seeds were sown in mud pots measuring 10 inches in diameter, each filled with 5 kg of soil. The AMF and PGPB consortia were sourced from the Department of Microbiology at the IARI, India, while the inoculum of *Piriformospora indica* was obtained from Amity University, Noida, India. During seeding, the soil was inoculated with 20 g of PGPB (10^8 Colony Forming Unit (CFU) per gram of charcoal carrier), 2 g of *P. indica* (75–80 spores per gram of talc), and 50 g of AMF (70–75 spores g^{-1}

of soil), while the uninoculated plants received an equivalent amount of sand. The soil was autoclaved for sterility and had the following characteristics: sandy loamy, pH 7.3, organic matter 1.3%, available Nitrogen (N): 172 g kg^{-1}, available Phosphorus (P): 45.1 g kg^{-1}, and available Potassium (K): 476 g kg^{-1}, Mg^{2+}: 132 mg g^{-1}, Zn^{2+}: 1.97 mg g^{-1}, Fe^{3+}:6.78 mg g^{-1}, Cu^{2+}: 2.00 mg g^{-1}, Mn^{2+}: 11.62 mg g^{-1}.

2.1 Experimental setup and growth regime

In a greenhouse experiment, the potential impact of microbes on mitigating drought stress in maize plants was investigated. The greenhouse maintained a humidity of $77 \pm 2\%$ and a temperature of 36 ± 2 °C. The experimental design was configured in a 4×4 factorial arrangement consisting of two factors: (i) four types of microbial inoculations (non-inoculated control group, maize inoculated with *P. indica*, maize inoculated with PGPB and maize inoculated with AMF) and (ii) four levels of drought conditions (75%, 55%, 45%, and 35% FC). This resulted in a total of 16 treatments replicated three times. Before subjecting the plants to drought stress, a 20-day growth period allowed for microbial colonization. The plants were subjected to distinct water regimes (75%, 55%, 45%, and 35% FC) on a weekly basis over a period of 60 days. Harvesting of the plants took place at the V10 stage, which corresponds to the tenth-leaf stage, after 80 days from sowing for further analysis (Supplementary material; Fig. S1).

2.2 *P. indica* and AMF root colonization

The maize roots were rinsed using distilled water. To assess colonization, root samples were softened in a 10% KOH solution, followed by immersion in 1% hydrochloric acid for 5 min. The samples were stained with 0.05% trypan blue and treated with lactoglycerol to remove excess stain (Phillips and Hayman 1970). Microbial colonization was assessed employing the gridline intersect technique (Giovannetti and Mosse 1980). Intersections at which fungal structures or hyphae intersected the grid lines were enumerated using an electronic binocular compound microscope (Nikon ECLIPSE E100).

$$\%\ \text{Root colonization percentage} = \frac{\text{Total count of intersections with fungal structure}}{\text{Total count of intersections}} \times 100$$

2.3 Assessment of plant growth metrics

Shoot and root lengths were assessed using a calibrated measuring instrument. Biomass determination involved oven-drying plant samples at 75 °C for three days, and biomass was calculated based on the weight of dried samples. Leaf area was quantified employing a leaf area meter (leaf area meter CI-202).

2.4 Analysis of mineral nutrients

Plant samples (0.1 g) were digested with HNO_3 (3 ml) and H_2O_2 (5 ml) in a microwave digestion unit (Multiwave PRO). Following the digestion process, the samples were diluted to a final volume of 25 ml using distilled water. Standards with concentrations ranging from 0.1 to 10 ppm were prepared using a 2% HNO_3 solution. Elemental analysis was conducted using Agilent Technologies 7700 series Inductively-Coupled Plasma Mass Spectrometry (ICP-MS) unit. The results were analyzed using the Mass Hunter workstation software.

2.5 Assessment of physiological process

Physiological parameters of plants, including transpiration rate (E), internal CO_2 content (Ci), photosynthesis rate (PN), and stomatal conductance (gs), were quantified using a portable photosynthesis system (LI-6400XT). Measurements were conducted between 8:30 am and 12:00 pm using a 2 cm^2 leaf chamber with a diode light source emitting red-blue light at a photosynthetic active radiation intensity of 1200 µmol $m^{-2}s^{-1}$ and a CO_2 concentration of 400 ppm. A light-response curve was constructed to ascertain the light saturation rate (Bidalia et al. 2018).

2.6 Measurement of chlorophyll and carotenoid

The estimation of photosynthetic pigments in plant samples followed the protocol established by Hiscox and Israelstam (1979). The samples (100 mg) were fragmented and introduced into a vial with 7 ml of dimethyl sulfoxide (DMSO). The vials were subjected to incubation at 65 °C until tissue decolorization was achieved. The solution volume was adjusted to 10 ml using DMSO. Absorbance readings were taken at 480 nm, 510 nm, 645 nm, and 663 nm. Calculation formulas established by Arnon (1949) were employed to quantify photosynthetic pigment content, with values expressed as mg g^{-1}fw.

$$\text{Chlorophyll a} = (12.7 \times D_{663} - 2.69 \times D_{645} \times V)/\text{Weight of sample} \times 1000$$

$$\text{Chlorophyll b} = (22.9 \times D_{645} - 4.68 \times D_{663} \times V)/\text{Weight of sample} \times 1000$$

$$\text{Total Chlorophyll} = (20.2 \times D_{645} + 8.02 \times D_{663} \times V)/\text{ Weight of sample} \times 1000$$

$$\text{Carotenoids} = 7.6 \times D_{480} - 1.49 \times D_{510} \times V/\text{Weight of sample} \times 1000$$

2.7 Analysis of osmolyte regulators

The quantification of glycine betaine content within the plant sample was carried out in accordance with the methodology of Grieve and Grattan (1983) and Singh et al.

(2023). The determination of the content of total sugar in plant samples was accomplished utilizing the anthrone method as detailed by Sadasivam and Manickam (2008). Plant samples (0.1 g) were treated with 2.5 N hydrochloric acid in a water bath for 3 h. After neutralization with sodium carbonate, the mixture was centrifuged at 5000 rpm for 15 min and the supernatant (500 µl) was treated with 4 ml of anthrone reagent. Absorbance at 630 nm was recorded, and sugar content was determined using a standard curve of glucose. Trehalose content in plant samples was determined by pulverizing them in liquid nitrogen (Ferreira et al. 1997). The resulting sample was boiled in ethanol (2 ml), evaporated and the remaining residue was reconstituted in a 5 ml aliquot of the mobile phase (acetonitrile and water). After centrifugation at 10,000 rpm for 10 min and filtration, the supernatant underwent heating, pH adjustment, and evaporation. The residue was solubilized in acetonitrile and water and analyzed using a High-Performance Liquid Chromatography (HPLC) unit (Agilent-1200 InfinitySeries) equipped with a column (C-8, 150 mm×4.6 mm, 5 µm) and flow rate of 500 µl/min. A 20 µl sample was injected, and the eluted compounds were analyzed using a refractive-index detector (Hewlett-Packard 1037 A), quantification was achieved by comparing it with a standard trehalose concentration.

2.8 Protein content

Protein content in plant samples was determined using the analytical Kjeldahl method according to the EN ISO 20483:2006 (AN 300) standard. The method involved the digestion of 1000 mg of the dried sample with 10 ml of concentrated sulphuric acid, 0.8 g of copper sulfate pentahydrate, and 7 g of potassium sulfate. The mixture was digested at 420 °C for 1 h (DT208 Labtech TM Line FOSS), which was connected to a scrubber (KSC 01, Labquest Borosil) to eliminate any hazardous fumes. After cooling, water and sodium hydroxide were added, and distillation was performed. In the receiver, a 4% H_3BO_3 solution containing indicators was placed. Prior to processing each batch of samples, a reagent blank was performed to account for potential background interference. The 0.1 M hydrochloric acid was used to titrate the content of the conical flask and the volume of hydrochloric acid required for neutralization was measured as the titre. This measurement was used to calculate the percentage of nitrogen content.

$$\% \text{ Nitrogen} = \frac{(\text{Sample titration} - \text{Blank titration}) \times \text{Normality of titrant} \times 14.007}{\text{Weight (mg)}} \times 100$$

$$\% \text{ Protein} = \text{Conversion Factor} \times \% \text{ Nitrogen}$$

Factor 6.25 used as conversion factor for calculating protein from N of seed/plant samples.

2.9 Estimation of membrane damage

Lipid peroxidation levels within the samples were assessed in accordance with the method established by Heath and Packer (1968). The relative permeability of leaf discs was assessed by incubating them in water and measuring the conductivity of the liquid using a conductivity meter. After subjecting the discs to a heating and cooling cycle, conductivity was once more measured (Zwiazek and Blake 1991). LOX activity was determined by grinding leaf samples in a phosphate buffer, centrifuging, and mixing the extracted content with a substrate solution. The content was then centrifuged for 15 min at 4 °C. The extracted (0.05 ml) content was mixed with 2.95 ml of substrate solution. The substrate solution had linoleic acid dissolved in H_2O having Tween-20. The absorbance of the content at a wavelength of 234 nm was quantified via spectrophotometric analysis (Doderer et al. 1992).

2.10 Assay of enzymatic antioxidants

The plant sample weighing 500 mg was grounded in a phosphate buffer having ethylenediamine tetraacetic acid. The content was centrifuged and the supernatant was utilized for measuring enzyme activity. CAT activity was assessed in accordance with the procedure detailed by Teranishi et al. (1974). The assessment of SOD activity was carried out as per the protocol described by Dhindsa et al. (1981). The analysis of APX activity followed the established protocol outlined by Nakano and Asada (1981).

2.11 Assay of non-enzymatic antioxidants

To estimate the α-tocopherol plant sample was pulverized in liquid nitrogen and ground using sulfuric acid (0.1 N), as per the procedure delineated by Sadasivam and Manickam (2008). Following this, the resultant homogenate was subjected to centrifugation at 10,000 rpm for a duration of 10 min. To the supernatant, the same amount of C_2H_6O was introduced and centrifuged again at 10,000 rpm for 5 min. An equal volume of xylene was introduced to the mixture, followed by a subsequent centrifugation step. The upper layer was aspirated and combined with an equivalent volume of 2,2´-bipyridyl. After incubating the content at room temperature for 15 min, $FeCl_3$ solution was poured in. At 520 nm the absorbance was recorded. The measurement of proline content in the samples was conducted following the method outlined by Bates et al. (1973) and Singh et al. (2023). Polyamine quantification in plant samples was performed following a modified protocol by Marcé et al. (1995). A 100 mg plant sample was treated with 5% $HClO_4$, and the aqueous phase obtained after centrifugation was stored at -20 °C (Minocha et al. 1994). The dansylation process

involved treating an aliquot with 0.05 M diamino-heptane, sodium carbonate, and dansyl chloride. After the incubation period, a proline solution was introduced, and subsequently, extraction was carried out using toluene. The organic fraction was gathered, organic phase was collected, evaporated, and solubilized in acetonitrile. The filtered solution was analyzed using an HPLC system (Agilent-1200 InfinitySeries) at1000 µl/min flow rate with an OROChem C-18 column and a fluorescence detector (Hewlett Packard 1046 A), employing acetonitrile and 1-heptane-sulphonic acid as the mobile phase.

2.12 Estimation of ultrastructural modulations

Plant samples were fixed in a solution of glutaraldehyde, paraformaldehyde, and phosphate buffer, followed by osmium tetraoxide fixation. Dehydration was performed using a series of propylene oxide and ethanol, and infiltration was done with epon araldite and propylene oxide mixtures. Embedded samples were sectioned with an ultramicrotome, and stained with uranyl acetate. After washing with 50% alcohol and Milli Q water, lead citrate was used to stain the sections, followed by a rinse with sodium hydroxide and drying. The sections were analyzed using a transmission electron microscope (Thermoscientific, Talos). This analysis aimed to investigate alterations in the plasma membrane and chloroplast ultrastructure of mesophyll cells.

2.13 Statistical analysis

The data analysis utilized IBM SPSS Statistics 21 software. A two-way analysis of variance was conducted, considering microbes and drought as independent treatments. Significant differences among means were further examined using a post hoc Duncan's test with a significance level of $P < 0.05$.

3 Results

3.1 Microbial colonization

The symbiotic associations between AMF and *P. indica* with maize roots were successful under different water stress levels. However, AMF colonization was higher than *P. indica* colonization during drought conditions. Non-inoculated plants showed no microbial colonization. Increasing water stress resulted in decreased microbial colonization (Table 1).

3.2 Plant growth and biomass

The growth characteristics of maize plants were examined under different water regimes with and without microbial inoculation (Table 1). Maize plants inoculated with microbes exhibited significant improvements in plant height, biomass, and leaf area compared to non-inoculated plants across all water regimes. Under extreme drought conditions (35% Field capacity), PGPB inoculation resulted in the highest enhancements in shoot and root length (27.76% % and 41.86% respectively), followed by AMF inoculation, which exhibited improvements of 22.74% and 31.30% in shoot and root length, respectively. *P. indica* inoculation showed increases of 17.51% and 9.58% in shoot and root length, respectively, in comparison to non-inoculated plants (Supplementary Material; Table S1). Furthermore, PGPB-inoculated plants subjected to 35% FC demonstrated an 84.86% increase in shoot biomass and a 66.25% increase in root biomass, whereas AMF inoculation led to improvements of 60.25% and 51.87% and *P. indica* by 53.32% and 20.06% in shoot and root biomass, respectively, in comparison to non-inoculated plants (Supplementary Material; Table S2). Maize leaf area increased substantially with PGPB (28.69%), followed by AMF (19.18%) and *P. indica* (10.31%) inoculation compared to non-inoculated plants at 35% FC drought condition (Supplementary Material; Table S3). Two-way analysis of variance indicated significant independent effects of water stress and microbial inoculations on plant growth (Table 3). PGPB-inoculated plants demonstrated higher drought resistance compared to AMF and *P. indica*-inoculated plants.

3.3 Mineral nutrients content

Table 2 provides comprehensive data on the macronutrient (Ca^{2+}, Mg^{2+}, K^+, and Na^+) and micronutrient (Zn^{2+}, Fe, Mn^{2+}, and Cu) composition of maize plants subjected to drought stress and microbial colonization. Both stress and microbial inoculation significantly influenced nutrient levels. Increasing drought stress resulted in notable reductions in Ca^{2+}, Mg^{2+}, K^+, Zn^{2+}, Fe, Mn^{2+}, and Cu levels, accompanied by an increase in Na^+ levels in both inoculated and non-inoculated plants. However, microbial-inoculated plants consistently exhibited higher levels of Ca^{2+}, Mg^{2+}, K^+, Zn^{2+}, Fe, Mn^{2+}, and Cu, along with lower Na^+ levels compared to non-inoculated plants under equivalent drought stress conditions (Table 2). At 35% field capacity (FC), PGPB-inoculated plants displayed the highest increase in Ca^{2+} (177.30%), followed by *P. indica*-inoculated plants (132.94%) and AMF-inoculated plants (121.94%). Similarly, at 35% FC, PGPB-inoculated plants exhibited the greatest increases in Mg^{2+} (469.95%) and K^+ (161.74%), followed by AMF-inoculated plants (231.75% and 122.25%, respectively) and *P. indica*-inoculated plants (280.93% and 30.44%, respectively). Furthermore, at 35% FC, PGPB-inoculated plants showed the highest decrease in Na^+ content (30.36%), followed by *P. indica*-inoculated plants (22.27%), and AMF-inoculated plants (16.42%). Regarding micronutrients, at 35% FC, PGPB, AMF, and *P. indica*-inoculated plants exhibited

Table 1 Impact of microbial inoculation and drought stress on root colonization, shoot length, root length, shoot biomass, root biomass, and leaf area of maize plants

FC (%)	Microbe	Shoot length (cm)	Root Length (cm)	Shoot biomass (g)	Root biomass (g)	Leaf area (cm^2)	Root colonization (%)
75	Control	67.60 ± 2.07def	24.97 ± 0.42j	6.41 ± 0.48gh	2.79 ± 0.07ghi	55.58 ± 0.87hi	
	P. indica	69.10 ± 0.17cde	29.00 ± 1.24f	7.58 ± 0.28de	3.83 ± 0.57 cd	63.31 ± 2.81ef	67
	PGPB	75.33 ± 2.01a	34.64 ± 1.07a	8.81 ± 0.63a	4.91 ± 0.54a	71.66 ± 2.52 a	
	AMF	73.22 ± 1.12ab	32.77 ± 2.11c	8.11 ± 0.08bcd	4.48 ± 0.13ab	69.16 ± 1.55abc	72
55	Control	66.83 ± 1.04efg	25.69 ± 0.68i	6.12 ± 0.43 h	2.73 ± 0.03ghi	55.33 ± 0.31hi	
	P. indica	69.04 ± 1.4cde	27.89 ± 0.39gh	7.46 ± 0.14ef	3.31 ± 0.20ef	662.88 ± 1.34efg	62
	PGPB	75.37 ± 1.64a	33.77 ± 2.13b	8.45 ± 0.39ab	4.72 ± 0.20ab	70.22 ± 0.26ab	
	AMF	73.33 ± 0.15ab	30.89 ± 2.59e	7.72 ± 0.11cde	4.27 ± 0.22bc	68.04 ± 1.48bcd	65
45	Control	64.30 ± 1.01 g	23.44 ± 0.74k	5.87 ± 0.50 h	2.60 ± 0.21hi	52.82 ± 2.01ij	
	P. indica	67.33 ± .055efg	27.44 ± 1.08 h	7.25 ± 0.22ef	3.07 ± 0.03 fg	60.04 ± 0.79 g	55
	PGPB	72.00 ± 0.38bc	32.00 ± 2.17c	8.36 ± 0.54ab	4.46 ± 0.22 b	68.51 ± 0.57bcd	
	AMF	69.63 ± 0.76cde	31.67 ± 1.25d	7.30 ± 0.06ef	3.99 ± 0.16 cd	66.82 ± 1.97 cd	57
35	Control	55.40 ± 1.13 h	21.50 ± 1.22 L	4.4 ± 0.26 i	2.40 ± 0.29i	50.96 ± 2.88j	
	P. indica	65.10 ± 0.43 fg	23.56 ± 0.78k	6.85 ± 0.52 fg	2.89 ± 0.14fgh	56.22 ± 2.47 h	32
	PGPB	70.78 ± 2.73bcd	30.50 ± 1.89e	8.26 ± 0.07abc	4.01 ± 0.04 cd	65.59 ± 0.55de	
	AMF	68.00 ± 2.01def	28.23 ± 0.15 g	7.16 ± 0.03ef	3.66 ± 0.14de	60.75 ± 1.08 fg	37
Significance							
Drought		***	***	***	**	***	***
P. indica		***	***	***	***	***	***
P. indica×Drought		**	***	*	ns	ns	***
PGPB		***	***	***	***	***	
PGPB×Drought		***	***	*	ns	ns	
AMF		***	***	***	***	***	***
AMF×Drought		**	***	**	ns	ns	***

The presented values represent the mean values obtained from multiple replicates. Letters assigned within a column indicate statistically significant differences at a significance level of $p < 0.05$. The notation "ns" denotes non-significant differences. The symbols "***", "**", and "*" indicate statistical significance levels of $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively

elevated levels of Zn^{2+} (2.42, 2.35, and 2.17-fold increase, respectively), Fe (1.30, 2.96, and 2.23-fold increase, respectively), Mn^{2+} (2.22, 1.21, and 1.16-fold increase, respectively), and Cu (4.67, 4.36, and 1.80-fold increase, respectively) compared to non-inoculated plants. Notably, AMF showed superior efficacy in mitigating drought stress in terms of Fe content when compared to *P. indica* and PGPB. Overall, PGPB-inoculated plants consistently exhibited enhanced levels of Ca^{2+}, Mg^{2+}, K^{+}, Zn^{2+}, Mn^{2+}, and Cu content, along with reduced Na^{+} content in leaves, relative to plants inoculated with AMF and *P. indica.*

3.4 Physiological response

Maize plants inoculated with microbes consistently exhibited higher levels of PN, E, gs, and Ci compared to non-inoculated plants, regardless of drought stress levels. PGPB and AMF inoculation significantly increased PN, with values of 357.03% and 282.44% respectively, under severe drought stress, while *P. indica* inoculation resulted in a 137.37% increase (Fig. 1A). The gs was significantly enhanced in PGPB-inoculated plants (197.08%), followed by AMF (184.58%) and *P. indica* (155%) (Fig. 1B). At 35% field capacity (FC), E levels in PGPB, AMF, and *P. indica*-inoculated plants increased by 243.98%, 174%, and 169.11% respectively (Fig. 1C). PGPB inoculation led to a substantial increase in Ci levels (369.03%) under the highest drought level, while AMF and *P. indica*-inoculated plants showed increases of 328.73% and 203.71% respectively (Fig. 1D). Two-way analysis of variance indicated significant independent effects of PGPB, AMF, *P. indica*, and drought on physiological characteristics (Table 3). PGPB exhibited higher efficacy in mitigating the impact of drought stress when compared to *P. indica* and AMF.

Table 2 Effects of microbial inoculation and drought stress on the macronutrient and micronutrient content in maize plant

FC (%)		Macronutrients (mg kg^{-1})				Micronutrients (mg kg^{-1})			
	Microbe	Ca^{2+}	Mg^{2+}	K^{+}	Na^{+}	Zn^{2+}	Fe	Mn^{2+}	Cu
75	Control	541.06 ± 32.37 g	2734.83 ± 101.62e	2493.63 ± 38.49 h	487.30 ± 13.06j	66.74 ± 4.55 cd	543.80 ± 24.62ef	54.81 ± 3.52e	93.80 ± 4.96de
	P. indica	1055.16 ± 26.47b	5183.37 ± 45.32 a	4945.35 ± 28.39a	483.62 ± 29.27j	107.68 ± 8.43ab	835.40 ± 24.62b	69.42 ± 3.53d	227.80 ± 6.67b
	PGPB	949.67 ± 44.37c	3798.56 ± 68.59b	4495.43 ± 74.64b	374.08 ± 29.22k	97.21 ± 5.40b	847.11 ± 10.76b	88.37 ± 2.16a	334.27 ± 4.96a
	AMF	837.95 ± 29.77d	3154.26 ± 30.58d	3400.94 ± 39.46d	451.37 ± 26.89j	113.04 ± 8.07a	741.52 ± 12.44c	74.89 ± 5.10 cd	95.14 ± 7.11de
55	Control	545.89 ± 45.54 g	2309.10 ± 127.71f	2480.20 ± 2.59 h	659.84 ± 26.89 fg	41.95 ± 8.63f	448.84 ± 17.35 g	53.69 ± 5.39e	66.91 ± 4.06ef
	P. indica	1136.63 ± 33.55a	2851.90 ± 118.47e	4072.05 ± 65.40c	609.88 ± 38.43ghi	76.40 ± 1.94c	825.11 ± 39.20b	69.05 ± 3.04d	127.18 ± 4.96c
	PGPB	731.74 ± 27.94e	3760.80 ± 19.89b	3327.45 ± 82.73d	556.29 ± 26.89i	72.13 ± 11.69c	779.39 ± 15.12c	73.48 ± 4.84 cd	127.58 ± 6.89c
	AMF	629.98 ± 23.24f	2266.02 ± 117.35 h	2657.13 ± 39.08 g	596.15 ± 38.43i	74.12 ± 12.85b	816.30 ± 10.58f	72.08 ± 4.25 cd	90.87 ± 6.89de
45	Control	414.92 ± 21.41 h	673.37 ± 107.88 k	1453.35 ± 29.79k	937.69 ± 25.51b	27.68 ± 7.54 g	327.40 ± 24.62 h	51.23 ± 4.90ef	55.25 ± 8.16 fg
	P. indica	821.83 ± 28.24d	1102.03 ± 56.26i	1760.18 ± 17.63j	671.50 ± 29.22ef	65.48 ± 3.51 cd	747.26 ± 32.20c	54.42 ± 4.67e	93.49 ± 9.95de
	PGPB	708.65 ± 15.85e	3499.70 ± 38.28c	3028.75 ± 26.15f	600.04 ± 41.27hi	56.32 ± 5.67de	756.25 ± 21.00c	67.34 ± 3.19de	103.34 ± 0.48 cd
	AMF	558.33 ± 23.81 g	2162.46 ± 107.78 g	2335.51 ± 27.48i	652.947 ± 46.52fgh	66.03 ± 8.47 cd	914.88 ± 16.77a	68.03 ± 5.32d	99.08 ± 4.52 cd
35	Control	230.84 ± 26.45i	287.56 ± 81.83 L	643.91 ± 2919 m	1030.80 ± 23.41a	23.17 ± 2.66 g	261.46 ± 29.58i	37.78 ± 4.48 g	20.65 ± 9.79 h
	P. indica	537.72 ± 33.10 g	1095.39 ± 86.83i	839.89 ± 36.46 L	861.54 ± 25.51c	50.38 ± 9.17ef	584.10 ± 10.58d	43.76 ± 6.67 fg	37.17 ± 4.04gh
	PGPB	640.11 ± 17.90f	1638.93 ± 47.44 h	1685.34 ± 23.24j	717.81 ± 27.27e	56.13 ± 2.97de	579.32 ± 17.26de	49.30 ± 4.85ef	96.47 ± 4.52de
	AMF	512.31 ± 11.76 g	953.97 ± 32.75j	1431.10 ± 14.38k	801.28 ± 28.93d	54.43 ± 9.54def	772.63 ± 12.44c	45.82 ± 5.13 fg	89.94 ± 4.52de
Significance									
Drought		***	***	***	***	***	***	***	**
P. indica		***	***	***	***	***	***	Ns	***
P. indica×Drought		***	***	***	***	ns	***	***	***
PGPB		***	***	***	***	***	***	***	***
PGPB×Drought		***	***	***	***	ns	***	***	***
AMF		***	***	***	***	***	***	Ns	*
AMF×Drought		***	***	***	***	ns	***	**	***

The presented values represent the mean values obtained from multiple replicates. Letters assigned within a column indicate statistically significant differences at a significance level of $p < 0.05$. The notation "ns" denotes non-significant differences. The symbols "***", "**", and "*" indicate statistical significance levels of $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively

3.5 Chlorophyll and carotenoid content

Figure 2A-D illustrates the impact of microbial inoculation on the photosynthetic pigment contents in maize plants under varying water regimes. Across all levels of water stress, plants inoculated with microbes consistently demonstrated significantly elevated pigment levels in comparison to their non-inoculated counterparts. As water stress intensified, pigment content decreased. Under the highest drought stress level, *P. indica*, AMF, and PGPB-inoculated maize plants showed substantial increases in chlorophyll a content (139.71%, 172.73%, and 190.18% respectively) when compared to plants that had not received microbial inoculation (Fig. 2A). Plants inoculated with PGPB displayed a 60.52% increase in chlorophyll b content, while AMF and *P. indica*-inoculated plants showed increases of 46.39% and 41.18%, respectively (Fig. 2B). PGPB-inoculated plants exhibited the highest total chlorophyll levels (111.19%) at the highest stress level, followed by AMF (95.77%) and *P. indica* (79.68%) (Fig. 2C). At 35% field capacity, carotenoid content was notably higher in PGPB-inoculated plants (58.76%), followed by AMF-inoculated plants (37.15%), and *P. indica*-inoculated plants (28.42%), as illustrated in Fig. 2D. These results imply that among the evaluated microbial agents, PGPB exhibits the highest effectiveness in alleviating drought-induced stress.

3.6 Osmolytes content

The concentration of glycine betaine, total sugars, and trehalose within the leaf tissues of both microbial-inoculated and non-inoculated maize plants exhibited a linear and significant increase with rising water stress levels. AMF-inoculated plants exhibited a 1.9-fold increase in glycine betaine content at 35% FC compared to 75% FC. Plants inoculated with *P. indica* exhibited a 1.6-fold increment, whereas non-inoculated plants displayed a 1.5-fold increase. Plants colonized by *P. indica* and AMF consistently exhibited elevated levels of glycine betaine compared to their non-inoculated counterparts across all levels of drought stress. Interestingly, PGPB inoculation did not affect glycine betaine content and it decreased with increasing drought levels (Fig. 3A). PGPB-inoculated plants had 2.5, 1.6, 1.7, and 1.8-fold higher total sugar content at 75%, 55%, 45%, and 35% FC, respectively, compared to

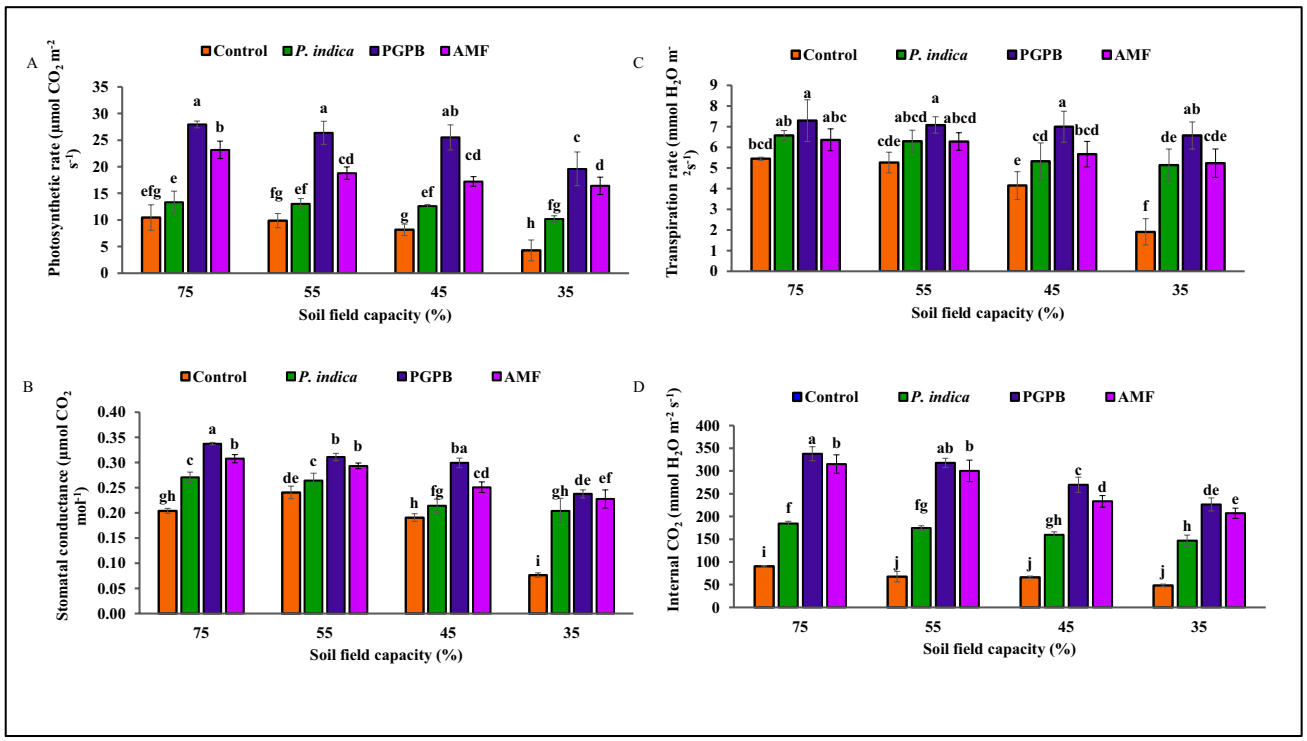


Fig. 1 The effects of microbial inoculation and drought stress on various maize parameters were assessed: **A** photosynthetic rate, **B** stomatal conductance, **C** transpiration rate, and **D** internal CO_2 concentration. The reported values represent the mean measurements derived from multiple replicates. Different letters denote statistically significant differences at a significance level of $P < 0.05$

non-inoculated plants (Fig. 3B). Likewise, plants inoculated with AMF exhibited 2.4, 1.5, 1.5, and 1.4-fold higher total sugar content at the corresponding drought stress levels. Inoculation with *P. indica* led to a 1.3-fold rise in the total sugar content across all FC levels. PGPB-inoculated plants showed trehalose contents 3.0, 1.8, 2.5, and 2.1 times greater than that observed in non-inoculated plants at 75%, 55%, 45%, and 35% FC (Fig. 3C). Similarly, AMF-inoculated plants exhibited trehalose contents 2.5, 1.6, 2.1, and 1.9-fold higher than non-inoculated plants at the corresponding stress levels. *P. indica*-inoculated plants had trehalose contents 2.4, 1.3, 1.6, and 1.3-fold higher than control plants at 75%, 55%, 45%, and 35% FC, respectively. The interaction of PGPB × drought. *P. indica* × drought and AMF × drought were significant on total sugars and trehalose content (Table 3). Across all water regimes, PGPB demonstrated superior efficacy in mitigating drought stress, particularly with regard to total sugars and trehalose, in comparison to AMF and *P. indica*.

3.7 Oxidative damage

Microbial inoculation and water stress exerted significant influences on the LOX activity, relative membrane permeability, and lipid peroxidation of maize plants. The imposition of water stress was found to result in a notable elevation of MDA levels, relative membrane permeability, and LOX activity in both microbial inoculated and control plants. However, plants that underwent microbial inoculation displayed reduced levels of electrolyte leakage, membrane permeability, and LOX activity when contrasted with their non-inoculated counterparts across all levels of water stress. At a FC of 35%, plants inoculated with AMF, *P. indica*, and PGPB exhibited reductions in MDA content by 35.10%, 23.01%, and 41.55%, respectively, in comparison to their non-inoculated counterparts (Fig. 3D). Under severe drought stress, PGPB, AMF, and *P. indica*-inoculated plants exhibited 38.36%, 36.86%, and 28.69% lower relative permeability, respectively (Fig. 3E). Specifically, in conditions of severe drought stress, plants inoculated with AMF, PGPB, and *P. indica* exhibited reductions of 16.14%, 31.23%, and 9.82% in LOX activity, respectively, when compared to control plants (Fig. 3F). The two-way analysis of variance substantiated the individual impacts of AMF, PGPB, *P. indica*, and drought stress, in addition to the interactive effects of PGPB×drought and *P. indica*×drought, on plasma membrane integrity in maize plants (Table 3). Consequently, it can be inferred that PGPB displays superior effectiveness in mitigating drought-induced membrane damage in maize plants,

Table 3 The impact of drought, AMF, PGPB, *P. indica*, drought × AMF, drought × PGPB, drought × *P. indica* on photosynthetic pigments, osmolytes contents, physiological processes, protein levels, membrane damage, and antioxidants in maize plants

Parameters	Drought	*P. indica*	Drought × *P. indica*	PGPB	Drought × PGPB	AMF	Drought × AMF
Photosynthetic rate	***	***	ns	***	ns	***	ns
Transpiration	***	***	**	***	**	***	**
Stomatal conductance	***	***	***	***	***	***	***
Internal CO_2	***	***	ns	***	***	***	***
Chlorophyll a	***	***	***	***	***	***	**
Chlorophyll b	***	***	ns	***	ns	***	ns
Total chlorophyll	***	***	***	***	***	***	***
Carotenoid	ns	***	ns	***	ns	***	ns
Protein	ns	***	ns	***	ns	***	ns
α-tocopherol	***	***	ns	***	ns	***	ns
Lipid peroxidation	***	***	ns	***	***	***	**
Relative permeability	***	***	**	***	***	***	***
LOX	***	***	ns	***	**	***	**
Trehalose	***	***	***	***	***	***	***
Total sugars	***	***	ns	ns	***	***	**
Proline	***	***	**	***	ns	***	ns
Glycine betaine	***	*	ns	***	***	**	ns
Putrescine	***	***	***	***	***	***	***
Spermidine	***	***	***	***	***	***	***
SOD	***	***	ns	***	ns	***	ns
CAT	***	***	ns	***	***	***	***
APX	***	***	ns	***	***	***	**

Significant interactions at $P < 0.05$. ns: not significant. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$. PGPB, plant growth-promoting bacteria; AMF, arbuscular mycorrhiza fungi; APX, ascorbate peroxidase, SOD, superoxide dismutase; LOX, lipoxygenase; CAT, catalase

surpassing the performance of both *P. indica* and AMF across various water regimes.

3.8 Protein content

Drought stress elicited a substantial increase in protein content in both maize plants, irrespective of microbial inoculation. However, under drought stress conditions, plants that received microbial inoculation exhibited markedly higher protein levels compared to their non-inoculated counterparts (Fig. 4D). Specifically, in comparison to non-inoculated maize plants, those inoculated with PGPB, AMF, and *P. indica* displayed increased protein content levels by 139.80%, 119.13%, and 86.79%, respectively, at an FC of 35%. Hence, it can be concluded that PGPB demonstrates superior efficacy in mitigating drought stress across all drought stress levels in comparison to AMF and *P. indica*-inoculated plants.

3.9 Antioxidant enzymes activity

In both water-limited and adequately watered treatments, the existence of microbial colonization demonstrated a beneficial impact on the activity of SOD in maize plants (Fig. 4A). Exposure to water stress led to an elevation in SOD activity in both maize plants with microbial inoculation and those without. Notably, under severe drought conditions (35% FC), maize plants inoculated with PGPB exhibited a 22.30% increase in SOD activity, followed by AMF-inoculated plants with a 19.36% increase, and *P. indica*-inoculated plants with a 12.47% increase when contrasted with plants that did not receive microbial inoculation. Furthermore, the response of maize plants to increasing water stress levels varied concerning CAT activity among the different microbial inoculations. Plants inoculated with *P. indica* and PGPB exhibited an enhancement in CAT activity as water stress levels increased. In contrast, plants inoculated with AMF displayed a decrease in CAT activity when compared to non-inoculated plants. More specifically, when the soil moisture level reached 35% of field capacity, plants inoculated with PGPB and *P. indica* exhibited increases in CAT activity by 54.13% and 24.21% respectively, when compared to plants that were not subjected to microbial inoculation (Fig. 4B). In terms of APX activity, plants subjected to microbial inoculation consistently exhibited elevated APX activity when compared to their non-inoculated counterparts, irrespective of the severity of drought stress levels. Across all levels of water stress, maize plants inoculated with PGPB

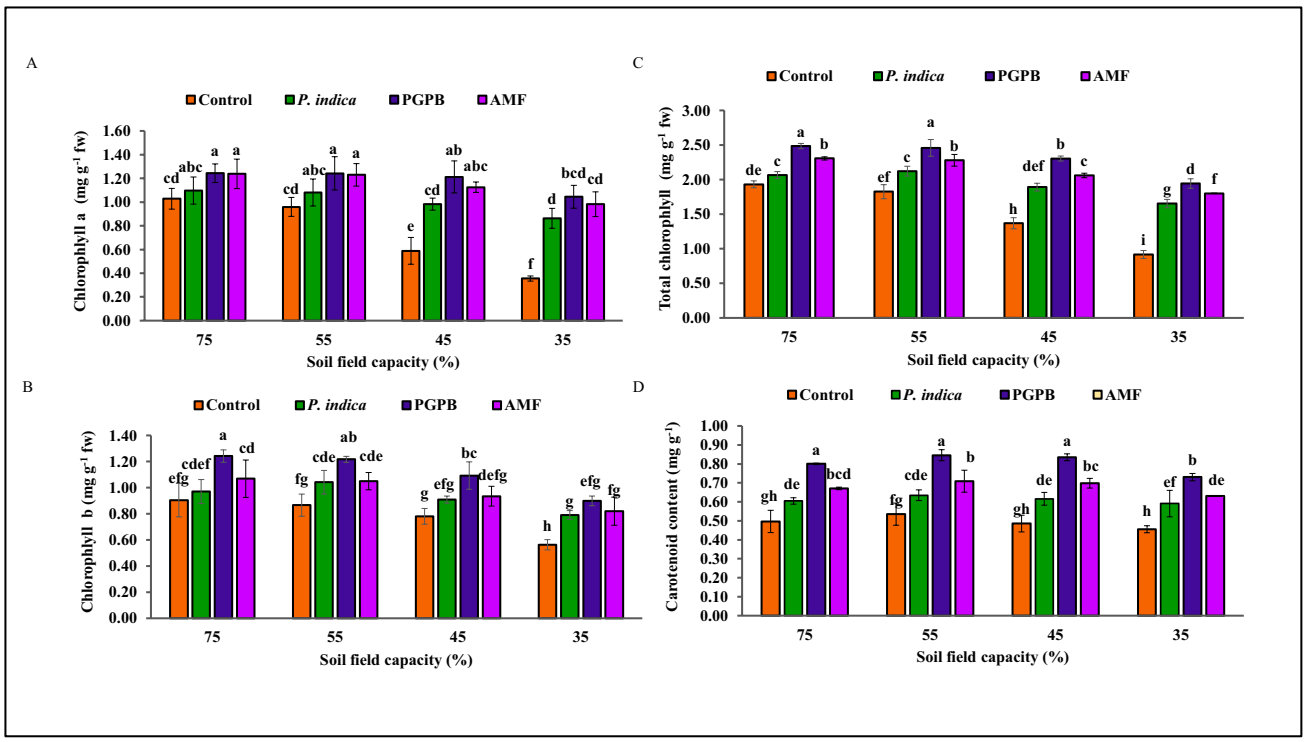


Fig. 2 The impacts of drought and microbial inoculation on various parameters in maize: **A** chlorophyll a, **B** chlorophyll b, **C** total chlorophyll, and **D** carotenoids. The values presented are the average measurements obtained from multiple replicates. Distinct alphabetic labels denote statistically significant variations at a significance level of $P < 0.05$

demonstrated notably higher APX activity in comparison to those inoculated with AMF and *P. indica*. Specifically, at FC of 75%, 55%, 45%, and 35%, PGPB-inoculated plants displayed APX activities that were 1.34, 1.23, 1.43, and 1.43-fold greater, respectively, than non-inoculated plants, while AMF-inoculated plants exhibited APX activities that were 1.24, 1.13, 1.37, and 1.28-fold higher, respectively, than non-inoculated plants (Fig. 4C). In all instances of water stress levels, maize plants inoculated with *P. indica* exhibited APX activity 1.1-fold greater than that of non-inoculated plants.

3.10 Non-enzymatic antioxidant molecules

Drought stress elicited a substantial augmentation in the levels of α-tocopherol, proline, and polyamines in maize plants, regardless of microbial inoculation. However, under drought conditions, inoculated plants exhibited significantly higher levels of antioxidants compared to control plants. PGPB-inoculated plants showed a 50.73% increase in α-tocopherol content at 35% FC, followed by AMF-inoculated plants with a 35.48% increase, and *P. indica*-inoculated plants with an 18.16% increase (Fig. 5A). PGPB-inoculated plants also had a 49.80% higher proline level, while AMF and *P. indica*-inoculated plants displayed increases of 36.73% and 32.34%, respectively (Fig. 5B). Polyamine content, particularly putrescine, and spermidine was influenced by drought stress and microbial inoculation. Nonetheless, it is noteworthy that spermine remained undetectable in both microbial-inoculated and non-inoculated plants across all levels of drought stress. At an FC of 35%, a notable increase in putrescine content was observed in PGPB-inoculated plants (3.1-fold), AMF-inoculated plants (2.9-fold), *P. indica*-inoculated plants (5.0-fold), and non-inoculated plants (2.7-fold) compared to well-watered plants (Fig. 5C). Additionally, at the same FC level, AMF-inoculated plants (1.3-fold), PGPB-inoculated plants (1.3-fold), and *P. indica*-inoculated plants (1.1-fold) displayed respective increments in spermidine content in comparison to non-inoculated plants (Fig. 5D).

The outcomes of a two-way analysis of variance presented in Table 3 reveal significant independent effects of AMF, *P. indica*, PGPB, and drought stress on polyamine and proline content. Furthermore, it is notable that AMF outperformed PGPB and *P. indica* in terms of polyamine content. In contrast, PGPB demonstrated superior efficacy in alleviating drought stress with respect to α-tocopherol and proline content compared to AMF and *P. indica* across all experimental conditions.

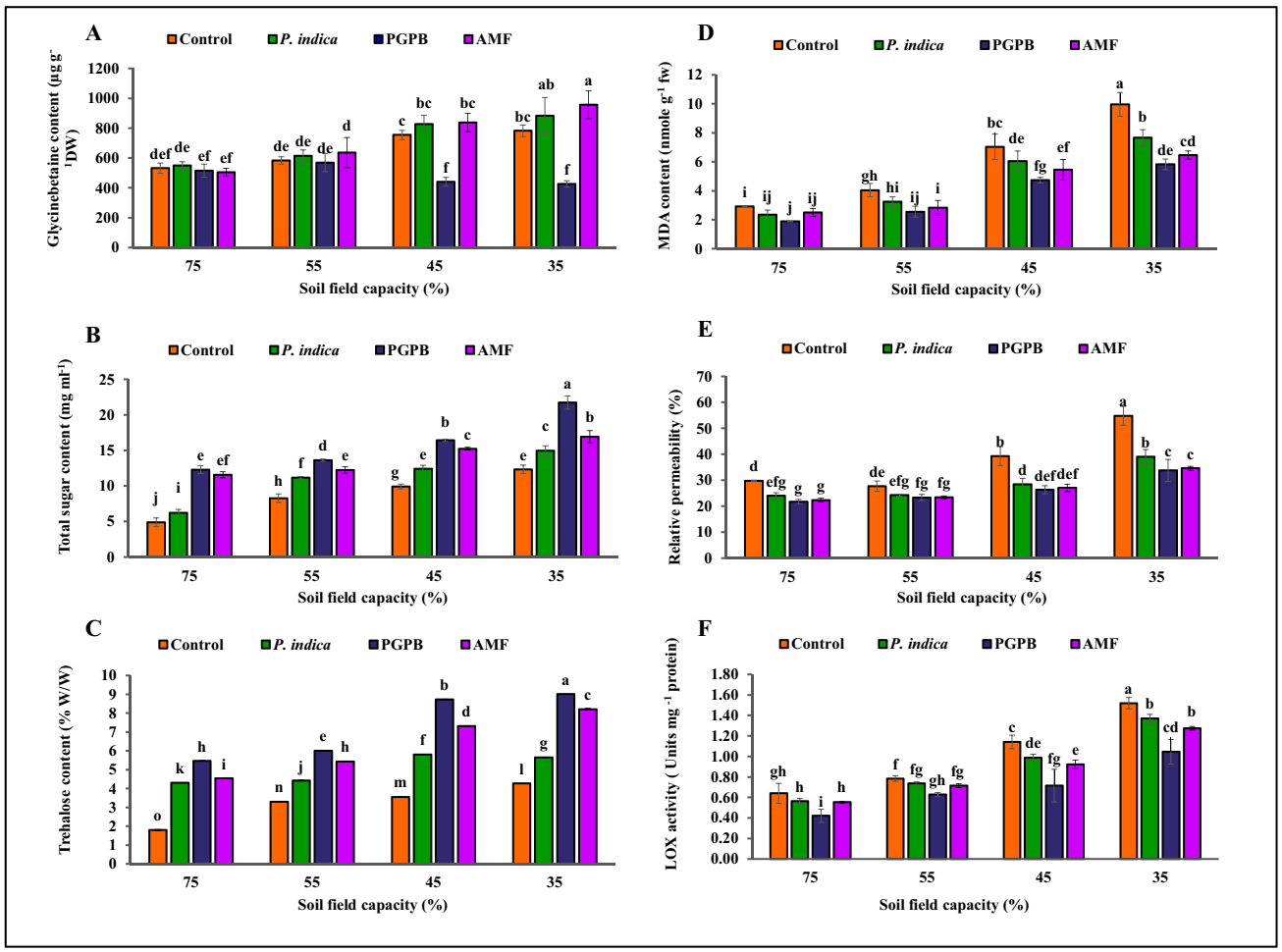


Fig. 3 The impact of microbial inoculation and drought in maize on the content of **A** glycine betaine, **B** total sugars, **C** trehalose **D** MDA content, **E** relative permeability, and **F** LOX activity. The values presented are the average measurements obtained from multiple replicates. Distinct letters signify statistically significant distinctions at a significance level of $P < 0.05$

3.11 Cell ultrastructural studies

At various water regimes, qualitative distinctions between microbial-inoculated and non-inoculated maize plants were evident in the electron micrographs.

3.11.1 Chloroplast

The response of chloroplast ultrastructure to water stress was examined in both microbial-inoculated and non-inoculated maize plants (Figs. 6 and 7). Under well-watered conditions (75% FC), all plants exhibited chloroplasts with well-developed grana and thylakoids enclosed by double-layered membranes. Plants inoculated with PGPB exhibited a significantly greater prevalence of grana and thylakoids when contrasted with plants inoculated with *P. indica* and AMF. Mild drought stress (55% FC) resulted in diminished grana and thylakoids, and contraction of the thylakoid membrane, leading to a decrease in the thylakoid lumen accompanied by an increase in plastoglobules in non-inoculated plants. The chloroplast structures were largely similar between microbial-inoculated plants at 55% FC and 75% FC. At 45% of field capacity, reduced soil moisture caused swelling of the thylakoid lumen, disintegration of grana, and accumulation of smaller plastoglobules in non-inoculated plants. Plants subjected to inoculation retained well-preserved thylakoid and grana structures, characterized by a reduced abundance of plastoglobules, albeit larger in size. PGPB-inoculated plants showed better preservation of thylakoids and grana compared to *P. indica* and AMF. Severe drought conditions (35% FC) resulted in disorganized thylakoids and grana in non-inoculated plants, while inoculated plants displayed preserved thylakoids, disarrayed grana, and a reduced number of plastoglobules that were larger in size. Under severe drought conditions, PGPB-inoculated plants exhibited higher preservation of thylakoids and larger plastoglobules compared to *P. indica* and AMF.

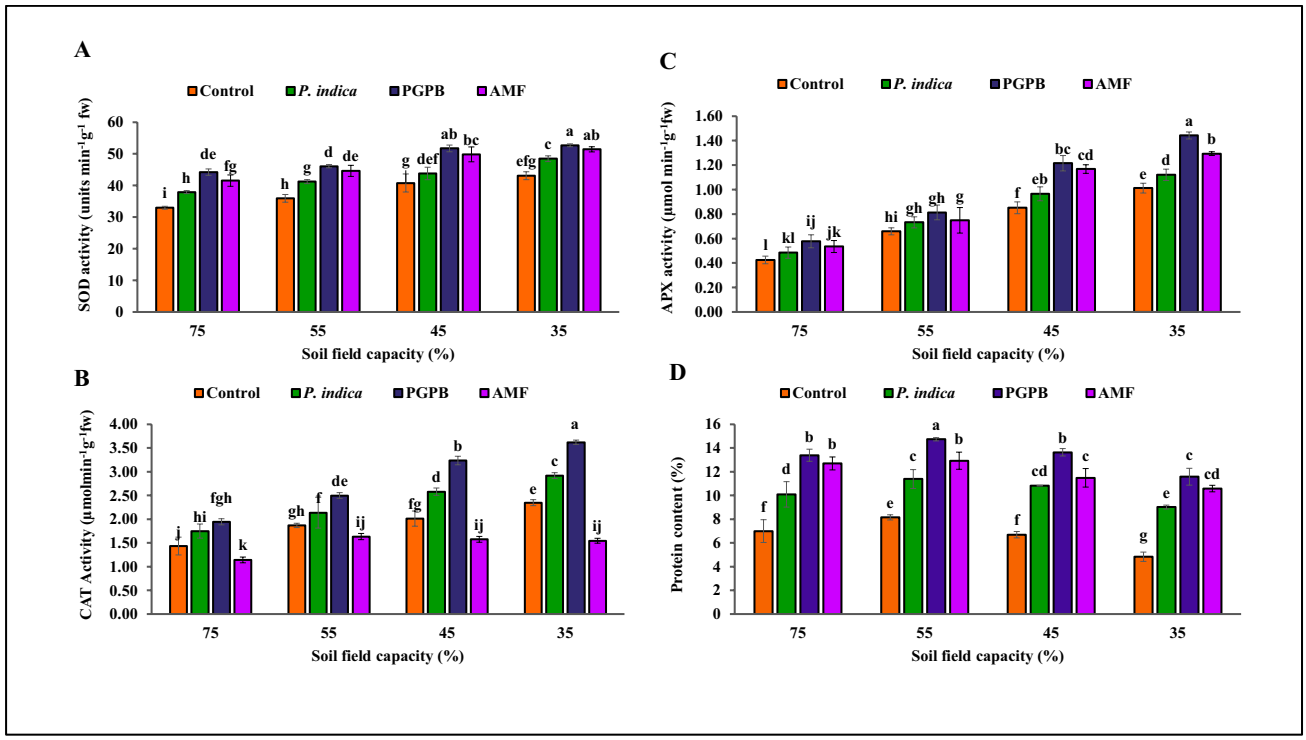


Fig. 4 The impact of microbial inoculation and drought in maize on the activity of **A** SOD, **B** CAT, **C** APX, and **D** protein content. The values presented are the average measurements obtained from multiple replicates. Distinct alphabetical symbols represent statistically significant distinctions at a significance level of $P < 0.05$

3.11.2 Plasma membrane

Under adequately watered conditions (75% FC), both microbial-inoculated and non-inoculated maize plants demonstrated close apposition of the plasma membrane to the cell wall (Figs. 8 and 9). However, the imposition of water stress resulted in the detachment of the plasma membrane from the cell wall, consequently generating an apoplastic space. As the intensity of drought stress heightened, a marked increase in the intercellular gap between the plasma membrane and the cell wall became apparent. Notably, microbial-inoculated plants displayed a reduced extent of plasma membrane detachment, leading to a narrower cell wall-plasma membrane gap. Across all levels of water stress, PGPB-inoculated plants exhibited a notably diminished degree of separation between the plasma membrane and the cell wall compared to plants inoculated with *P. indica* and AMF. Moreover, under elevated levels of drought stress, the plasma membrane underwent invaginations, leading to the formation of cytoplasmic vesicles. Notably, non-inoculated plants exhibited fewer vesicles compared to their microbial-inoculated counterparts.

4 Discussion

We observed a decline in AMF and *P. indica* colonization within maize roots as water stress intensified, consistent with prior research indicating drought hinders hyphal development and reduces mycorrhizal colonization (Ostadi et al. 2022). Notably, severe drought stress markedly decreased the colonization of both AMF and *P. indica*, exerting adverse effects on biomass levels.

In our investigation, both microbial-inoculated and non-inoculated maize plants showed growth suppression under drought stress, mainly due to the negative effects of drought on plant nutrient uptake and photosynthetic efficiency (Ullah and Farooq 2022). However, our study revealed that maize plants, when inoculated with *P. indica*, PGPB, and AMF, exhibited increased growth in shoots, roots, and overall biomass compared to non-inoculated plants. Similar findings have been documented in various species, such as *Camellia sinensis* (Liu et al. 2023), *Solanum lycopersicum* (Alam et al. 2023), and *Poncirus trifoliata* (Cao et al. 2023), under abiotic stress conditions.

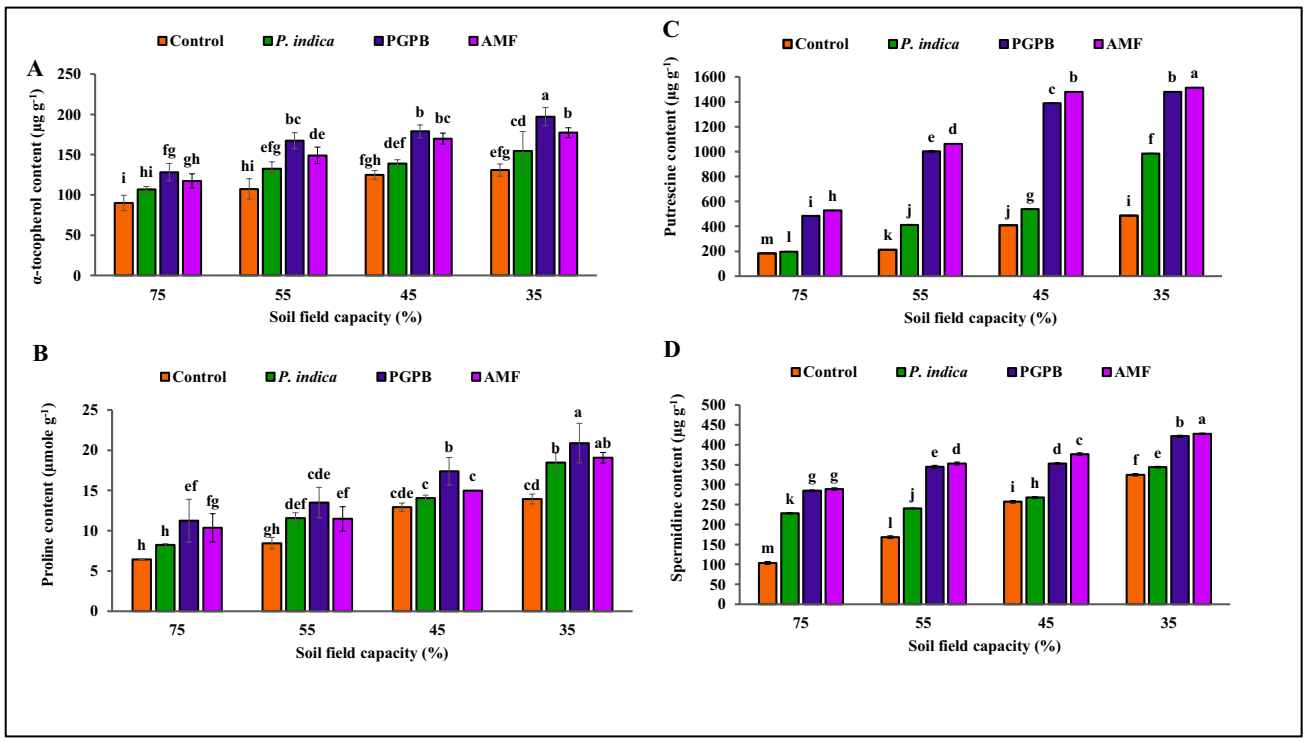


Fig. 5 The impacts of microbial inoculation and drought in maize on **A** α-tocopherol, **B** Proline, **C** Putrescine, and **D** Spermidine. The values presented are the average measurements obtained from multiple replicates. Distinct alphabetic symbols denote statistically significant variations at a significance level of $P < 0.05$

Furthermore, our study revealed that maize plants experiencing water stress, along with microbial inoculation, exhibited increased leaf surface area compared to non-inoculated plants. Similar results were reported by Oliveira et al. (2022) in *Glycine max*, where mycorrhizal inoculation also resulted in enhanced leaf area in comparison to non-inoculated plants. Augmented nutrient uptake (Rajput et al. 2022), osmolyte content, enzymatic-antioxidant activities, and reduced membrane permeability collectively contribute to the higher growth and drought resilience observed in microbial-inoculated plants (Tyagi et al. 2023). Importantly, in the present study, PGPB-inoculated maize plants demonstrated elevated drought resistance compared to those receiving *P. indica* and mycorrhizal treatments.

Our study found that as drought severity increased, both microbial-inoculated and non-inoculated maize plants experienced reduced nutrient contents due to limited water and nutrient availability and diminished root development. Nevertheless, microbial inoculation consistently led to improved nutrient acquisition compared to non-inoculated plants under all levels of drought-induced stress. This enhanced nutrient status in microbial-inoculated plants is facilitated by extraradical hyphae provided by *P. indica* and mycorrhiza, which increase the absorption surface (Ortiz et al. 2015). Calcium, essential for the structural integrity of membranes and cell walls as well as ion transport (de Bang et al. 2021), is depleted under water stress as Na^+ replaces Ca^{2+} in cell walls and plasma membranes. This substitution adversely affects cell turgidity and hydraulic conductivity, resulting in reduced water flow within the plant. Our study demonstrated that microbial-inoculated maize plants showed enhanced absorption of Ca^{2+} ions, consistent with findings by Devarajan et al. (2021).

K^+ is indispensable for preserving osmotic equilibrium, and membrane potential (Sardans and Peñuelas 2021), facilitating protein and sugar synthesis, and modulating enzymatic activity within plants (Hasanuzzaman et al. 2018). In this study, both microbial-inoculated and non-inoculated maize plants exhibited reduced K^+ levels under drought conditions, attributed to the competition for binding sites between Na^+ and K^+ inhibiting K^+-dependent metabolic processes. However, our study showed that maize plants inoculated with microbes showed higher K^+ uptake. Our findings suggest that increased K^+ accumulation in drought-stressed, microbial-inoculated plants enhances osmolyte levels, protein synthesis, and enzymatic antioxidant activities, supporting Hasanuzzaman et al. (2018).

Mg^{2+} is crucial for chlorophyll synthesis (Tang et al. 2023) and protein production in plants. Drought stress

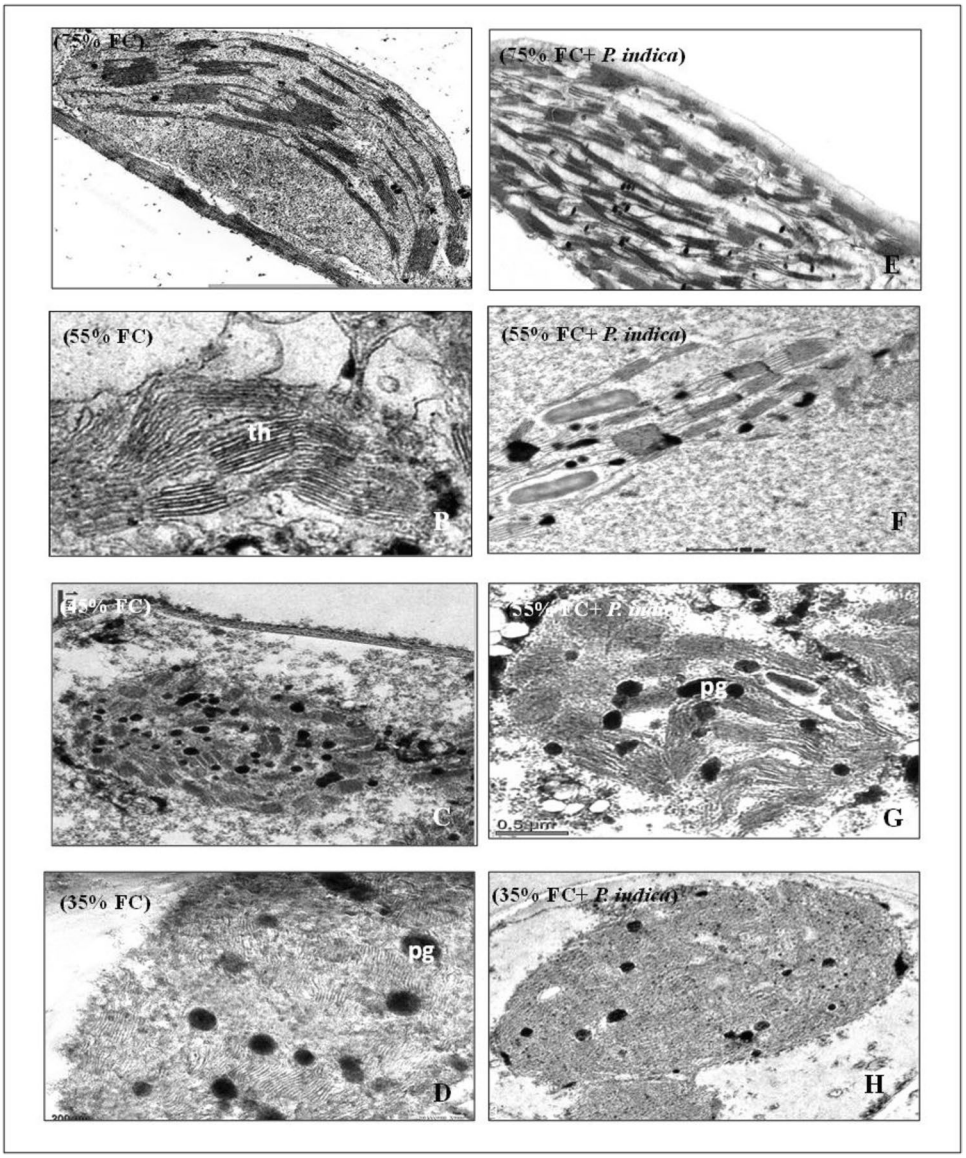


Fig. 6 Transmission electron micrographs of chloroplasts in mesophyll cells of maize plants under different drought conditions and *P. indica* inoculation **A** Control maize plant at 75% FC (non-drought condition); Well-defined double-layered membrane, distinct thylakoids, and grana. **B** Control maize plant at 55% FC; Reduced number of grana and thylakoids presence of plastoglobules. **C** Control maize plant at 45% FC; Disorganized grana and thylakoids, increased accumulation of plastoglobules. **D** Control maize plant at 35% FC; Disorganized grana and thylakoids, significant accumulation of plastoglobules. **E** *P. indica*-inoculated maize plant at 75% FC; properly developed thylakoids with properly stacked grana, and an intact membrane. **F** *P. indica*-inoculated maize plant at 55% FC; More grana observed, fewer plastoglobules. **G** *P. indica*-inoculated maize plant at 45% FC; Maintenance of thylakoid and grana structure, fewer plastoglobules. **H** *P. indica*-inoculated maize plant at 35% FC; Thylakoids intact, prominent plastoglobules, minor grana disintegration. Values in brackets indicate the level of imposed drought stress on plants

reduces chlorophyll and Mg^{2+} levels, impacting photosynthesis. However, *P. indica* and other plant growth-promoting microbes can enhance these parameters (Inbaraj 2021). In our study, drought and microbial inoculation significantly affected maize Mg^{2+} levels. Reduced Ca^{2+} availability under drought may compete with Mg^{2+} (Pandey et al. 2021), influencing membrane binding. In our study, microbial-inoculated maize showed higher chlorophyll, improved photosynthesis, and increased Mg^{2+} content under drought stress. These findings suggest a potential association between increased Mg^{2+} and enhanced photosynthesis in microbial-inoculated maize plants.

Microbial-inoculated plants exhibited lower levels of Na^{+} than control plants, with AMF excluding Na^{+} uptake and transfer within the plant and PGPB reducing Na^{+} absorption through exopolysaccharide binding. Simultaneously, PGPB enhances nutrient acquisition (Bhagat et al. 2021). Mycorrhiza retains Na^{+} within hyphae or sequesters it in root cell vacuoles, reducing Na^{+} content in roots and diluting ion concentration in plant tissues, promoting plant growth and physiological functions.

Micronutrients, including Zn^{2+}, Fe, Cu, and Mn^{2+}, hold pivotal significance in diverse physiological mechanisms within plants. Zn^{2+} is essential for protein synthesis, chlorophyll formation, and enzyme activity; Fe is vital for chlorophyll synthesis and chloroplast functioning (Kumar et al. 2021). Cu is involved in respiration, redox reactions, photosynthesis, and cell wall structure (Kumar 2015). Mn^{2+} contributes to the water-splitting system in photosynthesis. In our investigation, in the presence of drought conditions, the levels of these micronutrients declined in both maize plants with microbial inoculation and those without, primarily

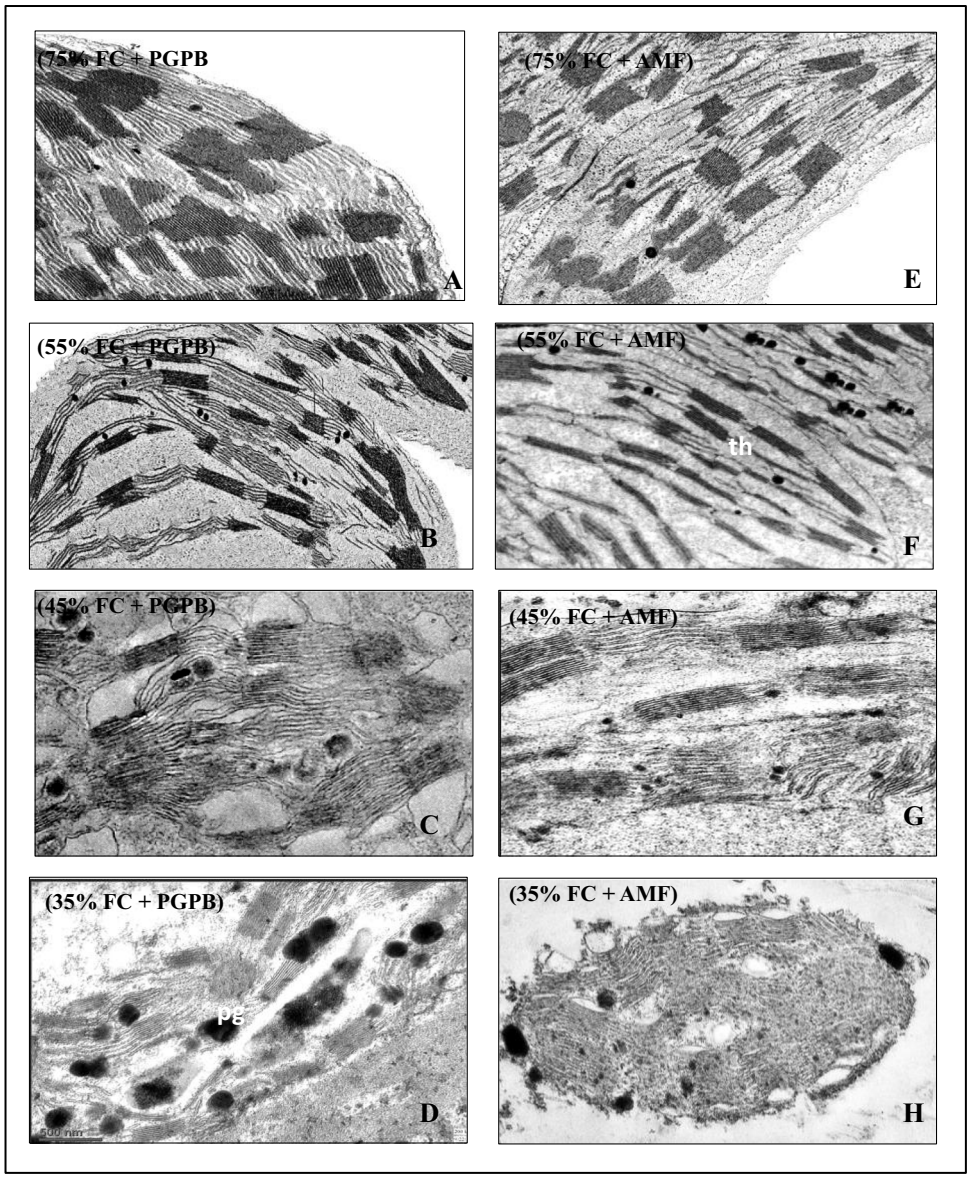


Fig. 7 Transmission electron micrographs of chloroplasts in mesophyll cells of maize plants inoculated with PGPB and AMF under varying drought conditions **A** PGPB-inoculated maize plant at 75% FC (non-drought condition); Prominent thylakoids and robustly formed grana enclosed within a membrane. **B** PGPB-inoculated maize plant at 55% FC; Intact grana and thylakoids with reduced plastoglobules. **C** PGPB-inoculated maize plant at 45% FC; Maintained grana organization and increased plastoglobule presence. **D** PGPB-inoculated maize plant at 35% FC; Accumulation of large plastoglobules while preserving the structure of grana and thylakoids. **E** Maize plant inoculated with AMF at 75% FC: Well-stacked grana in the stroma. **F** AMF-inoculated maize plant at 55% FC; Decreased thylakoids and plastoglobules. **G** AMF-inoculated maize plant at 45% FC; Intact thylakoids and enhanced plastoglobule accumulation within the chloroplast. **H** Maize plant inoculated with AMF at 35% FC: Large plastoglobule buildup, thylakoids intact, grana disorganized. Values in parentheses represent drought stress levels

attributed to decreased solubility and mobility within soil solutions. As a result, a depletion zone forms around maize plant roots, limiting the access of plants to these nutrients. However, microbial inoculation consistently enhanced the acquisition of micronutrients compared to non-inoculated plants, potentially through improved hyphal growth. These observations align with prior research and indicate a favorable association between microbial inoculation, elevated micronutrient content, chlorophyll levels, and photosynthetic performance in maize plants subjected to drought-induced stress.

Photosynthesis, a pivotal growth determinant in plants, reflects their drought resilience (Liang et al. 2019). Drought conditions lead to stomatal closure, which reduces gas exchange, internal CO_2 levels, photosynthetic pigments, and photosynthetic efficiency (Liang et al. 2019). Our study supports these effects by showing decreased transpiration, nutrient uptake, and water absorption in maize plants experiencing drought stress. In contrast, microbial-inoculated maize exhibited elevated photosynthetic parameters (PN, gs, Ci, and E) during drought, consistent with prior research demonstrating the enhancement of photosynthetic efficiency by PGPB, including increased Ci, gs, and E (Vishnupradeep et al. 2022). Additionally, our work showed enhanced carotenoids concentration subsequent to microbial inoculation. Carotenoids exhibit antioxidant properties that protect the photosynthetic apparatus from free radical damage, inhibit the peroxidation of lipids, stabilize membranes, and contribute significantly to the formation of light-harvesting complexes (Farooq et al. 2009). Water stress and microbial treatment jointly enhanced carotenoid content, which

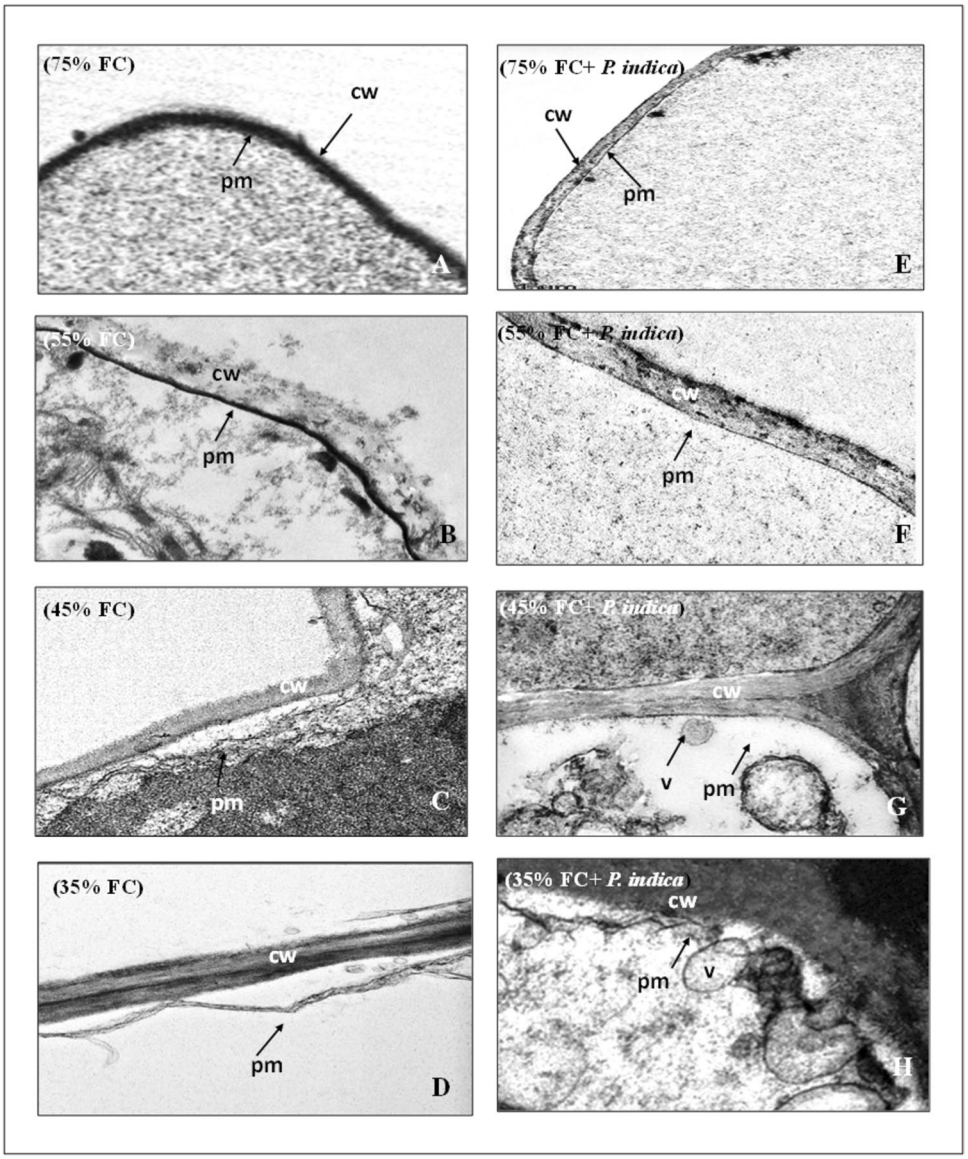


Fig. 8 Transmission electron micrographs of the plasma membrane and cell wall and in mesophyll cells of maize plants under different drought conditions and *P. indica* inoculation (**A**) Control maize plants at 75% FC (non-drought condition); Plasma membrane tightly adhered to the cell wall. **B** Control maize plants at 55% FC stress; Receding plasma membrane. **C** Control maize plants at 45% FC stress; Expanded gap between plasma membrane-cell wall, vesicle formation. **D** Control maize plants at 35% FC stress; Marked expansion of apoplastic space between cell wall-plasma membrane. **E** Maize plants inoculated with *P. indica* at 75% FC (non-drought); Intact plasma membrane tightly bound to the cell wall **F** *P. indica*-inoculated maize at 55% FC; Plasma membrane close to the cell wall. **G** *P. indica*-inoculated maize at 45% FC; Early plasmolysis. **H** *P. indica*-inoculated maize at 35% FC; Slight plasma membrane-cell wall gap increase. Values in parentheses denote water regime levels. v = vesicle; pm = plasma membrane; cw = cell wall

protects the photosynthetic apparatus from oxidative damage. In our investigation, the enhanced photosynthetic performance in microbial-inoculated plants can be attributed to an elevated chlorophyll content and an increased leaf surface area.

Chlorophylls and carotenoids, essential photosynthetic pigments in maize, exhibit a reduction in abundance during drought stress conditions (Khazaei et al. 2020), which is consistent with our research results. However, *P. indica*, PGPB, and mycorrhizal-inoculated maize displayed heightened pigment levels, suggesting a mitigating effect of microbial inoculation on water stress. This echoes earlier research, e.g., arbuscular mycorrhizal fungi augmenting chlorophylls in *Arizona* plants (Aalipour et al. 2023), and PGPB elevating *Zea mays* photosynthetic pigments (Vishnupradeep et al. 2022). In our study, microbial-induced antioxidants possibly fostered pigment accumulation, safeguarding chloroplasts and enhancing nutrient availability, including magnesium, in line with Khan et al. (2019).

During drought stress, both microbial-inoculated and non-inoculated maize showed increased osmolyte accumulation, with microbial-inoculated plants displaying even higher levels. Our results align with Hashem et al. findings (2016), highlighting that microbial-inoculated maize exhibits elevated osmolyte levels in response to drought stress. Kour et al. (2020) also corroborates this trend, illustrating drought alleviation in *Setaria italica* by *Penicillium sp*, *Calcoaceticus*, and *Acinetobacter*. This was achieved through the promotion of carbohydrate, glycine betaine, and proline accumulation, maintaining cell turgidity and membranes. In our study, we observed a decrease in glycine betaine content in PGPB-inoculated plants, and to date, there has

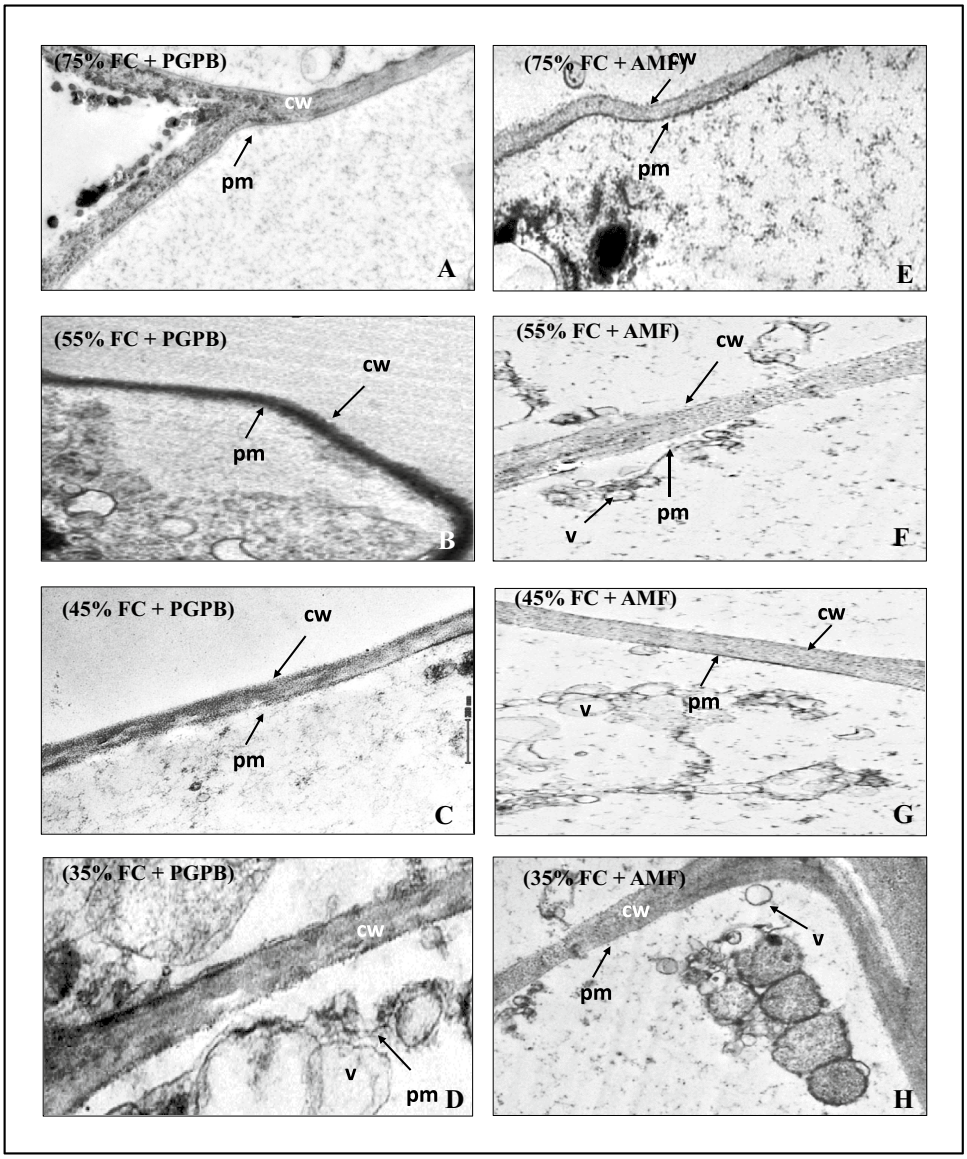


Fig. 9 Transmission electron micrographs of the cell wall and plasma membrane in mesophyll cells of maize plants inoculated with PGPB and AMF under varying drought conditions **A** PGPB-inoculated maize at 75% FC; Plasma membrane tightly bound to the cell wall. **B** PGPB-inoculated maize at 55% FC; Plasma membrane closely bound to the cell wall. **C** PGPB-inoculated maize at 45% FC; Intact plasma membrane. **D** PGPB-inoculated maize at 35% FC; Plasma membrane intact. **E** AMF-inoculated maize at 75% FC; intact plasma membrane near cell wall. **F** AMF-inoculated maize at 55% FC; Intact plasma membrane near cell wall. **G** AMF-inoculated maize at 45% FC; Slight plasma membrane shrinkage, and vesicle formation evident. **H** AMF-inoculated maize plants at 35% FC stress; Plasma membrane slightly detached from the cell wall. Values in parentheses indicate the level of water regimes imposed on plants. pm = plasma membrane; cw = cell wall; v = vesicle

been no reported evidence of such an occurrence in microbe-inoculated plants under stress conditions. Our investigation highlights the improved osmotic adjustment observed in maize plants inoculated with microorganisms when compared to their non-inoculated counterparts under water stress conditions.

Microbial-inoculated maize plants exhibited higher protein content compared to non-inoculated plants, which can be attributed to enhanced photosynthesis, increased chlorophyll levels, improved membrane integrity, and elevated osmolyte content. This discovery aligns with the research conducted by Boutasknit et al. (2020), which provided evidence of drought stress alleviation in carob plants through colonization by AMF, leading to elevated protein levels.

MDA, an outcome of membrane lipid peroxidation, serves as a pivotal physiological drought stress indicator in plants (Yasmeen et al. 2022). Drought-induced lipid peroxidation impacts the ultrastructural characteristics and permeability of cell membranes in maize. The enzyme LOX catalyzes lipid-to-hydroperoxyl fatty acid conversion, producing destructive free radicals. Averting lipid peroxidation and sustaining membrane permeability hold significance for drought resilience (Kour et al. 2020).

Microbial-inoculated maize displayed diminished electrolyte leakage and lipid peroxidation, in accordance with Begum et al. (2022). Surprisingly, LOX activity decreased in microbial-inoculated maize, contrary to a study where AMF and PGPB co-inoculation elevated LOX activity in *Nicotiana tabacum* (Begum et al. 2022). In our study, microbial-inoculated maize exhibited diminished membrane permeability and malondialdehyde levels, attributed to regulated adaptive responses involving α-tocopherol,

proline, carotenoids, and enzymatic antioxidants (APX, SOD, CAT). Consistent with prior research, PGPB-inoculated plants displayed elevated tocopherol, proline, and carotenoid levels, enhancing drought stress tolerance (Timmusk et al. 2014). Additionally, PGPB-inoculated plants showed increased APX, SOD, and CAT activity (Chiappero et al. 2019), resulting in curbed reactive oxygen species accumulation.

Our study unveiled distinctive antioxidant enzyme activity disparities between non-inoculated and inoculated maize, paralleling prior work showcasing increased growth and antioxidant enzyme activity through microbial inoculation under water stress (Tyagi et al. 2023). Microbial inoculation curbs membrane permeability and lipid peroxidation by amplifying enzymatic antioxidants (Tyagi et al. 2023). Analogously, *Zea mays* treated with microbial consortia exhibited augmented antioxidant enzyme activity amid abiotic stress (Abbas et al. 2020). Noteworthy, in our study, AMF-inoculated maize demonstrated reduced CAT activity as drought intensified, aligned with earlier findings of CAT reduction in microbial-inoculated drought-stressed maize (Siddique et al. 2022). Our results emphasize *P. indica*, AMF, and PGPB potential as bio-inoculants for bolstering antioxidant defenses and mitigating drought stress in maize.

This investigation demonstrates that the introduction of mycorrhiza, PGPB, and *P. indica* to maize plants results in increased levels of non-enzymatic antioxidants, including α-tocopherol, proline, and polyamines. These findings highlight the capacity of α-tocopherol to effectively scavenge free radicals and regulate lipid peroxidation under abiotic stress, in line with Kim et al. (2019) work. Fresno et al. (2023) research further supports this, revealing increased α-tocopherol levels due to AMF colonization in *Trifolium repens*, mitigating drought stress. Similarly, Liu et al. (2022) evidenced *Rhizobium* association with Chinese milkvetch augmenting drought tolerance via promoting putrescine, spermidine, and spermine. Consistent with Hashem et al. (2016), we observed elevated proline in drought-stressed maize treated with a microbial consortium. Overall, our study underscores that introducing mycorrhiza, plant growth-promoting bacteria, and *P. indica* to maize enhances water stress resistance through increased non-enzymatic antioxidants, fostering optimal photosynthetic efficiency, chlorophyll content, and reduced ion leakage.

The chloroplast, known for its susceptibility to the effects of water stress (Gunasekera and Ratnasekera 2023), exhibited modified ultrastructural features in response to drought stress in both microbial-inoculated and non-inoculated maize plants within our investigation. Mild drought stress induced thylakoid membrane contraction and reduced lumen volume, possibly due to osmotic collapse. As drought severity increased, thylakoid lumen swelling and grana destacking occurred. Several studies have provided evidence that osmotic stress can lead to membrane damage and thylakoid swelling (Aldesuquy et al. 2018).

However, in our study, microbial-inoculated maize showed less chloroplast damage compared to non-inoculated counterparts, potentially attributed to increased osmolyte accumulation (sugars, polyamines, glycine betaine) that may mitigate osmotic collapse. Polyamines binding to thylakoid sites may inhibit membrane degradation (Ali et al. 2020). This observation may be attributed to the existence of enlarged plastoglobules, potentially functioning as a protective mechanism against oxidative damage. During water stress, plastoglobules, which are lipid-protein bodies found in chloroplasts, have been observed to increase in size, as reported by Zhu et al. (2021). These expanded plastoglobules function as sites for α-tocopherol synthesis, particularly in situations characterized by oxidative stress, (Brehelin et al. 2007). In our investigation, microbial-inoculated maize displayed elevated α-tocopherol levels, guarding membrane lipids and photosystem II against oxidation. Plastoglobule accumulation correlated with higher polyamine levels (Shu et al. 2015). Overall, microbial-inoculated maize revealed preserved chloroplast ultrastructure, attributed to increased osmolyte levels and protective plastoglobules. Notably, PGPB outperformed mycorrhiza and *P. indica* in sustaining chloroplast integrity under drought across stress levels.

Cell membrane stability is a recognized drought tolerance indicator in plants (Abobatta 2019). Our investigation showed plasma membrane disintegration in maize cells under drought, with increasing severity expanding the membrane-cell wall gap, a finding similar to that reported in *Ormosia hosiei* seedlings (Liu and Wei 2021). Membrane damage results from free radical-peroxidized polyunsaturated fatty acids (Zhang et al. 2021) and protoplasm shrinkage during drought. However, in our study, microbial-inoculated maize had reduced damage and protoplasm shrinkage, attributed to osmolytes (trehalose, glucose, glycine betaine, proline, polyamines) stabilizing proteins, enzymes, membranes, and cell turgor (Sharma et al. 2019). Inoculated plants showed lower LOX activity, lipid peroxidation, and electrolyte leakage, essential for membrane protection. This can be attributed to the increased presence of nutrients (K^+ and Ca^{2+}), enzymatic antioxidants (APX, CAT, and SOD), and non-enzymatic antioxidants (polyamines, α-tocopherols, proline) in inoculated plants. Proline scavenges ROS, preserving cells from stress (Ashraf and Foolad 2007). Calcium bolsters membrane integrity and regulates ions (de Bang et al. 2021).

Polyamines enhance membrane stability (Wang et al. 2022). The α-tocopherol shields against lipid peroxidation (Niki 2021). Antioxidant enzymes (APX, SOD, and CAT) counter ROS (Ru et al. 2023) during adverse environmental conditions. Potassium plays a pivotal role in osmoregulation and membrane potential maintenance (Sardans and Peñuelas

2021). Consequently, microbial inoculation mitigates membrane injury by facilitating the accumulation of osmolytes, nutrients, and antioxidants, thereby diminishing lipid peroxidation in plants (Tyagi et al. 2023). Additionally, in our study, microbial-inoculated maize plants exhibited a higher degree of vesicle formation compared to non-inoculated plants in our study. Elevated vesicle formation constitutes a stress-responsive mechanism (Karim et al. 2014) aimed at augmenting membrane surface area. This augmentation aids in facilitating ion transport and storage, consequently affording protection against ionic imbalances within cells. Increased vesicle formation signifies enhanced drought resilience in plants subjected to microbial inoculation.

5 Conclusion

In conclusion, this research offers compelling evidence that the application of *P. indica* inoculum, AMF consortium, and PGPB consortium significantly enhances the performance of maize plants when subjected to drought stress. The introduction of microbial inoculants provides protection to maize plants against drought stress by improving their physiological responses and increasing the accumulation of osmolytes. This leads to enhanced growth and biomass compared to plants without inoculation. The increased osmolyte levels enable effective osmotic adjustment and more efficient scavenging of free radicals. Additionally, microbial colonization helps maintain optimal levels of essential macronutrients and micronutrients, thereby mitigating ionic imbalances caused by drought. Microbial-inoculated maize plants exhibit reduced lipid peroxidation, membrane permeability, and LOX activity. Simultaneously, there is an upregulation in the activity of antioxidant enzymes like APX, CAT, and SOD, accompanied by the accumulation of antioxidant metabolites such as polyamines, proline, and α-tocopherol. This microbial inoculation establishes a robust antioxidant defense system that effectively prevents oxidative damage.

Furthermore, microbial colonization contributes to the mitigation of structural changes induced by drought stress. This can be attributed to the improved antioxidant defense system, enhanced nutrient uptake, and an increased accumulation of osmolytes. Based on the findings of this investigation, it is evident that the utilization of *P. indica*, AMF consortium, and PGPB consortium holds great promise for enhancing the productivity and stress resilience of maize plants in drought conditions. When comparing the effectiveness of microbial inoculants in alleviating antioxidant responses, enhancing nutrient uptake, and mitigating ultrastructural changes in maize plants, it is worth noting that PGPB-inoculated plants demonstrated the highest degree of drought resilience, followed by AMF and *P. indica*-inoculated plants in succession. These microbial inoculants show significant potential for meeting the needs of sustainable agriculture in drought-prone environments.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13199-023-00945-5.

Acknowledgements The authors extend their sincere thanks to Professor Ajit Verma, AIMT, Amity University, Uttar Pradesh, India, for providing the *Piriformospora indica* inoculum. MS, JGS, and BG would like to express their gratitude to the Department of Microbiology at IARI, New Delhi, for generously providing the PGPB inoculum and AMF inoculum.

Author contribution MS, JGS and BG formulated the research framework and its conceptual design. MS executed all experimental procedures. MS authored the initial manuscript, which underwent subsequent refinement by JGS and BG. MS, JGS, and BG approved the final version of the manuscript.

Funding MS and JGS also acknowledge Delhi Technological University for their generous financial support in recording the TEM and ICPMS data.

Data availability All data generated or analyzed during this research have been incorporated into this research article. For any further inquiries, please contact the corresponding author.

Declarations

Conflict of interest Authors declare no conflict of interest.

References

A Saga of Progress (2012) Compendium of 50 years of achievements. Punjab Agricultural University, Ludhiana, p 230

Aalipour H, Nikbakht A, Sabzalian MR (2023) Essential oil composition and total phenolic content in Cupressus arizonica G. in response to microbial inoculation under water stress conditions. Sci Rep 13(1):1–11. https://doi.org/10.1038/s41598-023-28107-z

Abbas S, Javed MT, Shahid MI, Haider MZ, Chaudhary HJ, Tanwir K, Maqsood A (2020) Acinetobacter sp. SG-5 inoculation alleviates cadmium toxicity in differentially cd tolerant maize cultivars as deciphered by improved physio-biochemical attributes, antioxidants and nutrient physiology. Plant Physiol Biochem 155:815–827. https://doi.org/10.1016/j.plaphy.2020.08.024

Abobatta WF (2019) Drought adaptive mechanisms of plants–a review. Adv Agric Environ Sci 2(1):42–45. https://doi.org/10.30881/aaeoa.00021

Agnihotri R, Sahni S, Sharma MP, Gupta MM (2022) Facets of AM fungi in sequestering soil carbon and improving soil health. In: Rajpal VR, Singh I, Navi SS (eds) Fungal diversity, ecology and control management, Fungal Biology. Springer, Singapore, pp 327–344. https://doi.org/10.1007/978-981-16-8877-5_15

Akhilesh S, Basandrai AK, Kalia V (2000) Resistance in rabi maize to Fusarium stalk rot (Fusarium moliniformae) and common rust (Puccinia sorghi). Crop Res (Hisar) 20(2):221–225

Alam MZ, Choudhury TR, Mridha MA (2023) Arbuscular mycorrhizal fungi enhance biomass growth, mineral content, and antioxidant activity in tomato plants under drought stress. J Food Qual 2023:2581608. https://doi.org/10.1155/2023/2581608

Aldesuquy HS, Ibraheem FL, Ghanem HE (2018) Exogenously supplied salicylic acid and trehalose protect growth vigor,

chlorophylls and thylakoid membranes of wheat flag leaf from drought-induced damage. J Agric for Meteorol Res 1(1):13–20
Ali R, Gul H, Hamayun M, Rauf M, Iqbal A, Hussain A, Lee IJ (2022) Endophytic fungi controls the physicochemical status of maize crop under salt stress. Pol J Environ Stud 31:561–573. https://doi.org/10.15244/pjoes/134540
Ali R, Hassan S, Shah D, Sajjad N, Bhat EA (2020) Role of polyamines in mitigating abiotic stress. Roychoudhury A, Tripathi DK (ed) Protective chemical agents in the amelioration of plant abiotic stress: biochemical and molecular perspectives, vol 22. Wiley, New York, pp 291–305. https://doi.org/10.1002/9781119552154.ch13
Arnon DI (1949) Copper enzymes isolated chloroplasts, polyphenoloxidase in Beta vulgaris. Plant Physiol 24(1):1–15
Ashraf M, Foolad MR (2007) Roles of glycine betaine and proline in improving plant abiotic stress resistance. Environ Exp Bot 59(2):206–216. https://doi.org/10.1016/j.envexpbot.2005.12.006
Azizi M, Fard EM, Ghabooli M (2021) Piriformospora indica affect drought tolerance by regulation of genes expression and some morphophysiological parameters in tomato (Solanum lycopersicum L). Sci Hort 287:110260. https://doi.org/10.1016/j.scienta.2021.110260
de Bang TC, Husted S, Laursen KH, Persson DP, Schjoerring JK (2021) The molecular–physiological functions of mineral macronutrients and their consequences for deficiency symptoms in plants. New Phytol 229(5):2446–2469. https://doi.org/10.1111/nph.17074
Bates LS, Waldren RP, Teare ID (1973) Rapid determination of free proline for water-stress studies. Plant Soil 39:205–207. https://doi.org/10.1007/BF00018060
Begum N, Wang L, Ahmad H, Akhtar K, Roy R, Khan MI, Zhao T (2022) Co-inoculation of Arbuscular Mycorrhizal Fungi and the plant growth-promoting Rhizobacteria improve growth and photosynthesis in Tobacco under Drought stress by Up-Regulating antioxidant and Mineral Nutrition Metabolism. Microb Ecol 83:971–988. https://doi.org/10.1007/s00248-021-01815-7
Bhagat N, Raghav M, Dubey S, Bedi N (2021) Bacterial exopolysaccharides: insight into their role in plant abiotic stress tolerance.J. Microbiol Biotechnol 31(8):1045–1059. https://doi.org/10.4014/jmb.2105.05009
Bidalia A, Okram Z, Hanief M, Rao KS (2018) Assessment of tolerances in Mitragyna parvifolia (Roxb.) Korth. And Syzygium cumini Keels. Seedlings to waterlogging. Photosynthetica 56(2):707–717. https://doi.org/10.1007/s11099-017-0724-1
Bisht A, Gupta MM (2023) Arbuscular Mycorrhiza Fungi Resources for sustainable and climate-smart cultivation of Maize. In: Singh I, Rajpal VR, Nav SS (eds) Fungal resources for sustainable economy: current status and future perspectives. Springer, Singapore, pp 299–317. https://doi.org/10.1007/978-981-19-9103-5_11
Boutasknit A, Baslam M, Ait-El-Mokhtar M, Anli M, Ben-Laouane R, Douira A, El Modafar C, Mitsui T, Wahbi S, Meddich A (2020) Arbuscular mycorrhizal fungi mediate drought tolerance and recovery in two contrasting carob (Ceratonia siliqua L.) ecotypes by regulating stomatal, water relations, and (in) organic adjustments. Plants 9(1):80. https://doi.org/10.3390/plants9010080
Brehelin C, Kessler F, van Wijk KJ (2007) Plastoglobules: versatile lipoprotein particles in plastids. Trends Plant Sci 12:260–266. https://doi.org/10.1016/j.tplants.2007.04.003
Cao JL, He WX, Zou YN, Wu QS (2023) An endophytic fungus, Piriformospora indica, enhances drought tolerance of trifoliate orange by modulating the antioxidant defense system and composition of fatty acids. Tree Physiol 43(3):452–466. https://doi.org/10.1093/treephys/tpac126
Chiappero J, del Rosario Cappellari L, Alderete LG, Palermo TB, Banchio E (2019) Plant growth promoting rhizobacteria improve the antioxidant status in Mentha piperita grown under drought stress leading to an enhancement of plant growth and total phenolic content. Ind Crops Prod 139:111553. https://doi.org/10.1016/j.indcrop.2019.111553
Chourasiya D, Gupta MM, Sahni S, Oehl F, Agnihotri R, Buade R, Maheshwari HS, Prakash A, Sharma MP (2021) Unraveling the AM fungal community for understanding its ecosystem resilience to changed climate in agroecosystems. Symbiosis 84:295–310. https://doi.org/10.1007/s13199-021-00761-9
Devarajan AK, Muthukrishanan G, Truu J, Truu M, Ostonen I, Kizhaeral SS, Panneerselvam P, Kuttalingam Gopalasubramanian S (2021) The foliar application of rice phyllosphere bacteria induces drought-stress tolerance in Oryza sativa (L). Plants 10(2):387. https://doi.org/10.3390/plants10020387
Dhindsa RS, Plumb-Dhindsa P, Thorpe TA (1981) Leaf senescence: correlated with increased levels of membrane permeability and lipid peroxidation, and decreased levels of superoxide dismutase and catalase. J Exp Bot 32(1):93–101. https://doi.org/10.1093/jxb/32.1.93
Diatta AA, Min D, Jagadish SK (2021) Drought stress responses in non-transgenic and transgenic alfalfa—current status and future research directions. Adv Agron 170:35–100. https://doi.org/10.1016/bs.agron.2021.06.002
Doderer A, Kokkelink I, Van Der Veen S, Valk BE, Schram AW, Douma AC (1992) Purification and characterization of two lipoxygenase isoenzymes from germinating barley. Biochim Biophys Acta 1120(1):97–104. https://doi.org/10.1016/0167-48389290429-h
EN ISO 20483 (2006) AN 300 the determination of nitrogen according to Kjeldahl using block digestion and steam distillation. The International Organization for Standardization, Geneva
Fadiji AE, Santoyo G, Yadav AN, Babalola OO (2022) Efforts towards overcoming drought stress in crops: revisiting the mechanisms employed by plant growth-promoting bacteria. Front Microbiol 13:962427. https://doi.org/10.3389/fmicb.2022.962427
Farooq MA, Kobayashi WN, Fujita D, Basra SM (2009) Plant drought stress: effects, mechanisms and management. Agron Sustain Dev 29(1):185–212. https://doi.org/10.1051/agro:2008021
Ferreira JC, Paschoalin VMF, Panek AD, Trugo LC (1997) Comparison of three different methods for trehalose determination in yeast extract. Food Chem 60(2):251–254. https://doi.org/10.1016/S0308-81469600330-5
Fresno DH, Solé-Corbatón H, Munné-Bosch S (2023) Water stress protection by the arbuscular mycorrhizal fungus Rhizoglomus irregulare involves physiological and hormonal responses in an organ-specific manner. Physiol Plant 175(1):e13854. https://doi.org/10.1111/ppl.13854
Giovannetti M, Mosse B (1980) An evaluation of techniques for measuring vesicular arbuscular mycorrhizal infection in roots. New Phytol 84(3):489–500
Grieve CM, Grattan SR (1983) Rapid assay for determination of water soluble quaternary ammonium compounds. Plant Soil 70(2):303–307. https://doi.org/10.1007/BF02374789
Gunasekera D, Ratnasekera D (2023) Advancement in mitigating the effects of drought stress in wheat. In: Khan MK, Pandey A, Hamurcu H, Gupta OP, Gezgin S (eds) Abiotic Stresses in Wheat. Academic Press, Cambridge, pp 297–311. https://doi.org/10.1016/B978-0-323-95368-9.00023-0
Haghighi TM, Saharkhiz MJ, Kavoosi G, Zarei M (2023) Adaptation of Glycyrrhiza glabra L. to water deficiency based on carbohydrate and fatty acid quantity and quality. Sci Rep 13(1):1766. https://doi.org/10.1038/s41598-023-28807-6
Hari NS, Jindal J, Malhi NS (2008) Resistance of Cry1Ab maize to spotted stemborer Chilo partellus (Lepidoptera: Crambidae) in

India. Int J Trop Insect Sci 27:223–228. https://doi.org/10.1017/S1742758407850983

Hasanuzzaman M, Bhuyan MB, Nahar K, Hossain M, Mahmud J, Hossen M, Masud A, Moumita Fujita M (2018) Potassium: a vital regulator of plant responses and tolerance to abiotic stresses. Agronomy 8(3):31. https://doi.org/10.3390/agronomy8030031

Hashem A, Abd-Allah EF, Alqarawi AA, Egamberdieva D (2016) Bioremediation of adverse impact of cadmium toxicity on Cassia italica Mill by arbuscular mycorrhizal fungi. Saudi J Biol Sci 23(1):39–47. https://doi.org/10.1016/j.sjbs.2015.11.007

Hashemi M, Behboodian B, Karimi E, Oskoueian E (2022) Azotobacter chroococcum inoculation under low drought stress condition improves Trachyspermum ammi seeds' essential oil bioactivity. Biochem Syst Ecol 105:104537. https://doi.org/10.1016/j.bse.2022.104537

Heath RL, Packer L (1968) Photoperoxidation in isolated chloroplasts: I. kinetics and stoichiometry of fatty acid peroxidation. Arch Biochem Biophys 125(1):189–198. https://doi.org/10.1016/0003-98616890654-1

van Der Heijden MG, Martin FM, Selosse MA, Sanders IR (2015) Mycorrhizal ecology and evolution: the past, the present, and the future. New Phytol 205(4):1406–1423. https://doi.org/10.1111/nph.13288

Hiscox JD, Israelstam GFA (1979) Method for the extraction of chlorophyll from leaf tissue without maceration. Can J Bot 57(12):1332–1334. https://doi.org/10.1139/b79-163

Inbaraj MP (2021) Plant-microbe interactions in alleviating abiotic stress - a mini review. Front Agron 3:667903. https://doi.org/10.3389/fagro.2021.667903

Jarrett U, Miller S, Mohtadi H (2023) Dry spells and global crop production: a multi-stressor and multi-timescale analysis. Ecol Econ 203:107627. https://doi.org/10.1016/j.ecolecon.2022.107627

Jofre MF, Mammana SB, Lopez Appiolaza M, Silva MF, Gomez FJ, Cohen AC (2023) Melatonin production by rhizobacteria native strains: towards sustainable plant growth promotion strategies. Physiol Plant 175(1):e13852

Jyoti K, Ramesh K, Kumar RS, Sain D, Bilal B, Kamboj OP, Usha N, Yadav AK (2013) Response of maize (Zea mays L.) hybrids to excess soil moisture stress at different growth stages. Res Crops 14(2):439–443

Karim S, Alezzawi M, Garcia-Petit C, Solymosi K, Khan NZ, Lindquist E, Dahl P, Hohmann S, Aronsson H (2014) A novel chloroplast localized Rab GTPase protein CPRabA5e is involved in stress, development, thylakoid biogenesis and vesicle transport in Arabidopsis. Plant Mol Biol 84:675–692. https://doi.org/10.1007/s11103-013-0161-x

Khan SW, Yaseen T, Naz F, Abidullah S, Kamil M (2019) Influence of arbuscular mycorrhizal fungi (AMF) inoculation on growth and mycorrhizal dependency of (Lens culinaris L.) varieties. Int J Bioorganic Chem 4:47

Khazaei Z, Esmaielpour B, Estaji A (2020) Ameliorative effects of ascorbic acid on tolerance to drought stress on pepper (Capsicum annuum L) plants. Physiol Mol Biology Plants 26(8):1649–1662. https://doi.org/10.1007/s12298-020-00846-7

Kim SE, Lee CJ, Ji CY, Kim HS, Park SU, Lim YH, Park WS, Ahn MJ, Bian X, Xie Y, Guo X, Kwak SS (2019) Transgenic sweetpotato plants overexpressing tocopherol cyclase display enhanced α-tocopherol content and abiotic stress tolerance. Plant Physiol Biochem 144:436–444. https://doi.org/10.1016/j.plaphy.2019.09.046

Komatsu KJ, Esch NL, Bloodworth KJ, Burghardt KT, McGurrin K, Pullen JD, Parker JD (2023) Rhizobial diversity impacts soybean resistance, but not tolerance, to herbivory during drought. Basic Appl Ecol 66:31–39. https://doi.org/10.1016/j.baae.2022.12.004

Kour D, Rana KL, Yadav AN, Sheikh I, Kumar V, Dhaliwal HS, Saxena AK (2020) Amelioration of drought stress in Foxtail millet (Setaria italica L.) by P-solubilizing drought-tolerant microbes with multifarious plant growth promoting attributes. Environ Sustain 3(1):23–34. https://doi.org/10.1007/s42398-020-00094-1

Kumar N (2015) Effect of copper mining dust on the soil and vegetation in India: a critical review. Int J Mod Sci Eng Technol 2(2):73–76

Kumar R, Kaul J, Dubey RB, Singode A, Gk C, Manivannan A, Debnath MK (2015) Assessment of drought tolerance in maize (Zea mays L.) based on different indices. SABRAO J Breed Genet 47(3):291–298

Kumar S, Kumar S, Mohapatra T (2021) Interaction between macro- and micro-nutrients in plants. Front Plant Sci 12:665583. https://doi.org/10.3389/fpls.2021.665583

Liang D, Ni ZY, Xia H, Xie Y, Lv X, Wang J, Lin L, Deng Q, Luo X (2019) Exogenous melatonin promotes biomass accumulation and photosynthesis of kiwifruit seedlings under drought stress. Sci Hortic 246:34–43. https://doi.org/10.1016/j.scienta.2018.10.058

Liu Y, Guo Z, Shi H (2022) Rhizobium Symbiosis leads to increased Drought Tolerance in chinese milk vetch (Astragalus sinicus L). Agronomy 12(3):725. https://doi.org/10.3390/agronomy12030725

Liu CY, Hao Y, Wu XL, Dai F, Abd-Allah EF, Wu Q, Liu S (2023) Arbuscular mycorrhizal fungi improve drought tolerance of tea plants via modulating root architecture and hormones. Plant Growth Regul 4:1–10. https://doi.org/10.1007/s10725-023-00972-8

Liu Y, Wei X (2021) Dark septate endophyte improves the drought-stress resistance of *Ormosia hosiei* seedlings by altering leaf morphology and photosynthetic characteristics. Plant Ecol 222:761–771. https://doi.org/10.1007/s11258-021-01135-3

López-Galiano MJ, García-Robles I, González-Hernández AI, Camañes G, Vicedo B, Real MD, Rausell C (2019) Expression of miR159 is altered in tomato plants undergoing drought stress. Plants 8(7):201. https://doi.org/10.3390/plants8070201

Mahreen N, Yasmin S, Asif M, Yahya M, Ejaz K, Mehboob-Ur-Rahman Yousaf S, Amin I, Zulfiqar S, Imran A, Khaliq S, Arif M (2023) Mitigation of water scarcity with sustained growth of Rice by plant growth promoting bacteria. Front Plant Sci 14:1081537. https://doi.org/10.3389/fpls.2023.1081537

Marcé M, Brown DS, Capell T, Figueras X, Tiburcio AF (1995) Rapid high-performance liquid chromatographic method for the quantification of polyamines as their dansyl derivatives: application to plant and animal tissues. J Chromatogr B 666(2):329–335. https://doi.org/10.1016/0378-43479400586-T

Minocha R, Shortle WC, Long SL, Minocha SC (1994) A rapid and reliable procedure for extraction of cellular polyamines and inorganic ions from plant tissues. J Plant Growth Regul 13(4):187–193. https://doi.org/10.1007/BF00226036

Mirzabaev A, Kerr RB, Hasegawa T, Pradhan P, Wreford A, Cristina T, von der Pahlen M, Gurney-Smith H (2023) Severe climate change risks to food security and nutrition. Clim Risk Manage 39:100473. https://doi.org/10.1016/j.crm.2022.100473

Mori Y (2015) Functional polymeric membrane in agriculture. In: Cirillo G, Spizzirri UG, Iemma F (eds) Functional Polymers in Food Science: From Technology to Biology. John Wiley & Sons, vol 2, pp 33–45. https://doi.org/10.1002/9781119108580.ch3

Nakano Y, Asada K (1981) Hydrogen peroxide is scavenged by ascorbate-specific peroxidase in spinach-chloroplasts. Plant Cell Physiol 22:867–880

Naservafaei S, Sohrabi Y, Moradi P (2023) Effects of exogenous application of 24-Epibrassinolide on photosynthesis parameters, Grain Yield, and protein of Dragon's Head (Lallemantia iberica) under

Drought stress conditions. J Plant Growth Regul 18:1–13. https://doi.org/10.1007/s00344-023-10910-7

Nemeskeri E, Horvath KZ, Andryei B, Ilahy R, Takács S, Neményi A, Pék Z, Helyes L (2022) Impact of plant growth-promoting rhizobacteria inoculation on the physiological response and productivity traits of field-grown tomatoes in Hungary. Horticulturae 8(7):641. https://doi.org/10.3390/horticulturae8070641

Niki E (2021) Lipid oxidation that is, and is not, inhibited by vitamin E: consideration about physiological functions of vitamin E. Free Radic Biol Med 176:1–15. https://doi.org/10.1016/j.freeradbiomed.2021.09.001

Oliveira TC, Cabral JS, Santana LR, Tavares GG, Santos LDS, Paim TP, Müller C, Silva FG, Costa AC, Souchie EL, Mendes GC (2022) The arbuscular mycorrhizal fungus Rhizophagus clarus improves physiological tolerance to drought stress in soybean plants. Sci Rep 12(1):9044. https://doi.org/10.1038/s41598-022-13059-7

Ortiz N, Armada E, Duque E, Roldán A, Azcón R, Roldan A, Azcon R (2015) Contribution of arbuscular mycorrhizal fungi and/or bacteria to enhancing plant drought tolerance under natural soil conditions: effectiveness of autochthonous or allochthonous strains. J Plant Physiol 174:87–96. https://doi.org/10.1016/j.jplph.2014.08.019

Ostadi A, Javanmard A, Amani Machiani M, Sadeghpour A, Maggi F, Nouraein M, Morshedloo MR, Hano C, Lorenzo JM (2022) Co-application of TiO2 nanoparticles and arbuscular mycorrhizal fungi improves essential oil quantity and quality of sage (Salvia officinalis L.) in drought stress conditions. Plants 11(13):1659. https://doi.org/10.3390/plants11131659

Pandey GK, Sanyal SK, Pandey GK (2021) Calcium-from nutrition to signaling. Functional dissection of calcium homeostasis and transport machinery in plants, 1–9. https://doi.org/10.1007/978-3-030-58502-0_1

Phillips JM, Hayman DS (1970) Improved procedures for clearing roots and staining parasitic and vesicular-arbuscular mycorrhizal fungi for rapid assessment of infection. Trans Br Mycological Soc 55(1):158–IN18. https://doi.org/10.1016/S0007-15367080110-3

Rahmani V, Movahhedi Dehnavi M, Balouchi H, Yadavi A, Hamidian M (2023) Silicon can improve nutrient uptake and performance of Black Cumin under Drought and Salinity stresses. Commun Soil Sci Plant Anal 54(3):297–310. https://doi.org/10.1080/00103624.2022.2112590

Rajput S, Sengupta P, Kohli I, Varma A, Singh PK, Joshi NC (2022) Role of Piriformospora indica in inducing soil microbial communities and drought stress tolerance in plants. In: Singh H and Vaishnav A (eds) New and future developments in microbial biotechnology and bioengineering, sustainable agriculture: Microorganisms as Biostimulants. Elsevier, Amsterdam, pp 93–110. https://doi.org/10.1016/B978-0-323-85163-3.00003-X

Ramakrishna W, Rathore P, Kumari R, Yadav R (2020) Brown gold of marginal soil: plant growth promoting bacteria to overcome plant abiotic stress for agriculture, biofuels and carbon sequestration. Sci Total Environ 711:135062. https://doi.org/10.1016/j.scitotenv.2019.135062

Ru C, Wang K, Hu X, Chen D, Wang W, Yang H (2023) Nitrogen modulates the effects of heat, drought, and combined stresses on photosynthesis, antioxidant capacity, cell osmoregulation, and grain yield in winter wheat. J Plant Growth Regul 42:1681–1703. https://doi.org/10.1007/s00344-022-10650-0

Sadasivam S, Manickam A (2008) Biochemical methods (3rd edn). New Age International Publishers, New Delhi

Sardans J, Peñuelas J (2021) Potassium control of plant functions: ecological and agricultural implications. Plants 10(2):419. https://doi.org/10.3390/plants10020419

Shah MN, Wright DL, Hussain S, Koutroubas SD, Seepaul R, George S, Ali S, Naveed M, Khan M, Altaf MT, Ghaffor K (2023) Organic fertilizer sources improve the yield and quality attributes of maize (Zea mays L.) hybrids by improving soil properties and nutrient uptake under drought stress. J King Saud Univ-Sci 35(4):102570

Sharma A, Shahzad B, Kumar V, Kohli SK, Sidhu GPS, Bali AS, Handa N, Kapoor D, Bhardwaj R, Zheng B (2019) Phytohormones regulate accumulation of osmolytes under abiotic stress. Biomolecules 9(7):285. https://doi.org/10.3390/biom9070285

Shu S, Yuan Y, Chen J, Sun J, Zhang W, Tang Y, Zhong M, Guo S (2015) The role of putrescine in the regulation of proteins and fatty acids of thylakoid membranes under salt stress. Sci Rep 5(1):1–16. https://doi.org/10.1038/srep14390

Siddique S, Naveed M, Yaseen M, Shahbaz M (2022) Exploring potential of seed endophytic bacteria for enhancing drought stress resilience in maize (Zea mays L). Sustainability 14(2):673. https://doi.org/10.3390/su14020673

Silva S, Dias MC, Silva A (2023) Potential of MgO and MgCO3 nanoparticles in modulating lettuce physiology to drought. Acta Physiol Plant 45(2):1–9. https://doi.org/10.1007/s11738-022-03507-2

Singh M, Sharma JG, Giri B (2023) Microbial inoculants alter resilience towards drought stress in wheat plants. Plant Growth Regul. https://doi.org/10.1007/s10725-023-01059-0

Sinha SK, Sinha AK, Thakur D, Lakra A, Tripathi AK (2019) Maize research in Chhattisgarh: status and progress. Int J Curr Microbiol Appl Sci 8:2210–2230. https://doi.org/10.20546/ijcmas.2019.803.265

Srinivasa Rao C, Prabhakar M, Maheswari M, Srinivasa Rao M, Sharma KL, Srinivas K, Prasad JVNS, Rama Rao CA, Vanaja M, Ramana DBV, Gopinath KA, Subba Rao AVM, Rejani R, Bhaskar S, Sikka AK, Alagusundaram K (2016) National Innovations in Climate resilient agriculture (NICRA), research highlights 2015-16. Central Research Institute for Dryland Agriculture, Hyderabad, p 112

Tang L, Xiao L, Chen E, Lei X, Ren J, Yang Y, Xiao B, Gong C (2023) Magnesium transporter CsMGT10 of tea plants plays a key role in chlorosis leaf vein greening. Plant Physiol Biochem 17:107842. https://doi.org/10.1016/j.plaphy.2023.107842

Teranishi Y, Tanaka A, Osumi M, Fukui S (1974) Catalase activities of hydrocarbon-utilizing Candida yeasts. Agric Biol Chem 38(6):1213–1220. https://doi.org/10.1080/00021369.1974.10861301

Timmusk S, Abd El-Daim IA, Copolovici L, Tanilas T, Kännaste A, Behers L, Nevo E, Seisenbaeva G, Stenström E, Niinemets Ü (2014) Drought-tolerance of wheat improved by rhizosphere bacteria from harsh environments: enhanced biomass production and reduced emissions of stress volatiles. PLoS ONE 9(5):e96086. https://doi.org/10.1371/journal.pone.0096086

Tyagi J, Mishra A, Kumari S, Singh S, Agarwal H, Pudake RN, Varma A, Joshi NC (2023) Deploying a microbial consortium of Serendipita indica, Rhizophagus intraradices, and Azotobacter chroococcum to boost drought tolerance in maize. Environ Exp Bot 206:105142. https://doi.org/10.1016/j.envexpbot.2022.105142

Tyagi J, Shrivastava N, Sharma AK, Varma A, Pudake R (2021) Effect of Rhizophagus intraradices on growth and physiological performance of Finger Millet (Eleusine coracana L.) under drought stress. Plant Sci Today 8(4):912–923. https://doi.org/10.14719/pst.2021.8.4.1240

Ullah A, Farooq M (2022) The challenge of drought stress for grain legumes and options for improvement. Arch Agron Soil Sci 68(11):1601–1618. https://doi.org/10.1080/03650340.2021.1906413

Vishnupradeep R, Bruno LB, Taj Z, Karthik C, Challabathula D, Kumar A, Freitas H, Rajkumar M (2022) Plant growth promoting bacteria improve growth and phytostabilization potential of

Zea mays under chromium and drought stress by altering photosynthetic and antioxidant responses. Environ Technol Innov 25:102154. https://doi.org/10.1016/j.eti.2021.102154

Wang K, Li J, Yi Y, Lv B, Wu Y, Wang C, Li H, Li Y, Liu Y, Cai X, Meng X, Jiang X, Zheng X, Zhou Z, Bu W (2022) Polyamine-activated carbonyl stress strategy for oxidative damage therapy. Nano Today 42:101355. https://doi.org/10.1016/j.nantod.2021.101355

Wang Y, Zou YN, Shu B, Wu Q (2023) Deciphering molecular mechanisms regarding enhanced drought tolerance in plants by arbuscular mycorrhizal fungi. Sci Hort 308:111591. https://doi.org/10.1016/j.scienta.2022.111591

Waqar A, Bano A, Ajmal M (2022) Effects of PGPR bioinoculants, hydrogel and biochar on growth and physiology of soybean under drought stress. Commun Soil Sci Plant Anal 53(7):826–847. https://doi.org/10.1080/00103624.2022.2028818

Yadav S, Modi P, Dave A, Vijapura A, Patel D, Patel M (2020) Effect of abiotic stress on crops. In: Sustainable Crop Production, IntechOpen, London, pp 3–24. https://doi.org/10.5772/intechopen.88434

Yasmeen S, Wahab A, Saleem MH, Ali B, Qureshi KA, Jaremko M (2022) Melatonin as a Foliar Application and Adaptation in Lentil (Lens culinaris Medik.) Crops under Drought stress. Sustainability 14(24):16345. https://doi.org/10.3390/su142416345

Yu J, Fan N, Hao T, Bian Y, Zhuang L, Li Q, Yang Z (2023) Ethionine-mitigation of drought stress associated with changes in root viability, antioxidant defense, osmotic adjustment, and endogenous hormones in tall fescue. Plant Growth Regul 21:1–14. https://doi.org/10.1007/s10725-022-00944-4

Zhang K, Zhang Y, Sun J, Meng J, Tao J (2021) Deterioration of orthodox seeds during ageing: influencing factors, physiological alterations and the role of reactive oxygen species. Plant Physiol Biochem 158:475–485. https://doi.org/10.1016/j.plaphy.2020.11.031

Zhu J, Cai D, Wang J, Cao J, Wen Y, He J, Zhao L, Wang D, Zhang S (2021) Physiological and anatomical changes in two rapeseed (Brassica napus L.) genotypes under drought stress conditions. Oil Crop Sci 6(2):97–104. https://doi.org/10.1016/j.ocsci.2021.04.003

Zwiazek JJ, Blake TJ (1991) Early detection of membrane injury in black spruce (Picea mariana). Can J for Res 21(3):401–404. https://doi.org/10.1139/x91-050

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.

JOM
https://doi.org/10.1007/s11837-023-06220-8
© 2023 The Minerals, Metals & Materials Society

TECHNICAL ARTICLE

Morphology of Finished Brass Surface by Thermal Additive Centrifugal Abrasive Flow Machining Process Using Novel Electrode

ANANT BHARDWAJ ,[1,2] KROVVIDI SRINIVAS,[1] and RAJIV CHAUDHARY[1]

1.—Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India.
2.—e-mail: anant5892@gmail.com

Thermal additive centrifugal abrasive flow machining (TACAFM) is the latest hybrid of abrasive flow machining, which involves a combination of centrifugal action of rotating electrode and the spark to manage the limitation of low material removal (MR). The present investigation involves the study of the morphology of internal surface hollow cylindrical brass surface finished by TACAFM process. The researchers proposed a novel spline-shaped centrifugal force-generating electrode with a curved blade instead of straight blade for improving the process efficiency. The computation of MR and percentage improvement in surface finish ($\%\Delta R_a$) is done using the proposed electrode with curved blade, and the results are compared with the primitive electrode with straight blade against the process parameters (supply current, duty cycle, abrasive concentration, rotational speed of centrifugal force generating electrode and the extrusion pressure). The surface roughness plot using Taylor-Hobson tallysurf, micrographs taken by optical microscope and SEM images were used for analysis of the finished workpiece surface. The proposed electrode resulted in an 8.5% increase in MR and 27.38%ΔR_a compared to initial straight spline-shaped electrode used for TACAFM process. Also, XRD and EDAX analysis was done to predict the morphological change on the finished workpiece surface.

INTRODUCTION

Brass is an alloy of copper and zinc which applied in the fields of automobiles, aerospace, biomedical, electric and electronic components, machine tools, piping and fittings. The brass possesses adequate mechanical and thermal properties, which makes it suitable for manufacturing industries and modern machine equipment. Over the years, brass has been a prominent material analyzed by researchers, and its properties and application have been thoroughly analyzed in modern machining processes.[1–5] Many researchers reported the continuous application of brass in daily life. Thus, it is important to analyze the morphological change in the structure of brass when it undergoes various machining and fabrication processes.

Various conventional and non-conventional machining processes can be used for the machining of brass, including abrasive flow machining (AFM), magnetic abrasive finishing (MAF), magnetic float polishing (MFP), water jet machining (WJM), electronic beam machining (EBM), laser beam machining (LBM), electrical discharge machining (EDM), ion beam machining (IBM) and plasma beam machining (PBM), but when it comes to machining internal surfaces, abrasive flow machining (AFM) is a prominent machining option.[6] AFM is a non-traditional media finishing process, which uses an abrasive-based laden media to finish the internal surface of the workpiece. The process is applied in the finishing of intrinsic and inaccessible surfaces. The media is extruded within the hollow surface of the workpiece

(Received March 21, 2023; accepted October 9, 2023)

Published online: 06 November 2023

by means of pressure difference at the top and bottom of the cylinder resulting in the MR by abrasion process. The advantage of AFM process is that there are no structural or morphological defects on the workpiece surface. Also, compared to traditional finishing processes like lapping, honing and superfinishing, it does not require conventional tools for finishing, which increases its applicability.[7] Although low MR is the major limitation of the process, it is applied in the areas of automobiles, biomedicine, aerospace, cosmetics, defense manufacturing and heavy industry.[8,9] Ten hybrids of AFM have been developed so far to overcome the limitations of low MR based on the material removal mechanism and the input energy applied for material removal.[10] Basic hybridization is achieved by inserting a rotating rod co-axially with the internal surface of the workpiece, which provides the centrifugal action to the media along with the restricted path inside the workpiece, which enhances the abrasion process to foster material removal. Based on this principal, centrifugal force assisted AFM,[11,12] helical AFM,[13] drill-bit guided AFM[14] and rotational AFM[15] have been developed. The other hybrids were developed by utilizing different sources of energy by altering the composition of media, like carbonyl particles with abrasive laden media used in magneto AFM,[16] a hybrid of magneto and rotational AFM.[17] When the electrolyte is added to the media, the process of electrochemical machining along with conventional AFM takes place. This is electrochemical-aided AFM.[18] Ultrasonic AFM[19] includes additional ultrasonic vibrations in addition to the conventional extrusion of media. The advantages of electrochemical AFM and centrifugal force-assisted AFM are joined in the hybrid electrochemical and centrifugal force-assisted AFM.[20] TACAFM process is the latest hybridization of AFM, which uses the spark produced by the advantage of EDM and centrifugal force-assisted AFM.[20–25] Over the time period, the researcher has tried to improve the efficiency of AFM in different ways. Many scholars used improved fixture arrangement to enhancw the efficiency of the conventional AFM process;[26,27] many researchers have used simulation and mathematical modeling to improve the media flow path and the media interaction with the workpiece.[28–33] Many researchers focused on improvement of media development by changing the composition, constituents and concentration of media particles.[34–40] AFM is significantly used in modern applications, mainly the finishing of metallic alloys used in aerospace and automotive industries,[41–43] finishing of confocal channels, microchannel and thermal recast layers formed after machining by electric discharge machining (EDM), electron beam machining (EBM), laser beam machining (LBM) and selective laser machining.[44–46] AFM is prominently used in the finishing of additively manufactured parts mostly in the FDM (fused deposition modelling) process and various industrial parts.[47–49]

The recent hybrid of the TACAFM process is efficient in curbing the limitation of low MR of AFM, and the present investigation deals with improving its efficiency for finishing brass workpieces. The CFG electrode plays a vital role in the working of the CFAAFM and TACAFM process[50,51] as it provides centrifugal action to the media and acts as a cathode in the material removal process. Thus, the researcher proposed a novel spline-shaped electrode with a curved blade, which tends to increase the centrifugal force on the media compared to the primitive tool which results in the increase in active dynamic number of particles which enhances the material removal process.

TACAFM process is a hybrid of conventional AFM processes that uses the thermal energy produced by the spark (EDM effect) along with the abrasion mechanism of the abrasives striking the internal surface of the workpiece. A CFG electrode is placed co-axially between a hollow workpiece and is rotated to increase the centrifugal effect in the media (mixture of abrasive, polymer and gel), which is extruded within, up and down, between the workpiece. A constant gap is maintained between the workpiece and the electrode, and the pulsed DC supply is given to the electrode and the workpiece with the workpiece as cathode and electrode as anode. The electrons from the electrode move towards the workpiece and strike the media particle present between the gaps. The continuous striking of electrons ionized the media particles resulting in plasma formation. As the concentration of electrons increases, it results in spark formation. Figure 1 shows the mechanism of the TACAFM process. Figure 1a shows the arrangement of the electrode, media, abrasives and workpiece, whereas Fig. 1b shows the mechanism of spark formation.

MATERIALS AND METHODS

The present investigations aim to study the morphology of the internal surface of the machined brass workpiece machined using novel spline-shaped electrode with a curved blade, which would increase the centrifugal force on the media resulting in higher material removal. The initial investigations were done with the spline-shaped electrode with the straight blade. The experimentation was performed considering the optimum values of the process parameters were taken from the literature,[22] which were supply current, duty cycle, rotation speed, extrusion pressure and fraction of abrasive concentration.

The process parameters of TACAFM process are media flow rate, number of oscillations, type of abrasive particles, abrasive mesh size, current, rotational speed, extrusion pressure, etc. Among these process parameters, supply current, duty cycle, rotational speed of the CFG electrode, extrusion pressure and the abrasive concentration are considered in the investigation. The prominent

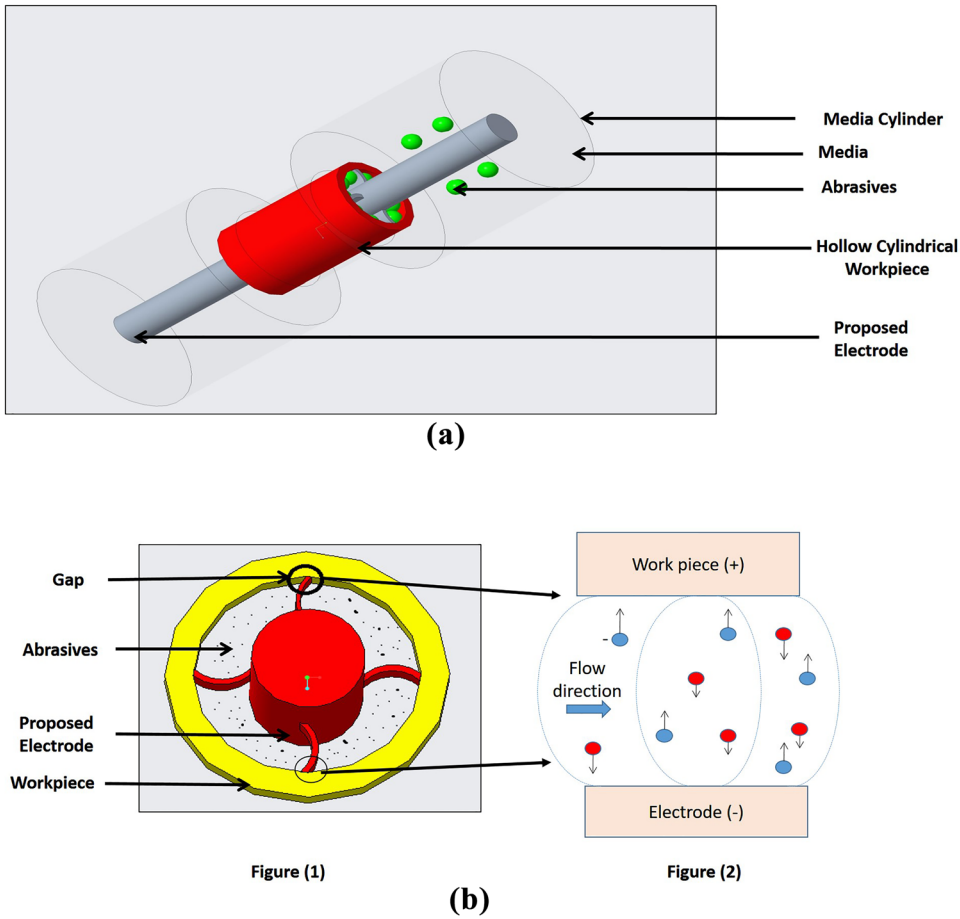


Fig. 1. Mechanism of TACAFM. (a) TACAFM arrangement. (b) Mechanism of spark formation. (b1) Top view of the electrode and work piece in the developed TACAFM process. (b2) Plasma channel formation in the gap between the rotating electrode and work piece surface.

reason for the selection of the above process parameters is their higher contribution in the material removal process. The supply current and duty cycle are the prominent parameters when we consider the material removal process in EDM and TACAFM processes. The combination of duty cycle and supply current creates an adequate potential difference and initiates the plasma formation, which results in the generation of sparks between the electrode and workpiece.[22,51] The MR significantly increases with the increase of current as it results in a deeper crater on the surface similar to the EDM process. On the other hand, with the increase in the duty cycle, the discharge increases, resulting in the increase in the energy concentration on the surface of the workpiece, which enhances MR. The rotational speed of the CFG electrode also increases the centrifugal action on the media, which increases the resultant force by which the abrasive impinges on the workpiece surface. Due to this, the abrasive hits with deeper penetration resulting in higher material removal.[50] Similarly, higher extrusion pressure results in higher axial force on the abrasive particle enhancing the value of the resultant force.[11] Higher abrasive concentration increases the number of dynamic active abrasives present in the media, which results in the higher MR.[10,12] Thus, the chosen parameters result in the increase in the MR which manages the major limitation of the conventional AFM process. The value of process parameters considered for this investigation are listed in Table III. To compare the performance of the proposed electrode compared with the conventional electrode, the proposed electrode is tested on the same parametric values as done on conventional electrode.[22]

The finished workpiece sample is analyzed under optical microscope and SEM for the morphological and structural study along with XRD analysis for validation of spark. Figure 2 shows the experimental setup and the proposed electrode. Figure 2a shows the TACAFM setup consists of conventional an AFM setup with 3 phase induction motor with power drive for rotation of centrifugal force generating (CFG) electrode and the EDM controller for providing optimal spark gap and the EDM power supply for the DC supply. Figure 2b shows the top view of the proposed electrode with rotating setup. Figure 2c shows the top view of the proposed electrode. Figure 2d shows the front view, and

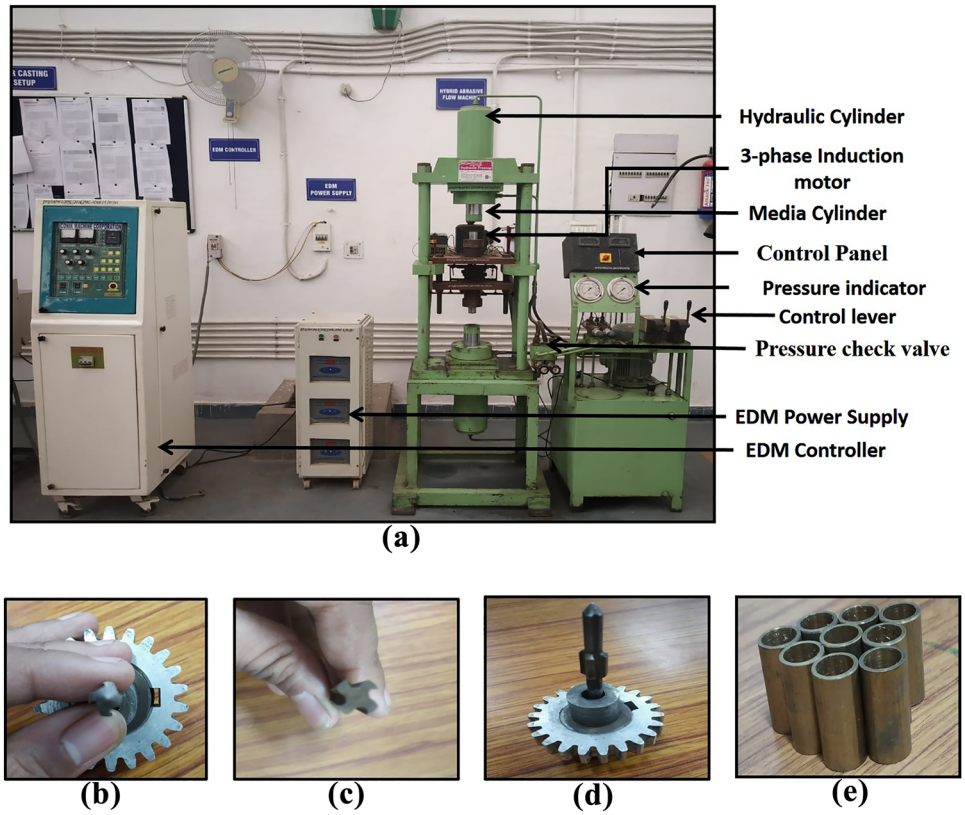


Fig. 2. Experimental setup. (a) TACAFM setup, (b) top view of the proposed electrode with gear, (c) shape of the curve blade of the proposed electrode, (d) front view of the proposed electrode, (e) hollow cylindrical brass workpiece.

Fig. 2e shows the hollow cylindrical brass workpiece with 10 mm OD, 8 mm ID and 16 mm height. The dimension of the proposed electrode is selected so that it can set into the ID of the workpiece with a 0.1-mm gap for the facilitation of spark. Table I shows the specification of the TACAFM setups.

MR and $\%\Delta R_a$ were calculated by the following expressions.

$$\%\Delta R_a = [(\text{Initial roughness} - \text{final roughness})/(\text{initial roughness})] \times 100 \quad (1)$$

$$\text{MR} = \text{Weight of the workpiece before machining} - \text{weight of the workpiece after machining} \quad (2)$$

RESULTS AND DISCUSSION

The experiments were performed on the single sample with the specified values of the process parameters, and the outcomes were observed. The $\%\Delta R_a$ is calculated based on Eq. 1, and the material removal (MR) is calculated based on Eq. 2. Taylor Hobson tallysurf was used to compute the value of surface roughness, and the precision weighing machine (SHIMADZU AUX220) was used for the computation of MR through difference in weight. Since the target surface was curved and it was difficult to mark the specific location on the internal surface before and after the machining process, so the value of surface roughness was taken at three locations on the workpiece before and after finishing, and the average was considered as a reading before and after finishing. Table II shows the surface roughness value before and after machining process. Figure 3 shows the surface roughness curves of machined and un-machined workpieces. Table III shows the experimentation details.

The value of MR obtained using the proposed novel electrode (curved blade) shows the value of MR as 40 g, which is 8.5% higher than that shown by the spline-shaped electrode with straight blades, which is 36.571 mg.[22] Also, the value of $\%\Delta R_a$ by the proposed electrode comes out to be 58.36% against the value of 42.38 for the straight spline-shaped electrode,[22] which is 27.38% higher, which proves the validity of the proposed novel electrode in improving the finishing efficiency of the developed TACAFM process.

The above results show the increase in the MR and $\%\Delta R_a$.. This effect is observed because of the curvature of the blade present in the proposed electrode. The machining in the TACAFM process is observed because of the combined action of the

Table I. Specification of equipment of TACAFM setup

Sr. no	Equipment	Specification
1	Conventional AFM setup	Working pressure = 20 MPa Media flow rate = 290 cm^3/min
2	Three-phase induction motor	Maximum speed = 800 rpm Maximum power = 1 HP
3	Power drive	Power = 0.75 KW/HP Input voltage (200–240v) for single phase
4	EDM power supply	Current = 25 Ampere Power supply = 415 Volt, three phase and 50 Hz
5	EDM controller	Microprocessor-based control system Maximum power-3 KVA

Table II. R_a of the workpiece before and after machining

Observations	Surface roughness value before machining micron	Surface roughness value after machining micron
1.	1.21	0.44
2.	1.00	0.43
3.	0.91	0.42
Average	1.04	0.43

thermal, centrifugal and conventional extrusion of the media within the internal surface of the workpiece. The thermal effect causes material softening due to spark formation, which results in deeper penetration of the abrasive particles over the surface of the workpiece. The total force exerted by the abrasive particle on the surface of the workpiece is the combination of force extorted because of conventional extrusion, thermal effect and Coriolis effect of CFAAFM process, which is given by Eq. 3[22]

$$F_{\text{TACAFM by Abrasive}} = \left[\sum_{i=\theta_a}^{\theta} (CA\sigma)\text{Sin}\theta_i \cdot N_d - F_{\text{thermal}}\right] \pm W \tag{3}$$

where $F_{(\text{TACAFM by abrasive})}$ is the total force acting on the workpiece by the abrasive particle during the TACAFM process, C is the flow constraint factor, A is projected area of abrasive particles, σ is uniaxial flow stress of the workpiece material, N_d is the number of dynamic active particle, $F_{(\text{thermal})}$ is the force on abrasive due to the thermal effect, and W is the weight of the abrasive particle. The negative and positive are used depending upon the direction of extrusion. Here, θ_i is the approach angle by which the abrasive particle hits the workpiece because of the resultant force produced by the centrifugal and axial force. The curve blade present on the novel proposed electrode increases the centrifugal force on the media particle by providing circular motion to the media which increases Coriolis force.[22,52,53] Due to this effect, the approach angle increases, which in turns increases total force exerted by the particle for indentation. Hence, the proposed tool with curvature effect results in the increase in penetration compared to the primitive tool used in the investigation. Figure 4 shows the mechanism of proposed spline-shaped electrode with curved blade compared to the traditional spline-shaped electrode with the straight blade used in the TACAFM process. Figure 4a shows the top view of the primitive electrode, and Fig. 4b shows the forces acting on the abrasive during the penetration inside the work surface. The resultant force is the vector (centrifugal and Coriolis) and axial force (extrusion and weight of the abrasive). This resultant force results in the angular impingement of abrasive particles on the workpiece surface. This angle is called the approach angle denoted by θ in Fig. 4b. Figure 4c shows the top view of the proposed electrode. Figure 4d shows the working mechanism of the proposed electrode.

Figure 4c shows the top view of the proposed electrode inside the workpiece highlighting the curved blade of the electrode. Figure 4d shows that, due to the curvature of the rod, the resultant force

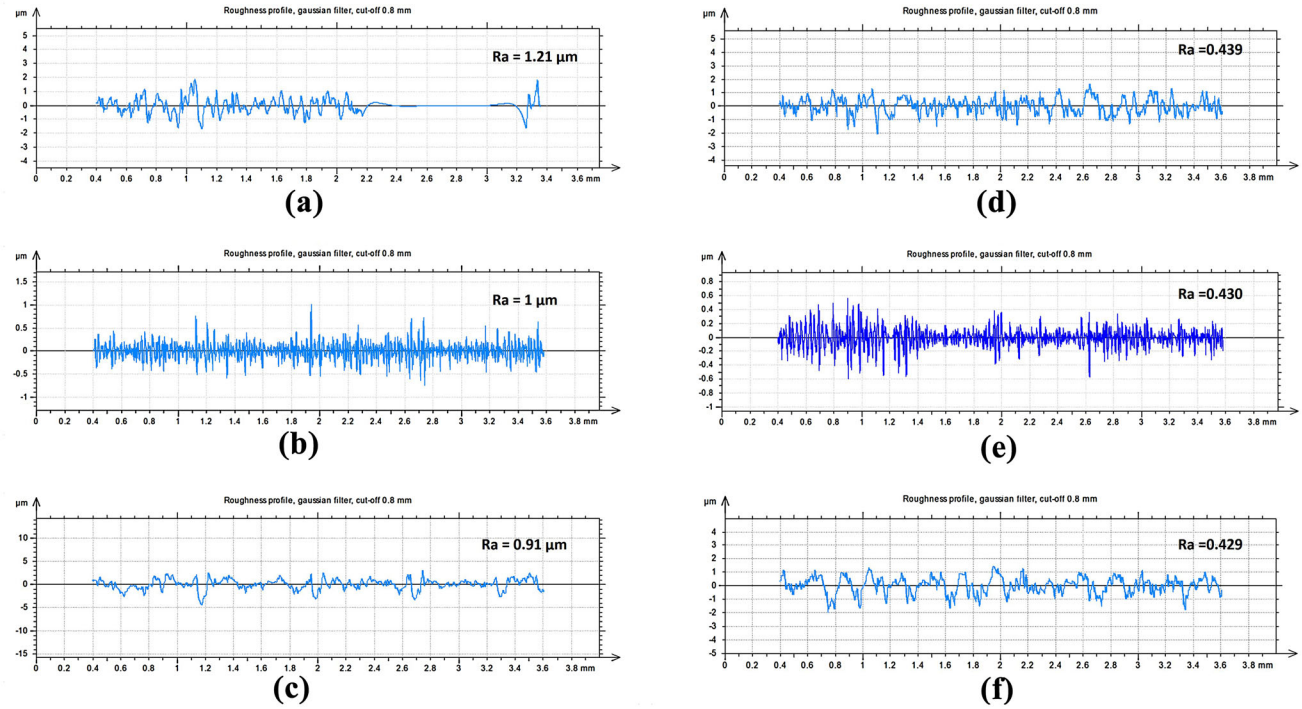


Fig. 3. Surface roughness plots (a), (b) and (c) before machining (d), (e) and (f) plots after machining.

Table III. Experimental results

Parameters	Value	IW (g)	FW (g)	MR (mg)	IR (μm)	FR (μm)	$\%\Delta R_a$
Current	12 Amp	4.22	4.18	40	1.04	0.433	58.36
Duty cycle	0.78						
Rotational speed of CFG Rod	250 rpm						
Pressure	10 MPa						
Abrasive concentration	0.73						

IW initial weight before machining, *FR* final weight after machining, *MR* material removal, *IR* initial surface roughness before machining, *FR* final Surface Roughness after machining.

increases because of the increase in Coriolis and centrifugal force, which increases the angle of impingement of the abrasive particle on the workpiece surface. Thus, the approach angle ϕ is more than θ, due to which total force exerted by the proposed electrode is greater than that of the conventional spline-shaped electrode. The MR by overall TACAFM process is given by Eq. 4[22]

$$\mathrm{MR}_{\mathrm{TACAFM}} = \rho N \left[\frac{d_g^2}{4} \sin^{-1} \frac{2\sqrt{x \cdot (d_g - x)}}{d_g} - \sqrt{x \cdot (d_g - X)} \cdot \left(\frac{d_g}{2} - x \right) \right] \cdot \sum_{i=1}^{2n} \left[1 - \frac{R_a^f}{R_a^i} \right] \cdot L_w + K.V.I.T_{\mathrm{on}} \quad (4)$$

where ρ is the density of the material, N represents the number of abrasive particles indenting on workpiece per stroke, d_g is the diameter of spherical abrasive particles, and x is the length of penetration, n is the number of cycles, R_a^f and R_a^i represent the final and initial surface roughness after and before machining, respectively, L_{w} is the length of the workpiece, K is the constant, V is voltage, I is current, and T_{on} is the pulse on time. Equation 4 shows the MR is a function of the ratio of final and initial surface roughness obtained after and before machining. The value $\left[1 - \frac{R_a^f}{R_a^i}\right]$ is < 1, which results in the decrease in MR. Hence, with a significant increase in $\mathrm{R_a}$ value, a marginal increase in MR is observed.

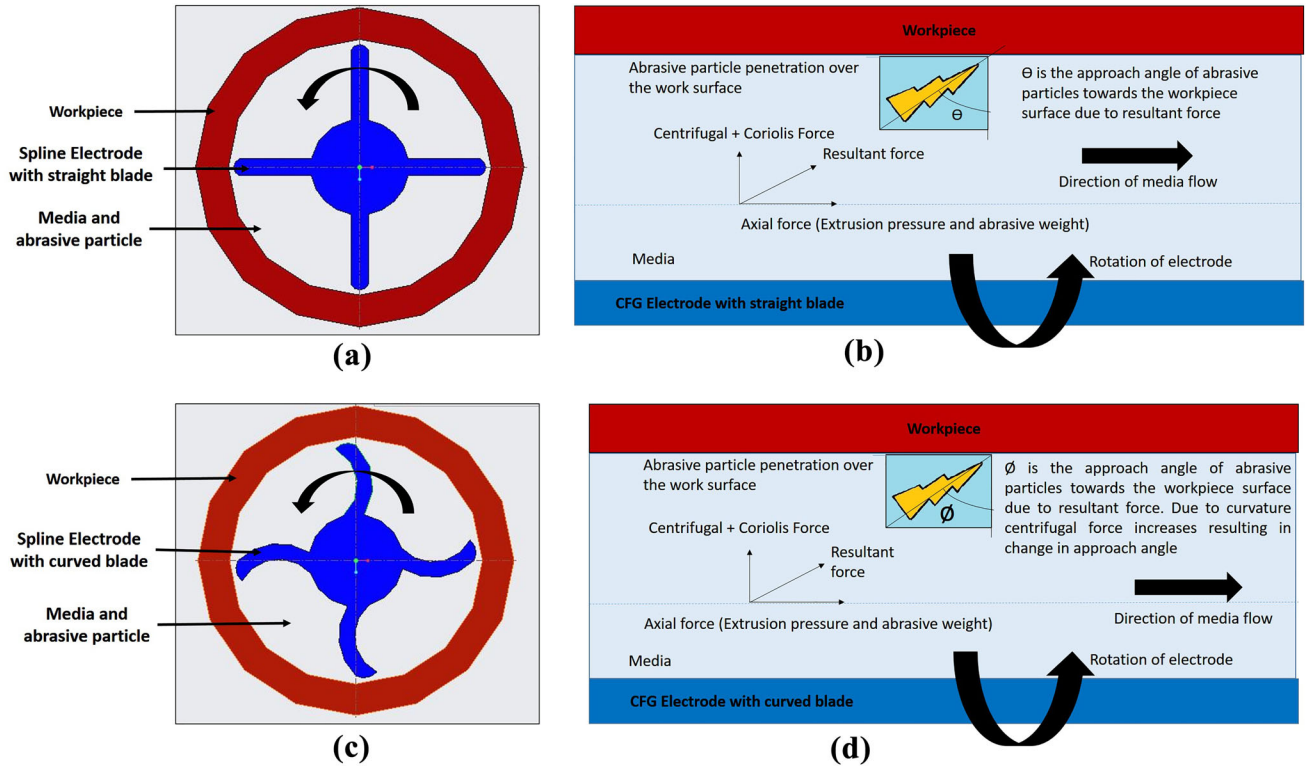


Fig. 4. Material removal mechanism of electrode: (a) spline-shaped electrode with straight blade, (b) mechanism of material removal with primitive tool, (c) top view of the spline-shaped electrode with curved blade, (d) mechanism of material removal by proposed electrode.

As discussed earlier, the TACAFM process is a combination of two machining processes involving thermal and abrasion process. The thermal effect due to spark formation softens the work surface due to which less effort is required for the machining process by the abrasion mechanism. Thus, the desired surface finish and MR are achieved in fewer cycles compared to individual thermal and centrifugal effect where more cycles are required for desired machining, which increases machining time.

MORPHOLOGY OF FINISHED SURFACE

The morphology of the finished surface is discussed with the help of a micrograph and SEM images. The TACAFM process involves the material removal mechanism through melting due to the heat produced by the spark and the abrasion action of the abrasives over the surface of the workpiece because of the extrusion of laden media due to the pressure difference. The rotation of CFG electrode also directs the media particle towards the wall of the workpiece, which enhances the dynamic number of abrasive particles near the wall which in turn enhances the finishing process.[54] Figure 5 shows the micrograph of the finished workpiece without polishing. Figure 5a and b is taken at 100× magnification while Fig. 5c and d is 200× magnification. The figure highlights the machined surface is spark proof. It is represented by a continuous dark band on the micrograph.

Figure 6 shows the micrograph of the polished surface at 100× and 200× magnification. Figure 6a shows the finished surface obtained after machining. The finished surface possesses abrasive marks because of the abrasion action of abrasive particles present in the media with the wall of the workpiece. This abrasion action is due to the resultant force produced by the combination of centrifugal force (Coriolis + tangential) and axial (abrasive weight and axial force), which enhances MR compared to conventional abrasive flow machining processes.[11,22] The black marks shown in Fig. 6b and c show the spark proof which is produced during the TACAFM process when the current is supplied between the electrode and workpiece. Since the spark is involving in the machining process, there is a chance of oxidation of the workpiece surface; also the heat produced by the spark initiates the melting of workpiece which results in the accumulation of some molten metal, which may appeared on the surface of the finished workpiece. However, due to continuous flow of media, the interaction of workpiece surface with atmosphere is the cutoff which limits the formation of oxide layer. Hence, a very less or negligible oxide layer would be seen on the surface of the finished workpiece.[21] Figure 6d shows the surface of the finished workpiece. Also, as the number of extrusion cycle increases, the interaction of abrasive particle with the workpiece increases, which results in the increase in the abrasion

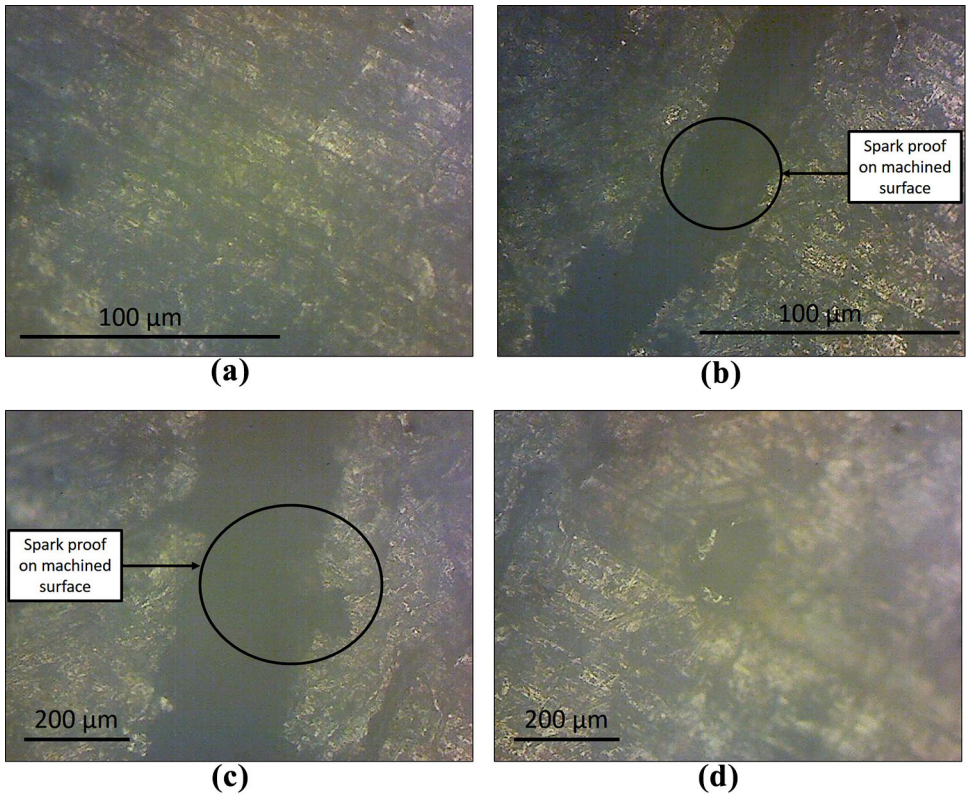


Fig. 5. Micrograph without polishing: (a) and (b) 100× magnification, (c) and (d) 200× magnification.

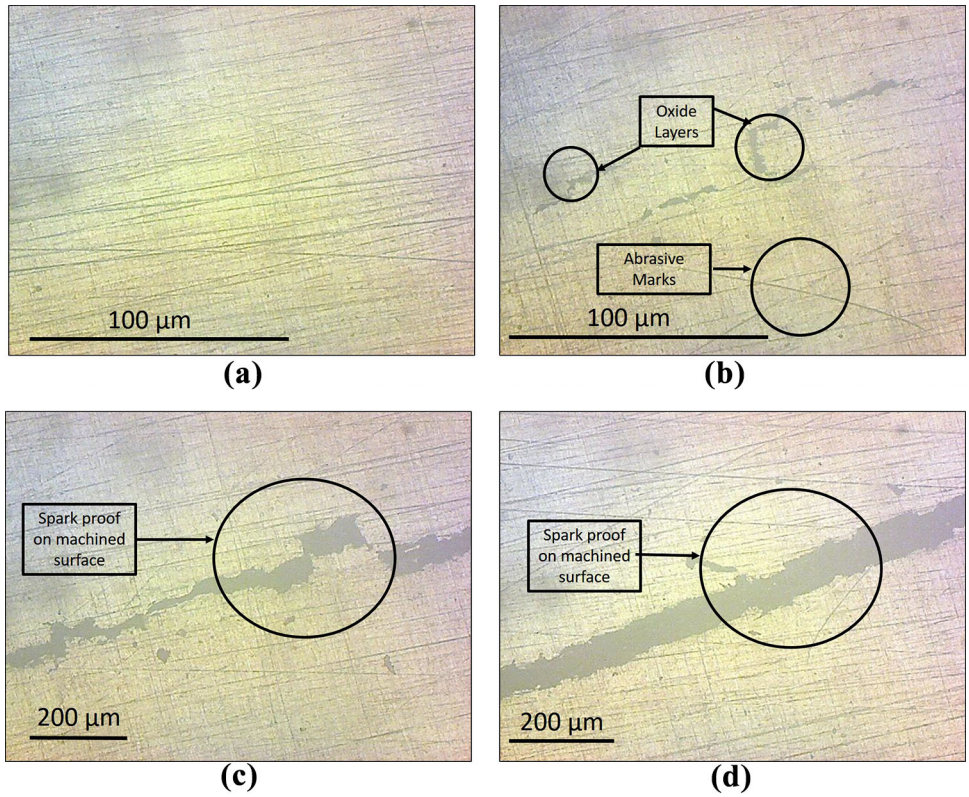


Fig. 6. Polished micrograph (a) and (b) 100× magnification (c) and (d) 200× magnification.

process. This interaction results in the removal of molten metal that is deposited on the finished surface because of melting and the oxide layer (if produce) during the spark action on the workpiece surface.[55] With the further increase in the number of cycles, the abrasive present in the media will become blunt and the abrasion process becomes weak. Thus, the material removal takes place mostly because of the thermal effect, and media flowing over the workpiece act as a polishing agent.[56]

Figure 7 shows the micrograph of the finished brass workpiece with polishing and after application of etchant to view the grain structure and grain boundaries. The ferric chloride and nitric acid were used to produce an effective etchant, and a grain boundary was obtained. Figure 7a and b shows the spark-proof and oxide layer at 100×, and Fig. 7c and d shows the finished machined structure at 200× magnification. Notably, the marks of the oxide layer and spark are spotted, which can be seen in Fig. 7d. Further characterization of the finished surface was done with the help of the SEM images.

To justify the morphology of the finished surface in more detail, the SEM image of the finished surface was done. The optical microscope has limited resolution with smaller depth of field compared with SEM. Also, an important phenomenon may be uncovered because of the reflection of light in the optical microscope. On the other hand, SEM provides the advantage of greater depth with sub-nanometer analysis of the sample and can analyze up to the higher mechanism of 100,000×.

Figure 8 shows the SEM images and XRD analysis of the finished surface. Figure 8a shows the oxide layer, molten metal, spark proof and random abrasive marks over the surface of the finished workpiece representing the involvement of spark and centrifugal action over the workpiece.[22,56,57] Figure 8b shows the uniform parallel arrangement of the spark. This is because of the rotation of the CFG electrode inside the media. The spark produced between the electrode and the workpiece is uniformly travelled around the workpiece because of the rotation of the electrode.[23,51] Also, as the number of cycles increased, the uniformity of surface also increased. This is because with the increase in the number of cycles the abrasive particle interaction with the workpiece surface increases resulting in higher MR; also, the media carry away the removed workpiece particles which also accelerates the abrasion process.[58,59]

The further confirmation of the spark formation was done by considering the XRD of the brass workpiece before and after machining of the workpiece, shown in Fig. 8c. The figure clearly indicates the difference in the XRD pattern before and after machining. The XRD pattern before machining represents the brass alloy, and the change in the pattern after machining is due to the formation of an oxide layer and molten metal, which is possible

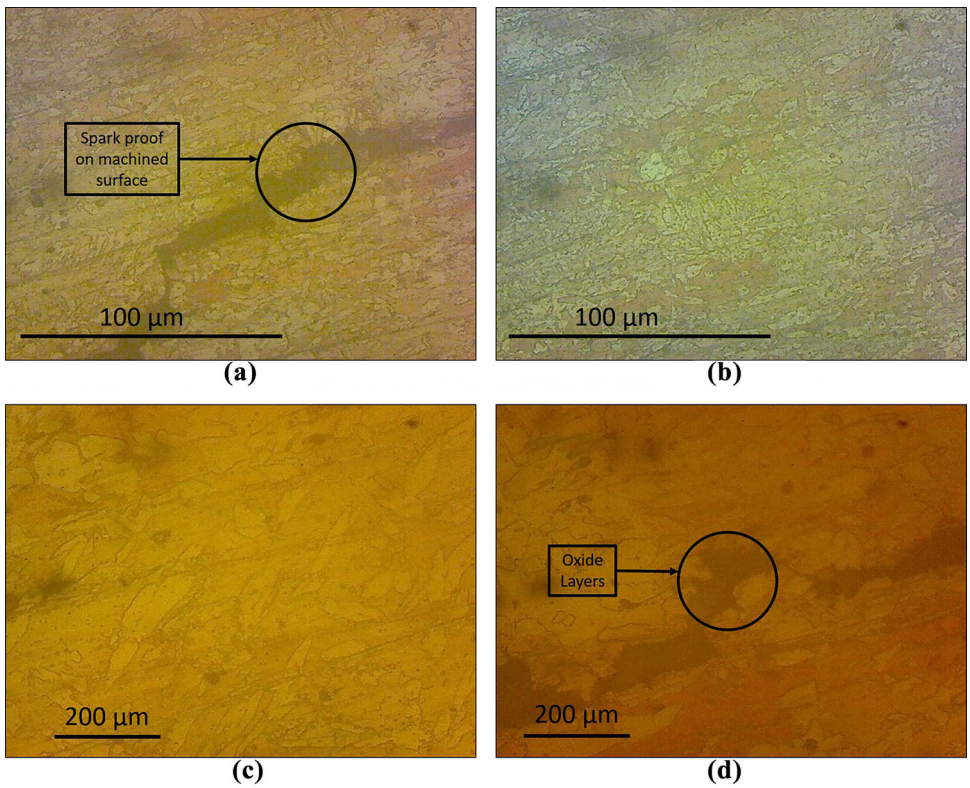


Fig. 7. Polished structure with etchant (a) and (b) 100× magnification (c) and (d) 200× magnification.

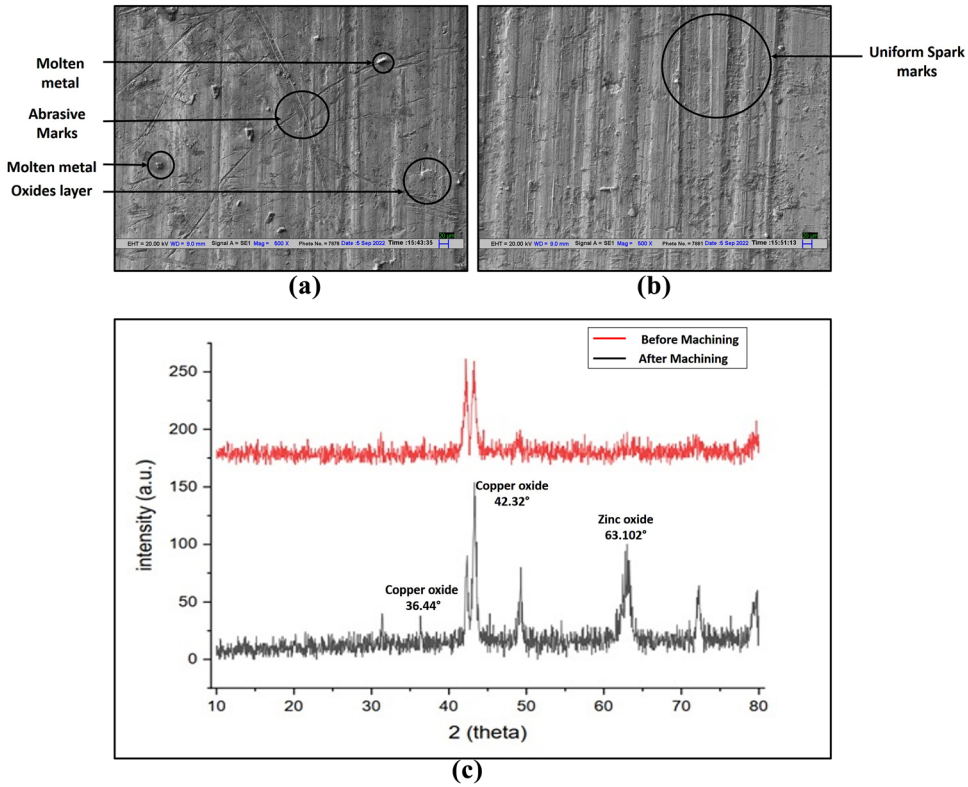


Fig. 8. SEM images and XRD analysis of machined workpiece. (a) SEM images of machined workpiece. (b) SEM image showing spark marks. (c) XRD analysis of machined workpiece.

because of the spark formation by the proposed novel electrode.[23] The XRD pattern for the unmachined workpiece is shown by a red curve, and the XRD pattern for the machined surface is shown with a black curve. The XRD results of the workpiece after machining show the peak at the 2θ value of 42.32°, which correlates with the copper oxide at (200) plane with cubic crystal system. Also, the copper oxide peak may also be seen at the 2θ value of 36.44° corresponding to the (111) plane corresponding to the cubic crystal system.[23,60] The confirmation of oxide layer may also be felt by the presence of zinc oxide, which is found at the 2θ value of 63.102°, which corresponds to the (103) plane corresponding to hexagonal crystal structure.[23,61]

Figure 9 shows the EDAX of the finished workpiece. Figure 9a shows the area of consideration of the finished workpiece; Fig. 9b shows the relevant graph highlighting the peak of elements. Figure 9c shows the major component with their percentage and symbols. The EDAX clearly highlights the presence of oxygen, which is formed because of the presence of an oxide layer due to spark formation.

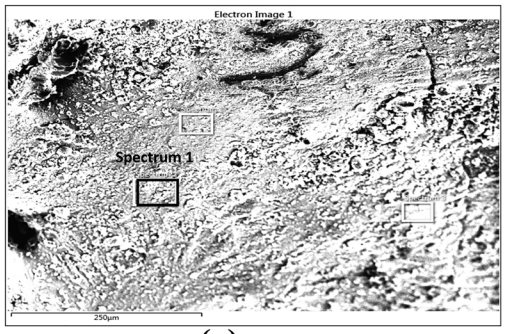
As discussed earlier, the TACAFM process is a combination of sparks along EDM with centrifugal force-assisted abrasive flow machining process, so it takes advantage of material removal by melting and evaporation process. The EDM process is successfully used in the finishing of modern alloy[44,62–64] but the formation of thermal recast layer and oxide layer over the EDM machined surface and slow rate of material removal are concerns regarding the EDM process. In the TACAFM process, the thermal recast layer, molten metal and oxide layer were produced on the surface of the workpiece because the EDM effect is removed by the media particle as the continuous extrusion of media results in the successive penetration of abrasive particles over the surface of the workpiece (which is soft because of the thermal effect) resulting in the increase in abrasion process.[23,51,65]

The other advantage in the TACAFM process is that it uses a leak-proof fixture for holding the workpiece. This is done to prevent the loss of media during the extrusion process. The advantage of this is that it cuts off the supply of oxygen on the internal surface of the workpiece, which restricts the oxide formation over the surface of workpiece.[66] Hence, the surface defects like the recast layer, molten metal and oxide layer are much less in proposed TACAFM process compared to EDM process.

CONCLUSION

The following conclusion can be drawn from the above investigations:

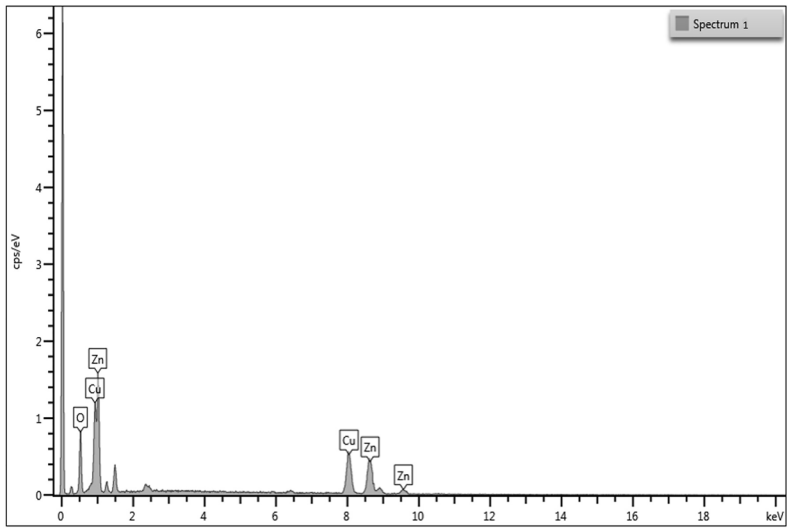
1. TACAFM is a recent hybrid of AFM which improves MR and reduces machining time com-



(a)

Symbols	Elements	% weight
O	Oxygen	15.25
Cu	Copper	44.03
Zn	zinc	40.72
Total		100

(c)



(b)

Fig. 9. EDAX of finished workpiece: (a) area of consideration, (b) EDAX curve, (c) composition of constituents by weight.

pared to the conventional AFM process by combining the advantages of spark energy, centrifugal action and conventional AFM process.

2. The curved blade of the proposed novel electrode increases the centrifugal force on the media compared to straight spline electrode, which enhances the MR and improves the surface finish because of the increase in the effective number of dynamic abrasive particles on the work surface.
3. The value of MR obtained using the proposed novel electrode (curved blade) is 40 mg, which is 8.5% higher than that shown by the spline-shaped electrode with straight blades.
4. The value of percentage improvement in surface finish by the proposed electrode is 58.36%, which is 27.38% higher, which proves the validity of the proposed novel electrode in improving the finishing efficiency of the developed TA-CAFM process.
5. The micrograph, SEM and XRD validate the spark generation and machining efficiency of the TACAFM process and justify the use of the proposed electrode in the developed TACAFM process.
6. The EDAX analysis shows the presence of oxygen element, which verifies the spark formation and validates the TACAFM process.

ACKNOWLEDGEMENTS

The authors thank Delhi Technological University for providing fellowship to cater the financial needs for the research.

CONFLICT OF INTEREST

On behalf of all authors, the corresponding author states that there is no conflict of interest.

REFERENCES

1. P. Chakraborty and S.P. Vejendla, *Mater. Today Proc.* 64, 1380 (2022).
2. A. Baraq, J.S. Breast, J. Abrantes-de-sampaio, F. Moitrier, and Y. Demarte, *Forces Mech.* 4, 100030 (2021).
3. X. Wu, H. Luo, M. Sheng, H. Liu, Z. Xiao, and H. Geng, *JOM* 69, 184 (2017).
4. K. Choi and H. Park, *JOM* 74, 878 (2022).
5. W.B. Qurix, L.C. Edomwonyi-Out, J.K. Aremu, G.J. Ibeh, S.G. Bawa, and D. Jise, *J. Mater. Res. Technol.* 9(5), 9831 (2020).
6. V.K. Jain, *J. Mater. Process. Technol.* 209, 6022 https://doi.org/10.1016/j.jmatprotec.2009.08.015 (2009).
7. P. Ali, S. Dhull, R.S. Walia, Q. Murtaza, and M. Tyagi, *Mater. Today Proc.* 4, 7208 https://doi.org/10.1016/j.matpr.2017.07.048 (2017).
8. A. Goyal, H. Singh, R. Goyal, R. Singh, and S. Singh, *Mater. Today Proc.* 56, 3065 https://doi.org/10.1016/j.matpr.2021.12.109 (2022).
9. S. Singh, H. Kumar, S. Kumar, and S. Chaitanya, *Mater. Today Proc.* 56, 3108 https://doi.org/10.1016/j.matpr.2021.12.273 (2022).
10. A. Bhardwaj, P. Ali, R.S. Walia, Q. Murtaza, S.M. Pandey, K. Shanker, R. Shankar, and R. Sindhwani, *Advances in Industrial and Production Engineering* (Springer, Singapore, 2019), pp41–67. https://doi.org/10.1007/978-981-13-6412-9_5.
11. R.S. Walia, H.S. Shan, and P. Kumar, *Int. J. Adv. Manuf. Technol.* 39, 1171 https://doi.org/10.1007/s00170-007-1301-8 (2008).
12. R.S. Walia, H.S. Shan, and P. Kumar, *Mach. Sci. Technol.* 10(3), 341 https://doi.org/10.1080/10910340600902157 (2006).
13. B.S. Brar, R.S. Walia, V.P. Singh, and M. Sharma, *J. Inst. Eng. India Ser. C.* 94, 21 https://doi.org/10.1007/s40032-012-0054-9 (2013).
14. M.R. Sankar, S. Mondal, J. Ramkumar, and V.K. Jain, *Int. J. Adv. Manuf. Technol.* 42, 678 https://doi.org/10.1007/s00170-008-1642-y (2009).

15. R. Sankar, V.K. Jain, and J. Ramkumar, *Int. J. Mach. Tools Manuf* 50, 637 https://doi.org/10.1016/j.ijmachtools.2010.03.007 (2010).
16. S. Singh and H.S. Shan, *Int. J. Mach. Tools Manuf* 42, 953 https://doi.org/10.1016/S0890-6955(02)00021-4 (2002).
17. R. Singh, R.S. Walia, and M.N. Suri, *IJRET* 04, 382 https://doi.org/10.15623/ijret.2015.0408065 (2015).
18. B.S. Brar, R.S. Walia, and V.P. Singh, *Int. J. Adv. Manuf. Technol.* 79, 329 https://doi.org/10.1007/s00170-015-6806-y (2015).
19. G. Venkatesh, A.K. Sharma, and P. Kumar, *Int. J. Mach. Tools Manuf* 89, 29 https://doi.org/10.1016/j.ijmachtools.2014.10.014 (2015).
20. R. Vaishya, R.S. Walia, and P. Kalra, *Mater. Today Proc.* 2, 3327 https://doi.org/10.1016/j.matpr.2015.07.158 (2015).
21. P. Ali, R.S. Walia, Q. Murtaza, and M.S. Ranganath, *Surf. Topogr. Metrol. Prop.* 8, 035013 https://doi.org/10.1088/2051-672X/abaffe (2020).
22. P. Ali, R.S. Walia, Q. Murtaza, and M.S. Ranganath, *J. Braz. Soc. Mech. Sci. Eng.* 42, 302 https://doi.org/10.1007/s40430-020-02375-6 (2020).
23. P. Ali, R.S. Walia, Q. Murtaza, and M.S. Ranganath, *IJEMS* 29(1), 29 (2022).
24. P. Ali, R.S. Walia, Q. Murtaza, and M.S. Ranganath, *Int. J. Mech. Prod. Eng. Res. Dev.* 8(6), 635 (2018).
25. P. Ali, R.S. Walia, Q. Murtaza, and M.S. Ranganath, *Int. J. Mech. Prod. Eng. Res. Dev.* 9(2), 415 (2019).
26. Y. Fu, H. Gao, Q. Yan, and X. Wang, *Precis. Eng.* 60, 497 https://doi.org/10.1016/j.precisioneng.2019.08.011 (2019).
27. R. Baraiya, A. Babbar, V. Jain, and D. Gupta, *J. Manuf. Process.* 50, 266 https://doi.org/10.1016/j.jmapro.2019.12.051 (2020).
28. A. Bhardwaj, K. Srinivas, and R. Chaudhary, *JER Special Issue*. https://doi.org/10.36909/jer.ICARI.15291 (2021).
29. A. Bhardwaj, K. Srinivas, and R. Chaudhary, *Int. J. Adv. Prod. Ind. Eng.* 6(2), 01 https://doi.org/10.35121/ijapie202104251 (2021).
30. S. Krovvidi and A. Bhardwaj, *IJAPIE* 3(4), 15 (2018).
31. A. Bhardwaj, computational and experimental analysis of parameters in centrifugal force assisted abrasive flow machining process, M. Tech Diss. 2019, Delhi Technological University.
32. S. Mohseni-Mofidi, L. Pastewka, M. Teschner, and C. Bierwisch, *Appl. Math. Model.* 101, 38 https://doi.org/10.1016/j.apm.2021.07.015 (2022).
33. A.W. Hashmi, H.S. Mali, A. Meena, K.K. Saxena, A.P.V. Puerta, C. Prakash, D. Buddhi, J.P. Davim, and D.S. Abdul-Zahra, *Metals* 12, 1328 https://doi.org/10.3390/met12081328 (2022).
34. P. Ali, S.M. Pandey, M.S. Ranganath, R.S. Walia, and Q. Murtaza, *Measurement* 151, 107133 https://doi.org/10.1016/j.measurement.2019.107133 (2020).
35. Z. Zhang, X. Zhang, X. Liu, P. Zhou, P. Yao, C. Huang, H. Liu, H. Zhu, and B. Zou, *CIRP Bio. Manuf. Conf.* 89, 282 https://doi.org/10.1016/j.procir.2020.05.153 (2020).
36. S. Ahmad, R.M. Singari, and R.S. Mishra, *Trans. IMF* 99(2), 94 https://doi.org/10.1080/00202967.2021.1865644 (2021).
37. N. Dixit, V. Sharma, and P. Kumar, *J. Manuf. Process.* 67, 101 https://doi.org/10.1016/j.jmapro.2021.04.053 (2021).
38. P. Yadav and S.C. Jayswal, *Mater. Today Proc.* 18, 3017 https://doi.org/10.1016/j.matpr.2019.07.173 (2019).
39. K.K. Kar, N.L. Ravikumar, P.B. Tailor, J. Ramkumar, and D. Sathiyamoorthy, *J. Mater. Process. Technol.* 209, 2212 https://doi.org/10.1016/j.jmatprotec.2008.05.012 (2009).
40. S.M. Basha, N. Venkaiah, and M.R. Sankar, *Addit. Manuf.* 61(2023), 103290 https://doi.org/10.1016/j.addma.2022.103290 (2023).
41. M.A. Manjunath, P. Vinod, N. Balashanmugam, and M.R. Sankar, *Mater. Today Proc.* 26, 1113 https://doi.org/10.1016/j.matpr.2020.02.223 (2020).
42. G.G. Krishna and N. Jinesh, *Mater. Today Proc.* 58(1), 335 https://doi.org/10.1016/j.matpr.2022.02.240 (2022).
43. E. Uhlmann, V. Mihotovic, S. Roßkamp, and A. Dethlefs, *Proc. CIRP* 46, 51 https://doi.org/10.1016/j.procir.2016.03.172 (2016).
44. H.J. Tzeng, B.H. Yan, R.T. Hsu, and H.M. Chow, *Int. J. Adv. Manuf. Technol.* 34, 649 https://doi.org/10.1007/s00170-006-0632-1 (2007).
45. S. Han, F. Salvatore, J. Rech, and J. Bajolet, *Precis. Eng.* 64, 20 https://doi.org/10.1016/j.precisioneng.2020.03.006 (2020).
46. S. Han, F. Salvatore, J. Rech, J. Bajolet, and J. Courbon, *J. Manuf. Process.* 59, 248 https://doi.org/10.1016/j.jmapro.2020.09.065 (2020).
47. C.W. Kuma, C.H. Wu, S. Wan, and C.W. Kang, *J. Mater. Process. Tech* 282, 116704 https://doi.org/10.1016/j.jmatprotec.2020.116704 (2020).
48. A.W. Hashmi, H.S. Mali, and A. Meena, *Mater. Today Proc.* 56, 1369 https://doi.org/10.1016/j.matpr.2021.11.428 (2022).
49. G. Venkatesh, A.K. Sharma, N. Singh, and P. Kumar, *Proc. Eng.* 97, 320 https://doi.org/10.1016/j.proeng.2014.12.255 (2014).
50. A. Bhardwaj, K. Srinivas, and R. Chaudhary, *Mapan* 38, 459 https://doi.org/10.1007/s12647-023-00626-0 (2023).
51. A. Bhardwaj, K. Srinivas, and R. Chaudhary, *Proc. Inst. Mech. Eng. Part E J. Process Mech. Eng.* https://doi.org/10.1177/09544089231195171 (2023).
52. A. Amjadimanesh, H. Ajam, and H. Nezhad, *J. Serb. Soc. Comp. Mech.* 8, 23 https://doi.org/10.5937/jsscm1401023A (2014).
53. Z. Wang and C. Zhu, *Int. J. Aerosp. Eng.* 2017, 1–9 https://doi.org/10.1155/2017/8695170 (2017).
54. R.S. Walia, H.S. Shan, and P.K. Kumar, *Int. J. Adv. Manuf. Technol.* 38, 1157 https://doi.org/10.1007/s00170-007-1184-8 (2008).
55. R.S. Walia, H.S. Shan, and P.K. Kumar, *Int. J. Adv. Manuf. Technol.* 44, 700 https://doi.org/10.1007/s00170-008-1893-7 (2009).
56. R.S. Walia, H.S. Shan, and P. Kumar, *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.* 220(11), 1775 https://doi.org/10.1243/09544054JEM325 (2006).
57. R.S. Walia, H.S. Shan, and P. Kumar, *Mater. Manuf. Process.* 21(4), 375 https://doi.org/10.1080/10426910500411645 (2007).
58. R. Singh and R.S. Walia, *Int. J. Res. Mech.* 2(1), 34 (2012).
59. D. Bähre, H. Brünnet, and M. Swat, *Proc. CIRP* 1, 419 https://doi.org/10.1016/j.procir.2012.04.075 (2012).
60. C.J. Newton and H.C. Vacher, *J. Res. Natl. Bureau Standards-C Eng. Instrum.* 65(1), 57 (1960).
61. L. Babouri, K. Belmokre, A. Kabir, A. Abdelouas, R. Khettabi, and Y. El Mendili, *Mater. High Temp.* 36(2), 165 https://doi.org/10.1080/09603409.2018.1499243 (2018).
62. M.K. Dikshit, S. Singh, V.K. Pathak, K.K. Saxena, M.K. Agrawal, V. Malik, K. Hazim Salem, and M.I. Khan, *J. Mater. Res. Technol.* 24, 223 https://doi.org/10.1016/j.jmrt.2023.03.005 (2023).
63. S. Patil, M. Patil, R. Kulkarni, and V. Malik, *Adv. Mater. Process. Technol.* https://doi.org/10.1080/2374068X.2023.2184313 (2023).
64. S. Patil, R. Kulkarni, M. Patil, and V.R. Malik, *Adv. Mater. Process. Technol.* https://doi.org/10.1080/2374068X.2023.2194519 (2023).
65. K.C. Cheng, A.C. Wang, K.Y. Chen, and C.Y. Huang, *Processes* 10(7), 1362 https://doi.org/10.3390/pr10071362 (2022).
66. A.W. Hashmi, H.S. Mali, A. Meena, N. Gupta, S. Ahmad, K.K. Saxena, and V. Malik, *Surf. Rev. Lett.* 30(06), 2350034 https://doi.org/10.1142/S0218625X23500348 (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.

J Mater Sci: Mater Electron **(2023) 34:2191**

Novel orange-red-emitting $Li_2Ba_5(WO_5)_3:Eu^{3+}$ phosphor for w-LEDs and non-contact thermometry applications

Anu[1], and A. S. Rao[1,*]

[1] *Department of Applied Physics, Delhi Technological University, Bawana Road, New Delhi 110042, India*

Received: 14 September 2023
Accepted: 6 November 2023
Published online: 20 November 2023

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

ABSTRACT

Novel Eu^{3+}-activated orange-red phosphors $Li_2Ba_{(5-x)}(WO_5)_3:xEu^{3+}$ (LBW:xEu^{3+}) (here, $1 \leq x \leq 9$, $\Delta x = 2$, and x is the mol%) were successfully prepared via high-temperature solid-phase reaction. The X-ray diffraction (XRD) patterns were recorded for structural analysis and phase identification of the synthesized phosphors. Morphological analysis was done using scanning electron microscopy (SEM). The optical bandgap value has been measured using diffuse reflectance spectra (DRS). The photoluminescence (PL) emission spectra recorded for the LBW:xEu^{3+} phosphors under 322 and 396 nm excitation reveal strong visible orange-red emission at 595 nm corresponding to $^5D_0 \rightarrow {}^7F_1$ transition of Eu^{3+} ions. The CIE chromaticity coordinates were estimated for all the phosphors and were found to be in the deep orange-red region. The PL decay spectral profiles recorded at 595 nm emission when excited with 322 nm light showed a consistent exponential pattern. The experimental lifetimes were observed to decrease as the concentration of Eu^{3+} ions increased. In order to study the luminescence intensity ratio (LIR) of the paired emissions lines of Eu^{3+} as a function of temperature, PL measurements were carried out in the temperature range of 298–448 K. With a maximum sensitivity of nearly 1.21% K^{-1} at 298 K, the synthesized phosphors are relatively more sensitive at low temperatures. The aforementioned results represent that the synthesized LBW:Eu^{3+} phosphor samples can be employed in non-contact optical thermometry as well as in w-LED applications.

1 Introduction

Phosphors are substances which absorb energy and then release it as UV, visible, or IR light. They are made up of an inorganic host that has been doped with activator ions; these ions are usually found in transition metals or rare earths (RE). RE-doped phosphors in particular are significant materials with numerous applications in artificial lighting, X-ray medical radiography, display technology, and high-power solid-state lasers [1]. Phosphor thermometry is a non-contact method for measuring temperature remotely using light signals. Thermographic phosphors (TGPs) are phosphors created especially for phosphor thermometry. TGPs are resistant to oxidation in high-temperature situations and do not react with harsh chemicals,

Address correspondence to E-mail: drsrallam@gmail.com

in contrast to standard techniques of monitoring temperature (thermocouples, thermistors, thermal paints). The type of RE impurity and the host both affect the temperature-dependent changes in the TGPs' emission properties. Both the decay observations and the intensity ratio measurements serve as the foundation for phosphor thermometry. The latter has gained a lot of acceptance since it eliminates the possibility of several mistakes caused by variations in the excitation light source, temperature changes in the excitation bands, and uneven dopant concentrations [2] and has been used in this study. It has recently been difficult to get fluorescence thermal sensing technology results for reliable temperature estimation. As a result, the need for appropriate rare-earth ion-doped optical temperature-sensing materials exists. Additionally, suitable methodologies should be employed to establish the relation between temperature and fluorescence characteristics. Although many rare-earth ions, including Er^{3+}, Pr^{3+}, Yb^{3+}, Gd^{3+}, and Ho^{3+}, can be employed for optical temperature sensing [3–6] their poor luminescence efficiency reduces the precision of temperature measurements. The luminous phosphor that is doped with Eu^{3+} ions typically has a high luminescence efficiency [7–9]. Moreover, energy levels of $^5D_0 \rightarrow {}^4F_1$ and $^5D_0 \rightarrow {}^4F_2$ of Eu^{3+} can be used as the thermal coupling energy levels (TCLs).

Moreover, now-a-days there is a lot of focus on luminous rare-earth (RE) ion-doped oxide materials with excellent thermal and chemical stability. These materials are widely utilized in lighting equipment, solid-state lasers, agriculture fields, medical devices, displays, and lighting devices [10–12]. In all these materials, the phosphor-converted white light-emitting diodes (pc w-LEDs) are more likely to replace traditional incandescent and fluorescent lamps than any other solid-state lighting (SSL) source due to their advantages like environmental friendliness, durability, high luminous efficiency, low power consumption, prolonged operation time, high brightness, and high-temperature tolerance [13]. W-LEDs use lanthanide-doped phosphors that can be stimulated by the UV/n-UV or blue LED chip and are coated on the walls of the encapsulated tube or on the LED chip. The typical behavior of lanthanides is to get excited by absorbing UV/n-UV radiation and then emit visible light. The f-block series of elements in the periodic table are referred to as rare earths (RE) [14].

At present, the approach used to create white light was developed by Nichia Chemicals Co. in 1996, which involves coating blue chips with yellow phosphor (YAG: Ce) (InGaN). However, due to the lack of a red component and poor reproducibility, w-LEDs created using this approach have a low color rendering index (CRI). Red, green, and blue (RGB) LEDs combined with RGB phosphor-coated UV-LED chips could resolve this problem. Although this method increases CRI, it comes at the cost of complicated manufacturing methods (extra RGB LED components that require variable driving current) or rather a low efficiency in the case of UV-excited tri-color phosphor LEDs. Due to their restricted applicability and lack of an efficient red-emitting phosphor, these approaches are less popular [15–18]. Also, these techniques have a high Corrected Color Temperature (CCT > 7000 K) due to the absence of a red component, which also restricts their range of uses [19, 20]. The single-phase white light-emitting phosphor, which is typically triggered by UV radiation, can be used to combine several activator ions in the same host matrix to address these drawbacks. These luminescent materials are excellent alternatives for the use in LEDs and w-LEDs due to high CRI, good luminescence efficiency, low CCT, and lower manufacturing costs [21].

Recently, many studies have been conducted on the introduction of RE^{3+} activator ions into various hosts for usage as w-LEDs, resulting in substantial advancements. But due to a number of drawbacks, such as poor chemical and thermal stability, and poor fluorescence efficiency, orange and red phosphors still need to be improved [22–24]. RE^{3+} ions-doped phosphors have drawn the considerable attention of researchers as such studies reveal the synthesis of phosphors followed by the information pertaining to the physical processes of energy absorption and relaxation. Further, due to their employment as luminous optical materials emitting in the visible and NIR regions, the absorption and emission of RE-activated crystalline hosts via f–f transition are of great interest [25]. Due to the remarkable luminous characteristics and narrow emission bands, the RE-activated inorganic materials have received significant attention [26, 27]. Amid all the RE^{3+} ions, trivalent europium (Eu^{3+}) can act as a strong activator [28–30]. Also, oxide phosphors are more stable than sulfide phosphors. Oxide-based phosphors have received a lot of attention in recent decades as potential candidates for use in w-LED applications [27].

Recently, Parauha et al. synthesized Eu^{3+} ions-doped Na_2CaSiO_4 phosphor that exhibits prominent

orange and red emission bands due to the $^5D_0 \rightarrow {}^7F_{1,2}$ transitions, potentially leading to applications in lighting and display devices [31]. S. Wang and his co-workers prepared a thermally stable $Y_2SiWO_8:Eu^{3+}$ for w-LEDs [32]. Shi et al. suggested blended phosphors for white light by combining ($Gd_6WO_{12}:Eu^{3+}$) red and ($YAG:Ce^{3+}$) yellow phosphors to enhance the color rendering of w-LEDs [33]. The aforementioned research work in the Eu^{3+} ions-activated tungstate phosphors motivated us to explore this field.

Now, selecting the appropriate host medium is also essential when preparing luminous materials for a variety of applications. The phosphor host materials typically consist of tungstates, borates, phosphates, aluminates, silicates, molybdates, vanadates, etc. Since tungstates have strong covalent interaction, high stability, and strong visible luminescence, they are suitable host materials for RE^{3+} ions and have potential uses in lasers, optical fibers, w-LEDs, plasma display panels, quantum electronics, phosphors, scintillators, and solid-state optoelectronic devices [34, 35]. Also, the alkali metal ions, for instance Li^+, Na^+, K^+, etc. improve the luminescence emission intensity and color purity of the host matrix. A few literature [36, 37] show that due to presence of Li^+ ions in the host matrix, a better luminescence property is achieved instead of other alkali ions making them suitable as modifier in the host matrix. Additionly, tungstate with alkaline-earth metal ions (i.e., Ba^{2+}, Sr^{2+}, Ca^{2+}, Mg^{2+} etc.) are a fascinating group of compounds that exhibit blue or green luminescence associated with the tetrahedral WO_4 groups [38]. These substances find extensive applications as X-ray phosphors, scintillators, laser crystals, and matrices for electron spin resonance [39].

To the best of our knowledge, there has been no report on the structural and luminescence properties of Eu^{3+} ions-doped $Li_2Ba_5(WO_5)_3$ (LBW) phosphors. In the present work, we have reported the synthesis, structural, morphological, and PL characteristics of Eu^{3+} ions-activated LBW phosphors for non-contact optical thermometry as well as in w-LED applications. A high-temperature solid-state reaction method was used to synthesize the phosphor samples under investigation and further characterized using various techniques such as X-ray diffraction (XRD), scanning electron microscopy (SEM), energy-dispersive x-ray analysis (EDAX), thermogravimetric analysis (TGA), PL, time-resolved PL (TR-PL), and temperature-dependent PL (TD-PL). The optical bandgap was evaluated using DRS and the Kubelka–Munk function. The efficient excitation and emission bands are obtained. The TR-PL and CIE chromaticity coordinates of the prepared phosphor samples have been studied to understand the utility of the titled phosphors in n-UV/blue-pumped w-LEDs. The FIR and relative sensitivity parameters were evaluated using the TD-PL data, which approves its suitability for non-contact optical thermometry.

2 Experimental studies

2.1 Synthesis

The LBW phosphors doped with Eu^{3+} ions were synthesized using high temaperature solid-phase reaction method. The raw constituents $BaCO_3$ (98%), Li_2CO_3 (98%), WO_3 (99.9%), and Eu_2O_3 (99.9%) having high purity were used for the phosphor synthesis. Using an electric balance, the raw elements for all compositions were taken in proper stoichiometric ratio and mixed in an agate mortar using pestle for 1 h to get a homogeneous mixture. These mixtures were shifted in alumina crucibles followed by sintering at 600 °C for 2 h in an electric high-temperature (HT) furnace. After the first heating, the mixtures were ground thoroughly to maintain homogeneity and then heated again in the electric HT furnace at 1200 °C for 6 h. When the reaction was complete, the phosphor components were naturally cooled to room temperature (RT). Finally, the prepared $LBW:xEu^{3+}$ phosphor samples were ground one more time for characterizations. Figure 1 represents all the steps involved during the preparation of the phosphor samples by using solid-state reaction method.

2.2 Instrumentation

The structural analysis was performed using X-ray diffraction (XRD) method on a Bruker D-8 Model Advance diffractometer. The instrument was equipped with a Nickel-based filter and used a CuKα radiation source with a wavelength of $\lambda = 1.54$ Å. The analysis was conducted within the range of $20° \leq 2\theta \leq 70°$ for the 2θ angle. The thermal behavior of the materials was examined via thermogravimetric analysis (TGA) using TGA/DSC/DTA equipment from Perkin Elmer. JSM-661OLV SEM was used to investigate the SEM and EDAX elemental analyses. The analysis of

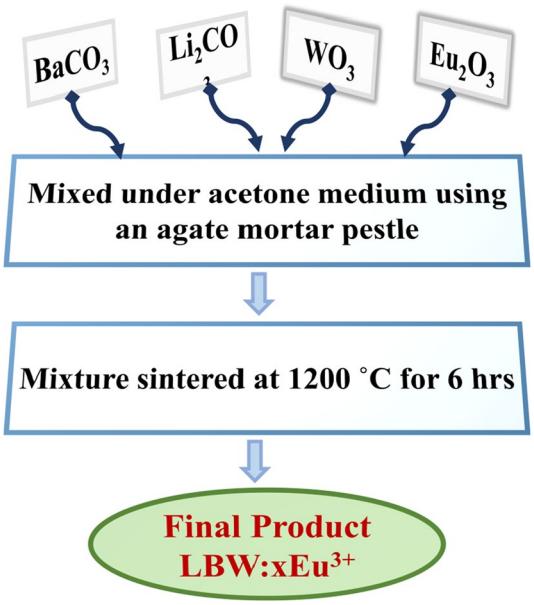


Fig. 1 Synthesis procedure for LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ and x is the mol%) phosphors using solid-state reaction method

diffuse reflectance spectroscopy (DRS) was carried out using a Jasco V-770 spectrophotometer with a resolution of ± 2.0 nm. The instrument utilized both deuterium and halogen light sources. To validate the photoluminescence (PL) excitation and emission spectra, a Jasco 8300 FP spectrofluorophotometer with a 300 W Xenon lamp as the excitation source and a resolution of ± 1.0 nm was employed. The PL decay measurement was conducted using the Edinburgh FL920 Fluorescence Spectrometer. All the above measurements were conducted at room temperature. The temperature-dependent PL characteristics were recorded using FLAME-S-XR1-ES having an ocean optics spectrometer and a heating system component.

3 Results and discussion

3.1 Structural analysis

3.1.1 XRD measurements

The X-ray diffraction (XRD) patterns were utilized to analyze the crystalline structure and phase purity of the prepared phosphor samples. The XRD patterns of un-doped LBW and LBWEu9.0 phosphor samples follow the standard JCPDS data (01-072-1717), displayed in Fig. 2a. The recorded XRD data support that the as-prepared phosphor samples are crystallized in Hexagonal, with $P6_3mc$ space group (space group number is 186) and the lattice parameters of $a = 5.760$ Å $= b$, $c = 23.742$ Å, $\alpha = \beta = 90°$ and $\gamma = 120°$ and $V = 682.17$ Å^3. The diffraction patterns for the LBW samples doped with Eu^{3+} ions are alike to the standard data, indicating that the Eu^{3+} ions were efficiently incorporated in the host material without causing any structural alterations. It is also required to determine the radius percentage difference (D_r) between substituted ions and activator ions in order to better understand the dopant Eu^{3+} ions occupying Ba^{2+} sites in the LBW host lattice. The value of D_r is defined by the following expression [31]:

$$D_r = \frac{R_1(\mathrm{CN}) - R_2(\mathrm{CN})}{R_1(\mathrm{CN})} \times 100\%, \quad (1)$$

here, R_1(CN)(= 1.35Å) and R_2(CN)(= 0.95Å) are the radii of Ba^{2+} ions and activated Eu^{3+} ions, respectively. The value of D_r was determined using Eq. (1) to be 29.85%, indicating that Eu^{3+} could effectively replace Ba^{2+} ion positions in LBW host lattice [31, 40].

The following Debye–Scherrer equation is used to evaluate the crystallite size (D) using the diffraction data [41]:

$$D = \frac{K\lambda}{\beta_{\mathrm{hkl}} \cos\theta} \quad (2)$$

here, K denotes the shape factor and λ (= 1.54 Å) is the X-ray (CuK$_\alpha$ radiation) wavelength. K can have values anywhere from 0.62 to 2.08. Here, we use $K = 0.94$. The values of FWHM (β) and Braggs diffraction angle (θ) were resulting from the sample XRD patterns to evaluate the crystallite size. The average crystallite size is nearly 83 nm.

Additionally, peak shifting and line broadening in XRD spectral patterns are due to the existence of microstrain in the phosphors. Williamson–Hall (W–H) equation was used to determine the strain present in phosphors [42]:

$$\beta_{\mathrm{hkl}} = \frac{K\lambda}{D\cos(\theta)} + 4 \in \tan(\theta) \quad (3)$$

By rearranging the above equation, we get

$$\beta_{\mathrm{hkl}} \cos(\theta) = \frac{K\lambda}{D} + 4 \in \sin(\theta), \quad (4)$$

where K and D are constants and signify the strain-induced broadening in the powdered materials, β is the full width half maxima. A linear fit graph between

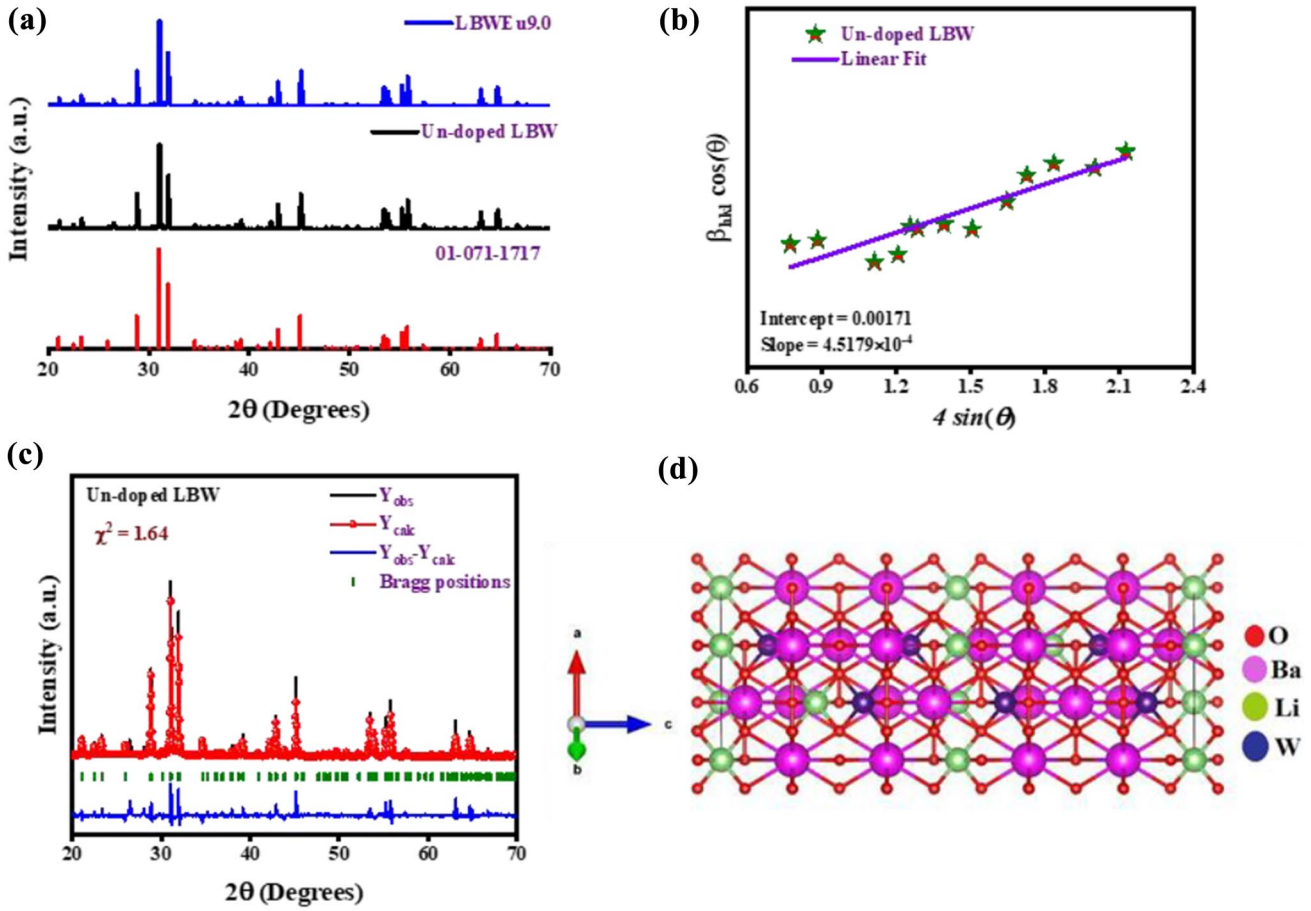


Fig. 2 **a** The XRD patterns of un-doped LBW and LBWEu9.0 phosphors, **b** a linear fit graph between βcosθ (y-axis) and 4sinθ (x-axis), **c** rietveld refinement of un-doped LBW sample, **d** crystal structure of LBW phosphor

$\beta_{hkl}\cos(\theta)$ (y-axis) and $4\sin(\theta)$ (x-axis) is shown in Fig. 2b. Crystallite size was determined from the intercept of the linearly fitted plot and microstrain from the slope value. The values of D and $\in$ were found to be nearly equal to 84 nm and 4.51×10^{-4}, respectively. The values of crystallite size computed using the above-discussed two methods varied slightly. This is because the strain component was assumed to be zero in Scherrer's formula and the apparent broadening of the XRD peaks was exclusively attributed to grain-size reduction. The un-doped LBW phosphor sample undergoes Rietveld refinement using the FullProf Suite programme in order to investigate more thoroughly. The backgrounds were improved using fifteen cycles of linear interpolation, and for each composition, pseudo-Voigt function was employed to fit the experimental data. Figure 2c shows the Rietveld refinement map for un-doped LBW sample and Table 1 lists the related crystal structure parameters.

The black line in the refined map for each composition represents the observed intensity data, the red line with a circle represents the refined calculated intensity, the blue line illustrates the difference between the calculated and the observed intensities, and the green perpendicular line symbolizes the positions of Bragg's peaks. The goodness value for the refined data falls between 1 and 2, is the most appropriate. The VESTA software was used to analyze the hexagonal crystal structure for the host lattice, shown in Fig. 2d.

Table 1 Refinement parameters for un-doped LBW phosphor

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
$\chi 2$	–	1.64

3.1.2 SEM study

A key element for successful application in commercial w-LEDs is the surface morphology. The SEM and EDAX data were recorded to completely analyze the morphological behavior of the as-prepared phosphor samples. The SEM images of the as-prepared LBW:xEu^{3+} ($x = 0$, 3) phosphor sample was taken at distinct scales and resolutions to examine the morphological aspects and have been shown in Fig. 3a, b, c, d. Figure 3a, c represent the SEM images for an un-doped LBW sample and Fig. 3b, d represents the Eu3.0 doped LBW phosphor with different magnifications and scales, which depict different particle morphologies with irregular shapes. The particles of phosphor with a size in microns (2–4 μm) have been aggregated as a result of hand grinding and high-temperature sintering. Smaller particles with a larger surface-to-volume ratio than bulk particles are considered to be the cause behind agglomeration [43]. Smaller particles can therefore have a lot of surface energy; as a result, they tend to group together to form larger particles in order to lower this energy and increase stability. The phosphor particles with size ranges in micro meters (μm) have been considered more convenient for use in pc w-LEDs [44].

The atomic and weight percentages of the elements that are present in the LBW:xEu^{3+} (x = 0, 3 mol%) phosphors have been depicted in EDAX as shown in Fig. 4a and b. From Fig. 4, it is conspicuous that, the elements such as Barium (Ba), tungsten (W), oxygen (O), and europium (Eu) are present in the as-prepared phosphor. However, the elements pertaining to Lithium (Li) is not observed in the EDAX spectrum as it is a light particle. From literature it is evident that, the light particles such as Li with very low energy of characteristic radiation, does not exhibit any peak in the EDAX spectra [45].

3.2 Thermal properties

The temperature-dependent weight loss of the as-prepared un-doped LBW material in thermogravimetric analysis (TGA) was studied to understand

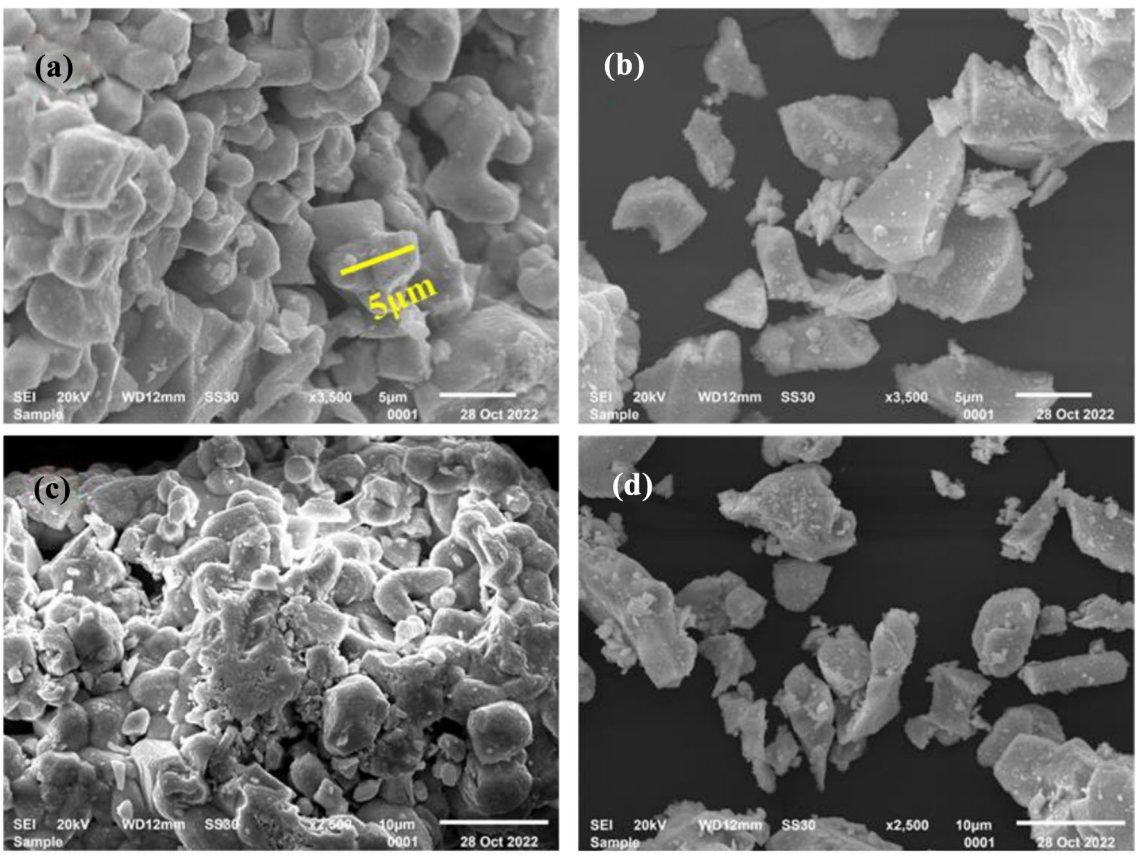


Fig. 3 SEM images for **a**, **c** un-doped LBW sample; and **b**, **d** LBWEu3.0 phosphor

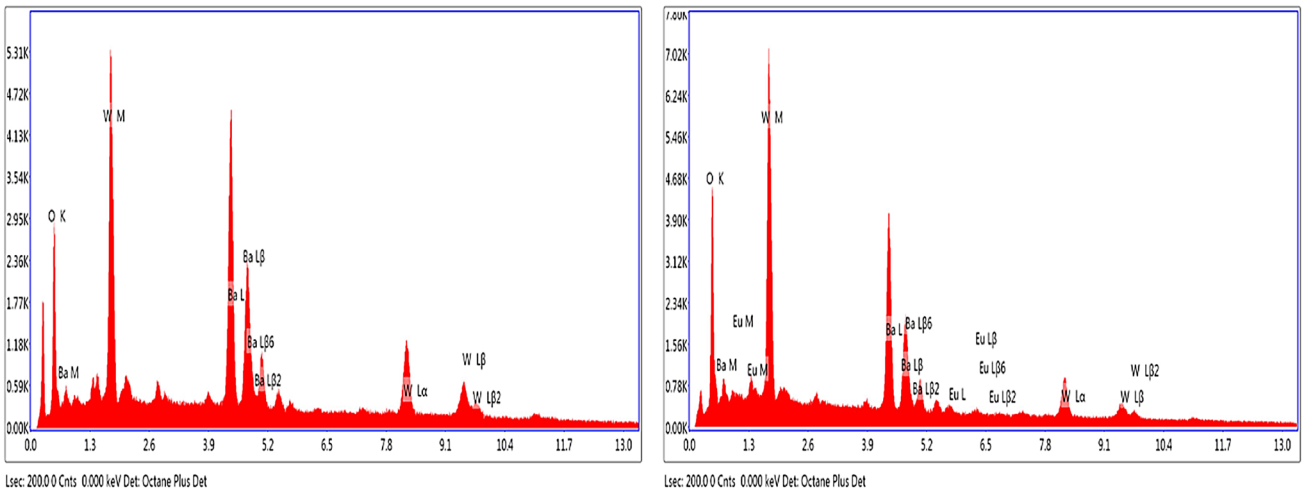


Fig. 4 EDAX spectra for un-doped LBW and LBWEu3.0 phosphors

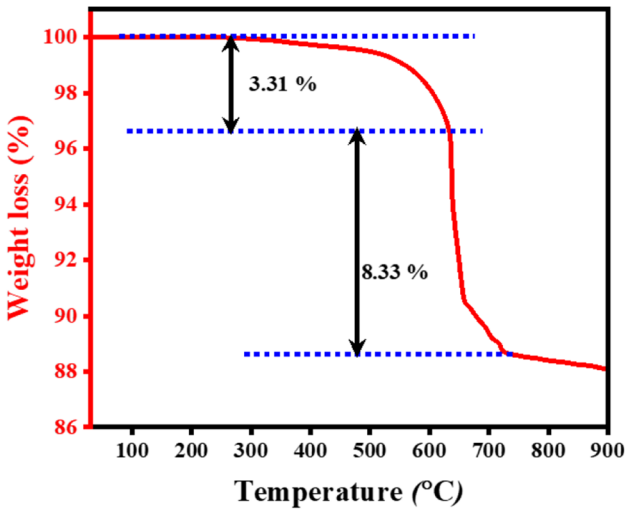


Fig. 5 Thermogravimetric analysis (TGA) of un-doped LBW sample

the thermal behavior. The TGA pattern, given in Fig. 5, shows a total weight loss of 12.25% over a temperature range of 30 to 900 °C, leaving 87.75% of the original weight. The TGA curve indicates that weight loss with temperature occurs in two stages. From 316 to 632 °C, an initial weight loss of almost 3.31% was detected due to decomposition of the raw materials [46]. At temperatures between 632 and 736 °C, there was a consistent drop and 8.31% weight loss of the as-prepared LBW phosphor material was detected. This temperature range may be where the LBW phase forms. The weight loss ceased and maintained a consistent trend when the temperature reached around 800 °C, indicating that the calcination stage was over 800 °C.

3.3 Optical analysis

3.3.1 *Diffuse reflectance study (DRS) and optical bandgap measurements*

The optical bandgap (E_g) evaluating form the DRS is a crucial parameter for phosphor materials used in w-LEDs. The DRS was measured for an un-doped LBW and LBWEu9.0 phosphor in between 200 and 1000 nm wavelength range and has been presented in Fig. 6. This spectrum consists of a broad absorption band with a maximum wavelength (λ_{max}) of 306 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, causes the absorption zone in the range of 280–330 nm. LBWEu9.0 sample also shows several narrow characteristic absorption lines at 464, 530 and 596 nm, which correspond to the $^7F_0 \rightarrow {}^5D_{2,1,0}$ transitions of Eu^{3+} ions, respectively [47].

The reflectance spectrum is transformed in to absorbance data using the Kubelka–Munk function given as [48]

$$F(R) = \frac{(1-R)^2}{2R} = \frac{K}{S}, \tag{5}$$

which relates the ratio of the absorption coefficient (K) and scattering coefficient (S) of the sample with diffuse reflectance (R).

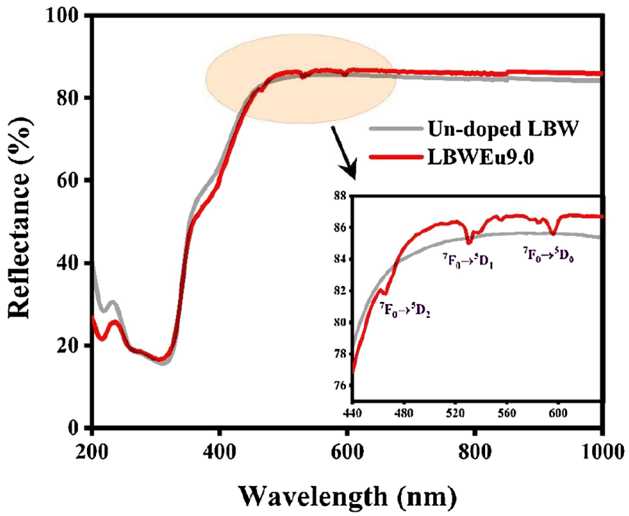


Fig. 6 The diffuse reflectance spectrum (DRS) for an un-doped LBW and LBWEu9.0 phosphor

In a parabolic band structure, the optical bandgap (E_g) and the semiconductor oxides absorption factor (B) can be interrelated as follows [49]:

$$\alpha(\nu) = \left(\frac{B}{h\nu}\right)(h\nu - E_g)^n, \tag{6}$$

where n is a constant related to various types of electronic transitions, $h\nu$ is the energy of a photon, C is a proportionality constant, E_g is the optical bandgap, and α is the material's linear absorption coefficient. Here, $n = 1/2, 2$ 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, $F(R)$ [50]. Therefore, in the calculations, α is changed to $F(R)$, and the optical bandgap (E_g) of LBW:xEu^{3+} ($1 \le x \le 9$, Δx = 2 mol%) is then determined using the following expression [48, 51, 52]:

$$F(R)h\nu = C\left(h\nu - E_g\right)^n, \tag{7}$$

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 7 represents a graph between $h\nu$ (incident energy) and $[F(R)h\nu]^2$, called Tauc Plot for Eu^{3+} ions-doped LBW phosphor samples. By extrapolation of the linear portion of the graph to the x-axis, the bandgap values of LBW:xEu^{3+} samples were evaluated. The results show that the E_g decreases with increasing Eu^{3+} ions concentration due to the formation of the intermediate energy levels of the Eu^{3+} ions in the LBW host matrix. The decrease is very less with increasing dopant ion concentration, revealing that the dopant ions do not have much effect on the structural behavior of the as-prepared LBW

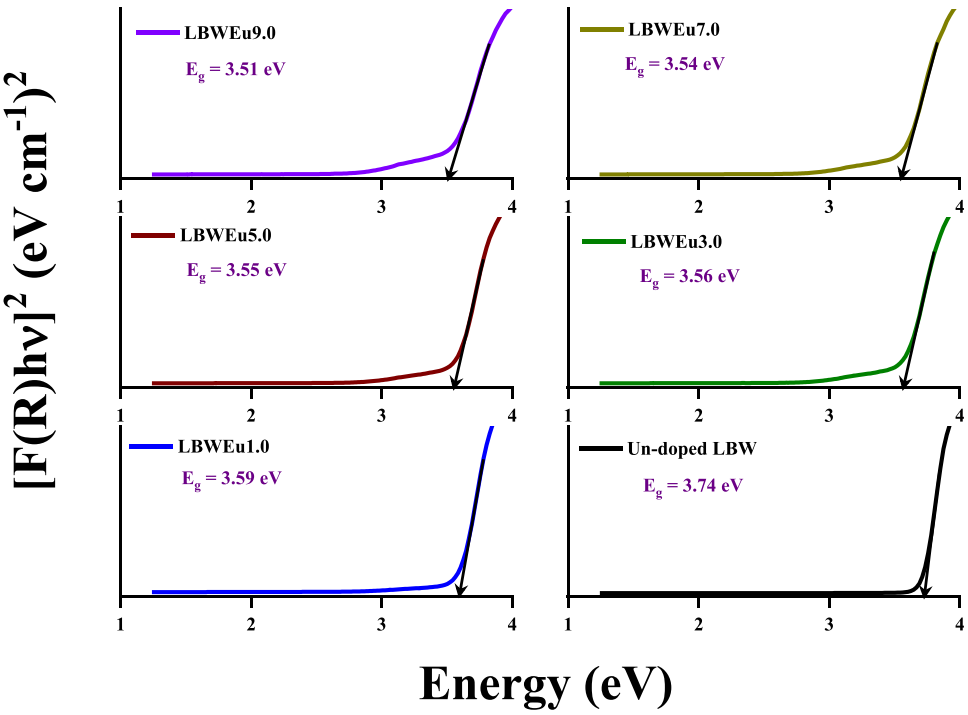


Fig. 7 Tauc's plot for un-doped LBW and LBW:xEu^{3+} ($1 \le x \le 9$) phosphors

Table 2 Bandgap E_g (eV) and Urbach's Energy ΔE (eV) of the as-prepared un-doped LBW and LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ mol%) phosphors

S. No.	System name	Bandgap (E_g (eV))	Urbach's energy (ΔE (eV))
1	Un-doped LBW	3.74	0.57
2	LBWEu1.0	3.59	0.48
3	LBWEu3.0	3.56	0.36
4	LBWEu5.0	3.55	0.34
5	LBWEu7.0	3.54	0.30
6	LBWEu9.0	3.51	0.24

phosphor materials. The variation of bandgap with Eu^{3+} ions concentration is presented in Table 2.

The degree of disorder present in a structure is represented by Urbach's law and the following formula was used to calculate the Urbach's energy (ΔE) [48]:

$$\alpha(\nu) = C \exp\left(\frac{h\nu}{\Delta E}\right), \tag{8}$$

where C, ΔE , and ν are constant, Urbach's energy, and the radiation frequency, respectively. The magnitude of Urbach's energy is calculated using the inverse of the slope of the linear section of the ln(α) vs. $h\nu$(eV) graphs. According to calculations, the Urbach's energies for LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ mol%) phosphors are 0.48, 0.36, 0.34, 0.30, and 0.24 eV, respectively, shown in Table 2. The Urbach's energy decreases with an increase in dopant ions concentration in the LBW phosphors indicating the decrement of defects and disorder in the host lattice with Eu^{3+} ions concentration. This means that there is relatively less chaos and lesser defects making the phosphors useful in enhancing its luminescence efficiency.

3.4 Photoluminescence properties

To know more about the comprehend emission spectral features of the Eu^{3+}-doped LBW phosphors, it needs to be excited with a suitable wavelength of excitation and for this an excitation spectrum needs to be recorded by fixing the prominent emission band of Eu^{3+} ions. The excitation spectrum of LBW:xEu^{3+} phosphors was recorded at an emission wavelength (λ_{em}) of 595 nm in the wavelength range from 250 to 500 nm. The excitation spectrum for LBWEu1.0 phosphor is presented in Fig. 8, containing the peaks of

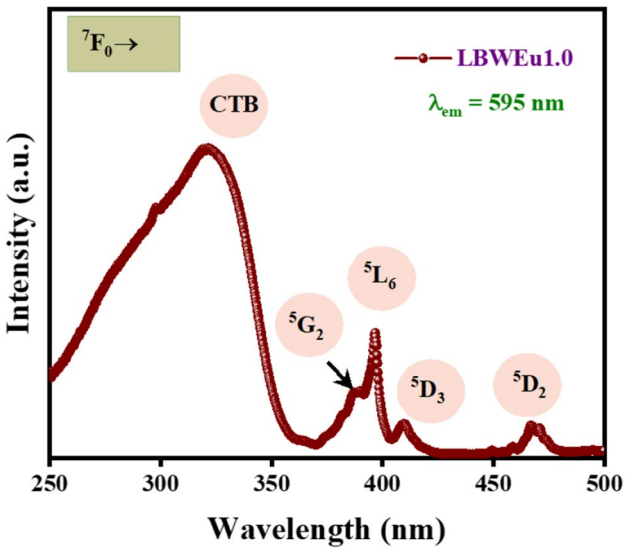


Fig. 8 The excitation spectrum for LBWEu1.0 phosphor

Eu^{3+} ions at 388, 396, 409, and 467 nm corresponding to the transition $^7F_0$ to $^5G_2$, $^5L_6$, $^5D_3$, and $^5D_2$, respectively [53]. A highly intense band is observed at 322 nm corresponds to a charge transfer band (CTB) and, therefore, is used as an excitation wavelength to study the PL emission spectrum.

Additionally, the band at 322 nm is stronger than the other peaks, suggesting that n-UV light is the best source for phosphor excitation. The results of the excitation-wavelength-dependent PL analysis, as displayed in Fig. 9a, confirm the above finding.

To determine the effect of Eu^{3+} ions on the PL intensity, the emission spectra for variable Eu^{3+} ions-doped LBW phosphors were recorded. Figure 9b illustrates the PL spectra of LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ mol%) phosphors at $\lambda_{ex} = 322$ nm in 550 to 700 nm wavelength range. The spectra contain peaks at 584, 595, 613, and 663 nm corresponding to the electronic transitions $^5D_0 \rightarrow {}^7F_0$, $^7F_1$, $^7F_2$, and $^7F_3$, respectively. The $^5D_0 \rightarrow {}^7F_2$ (Electric dipole (ED)) transition is highly sensitive to the surroundings of the dopant (here Eu^{3+} ions), and as a result, is hypersensitive to the distortion of the inversion symmetry, however, the $^5D_0 \rightarrow {}^7F_1$ (Magnetic dipole (MD)) transition is independent of the environment [27]. Previous research confirmed that Eu^{3+} ions inhabit low symmetry sites without an inversion center since the emission spectra clearly showed that the MD transition is dominant over the ED transition [31, 54, 55]. Ba^{2+} ions with low symmetry can be effectively replaced by Eu^{3+} ions because their ionic radii

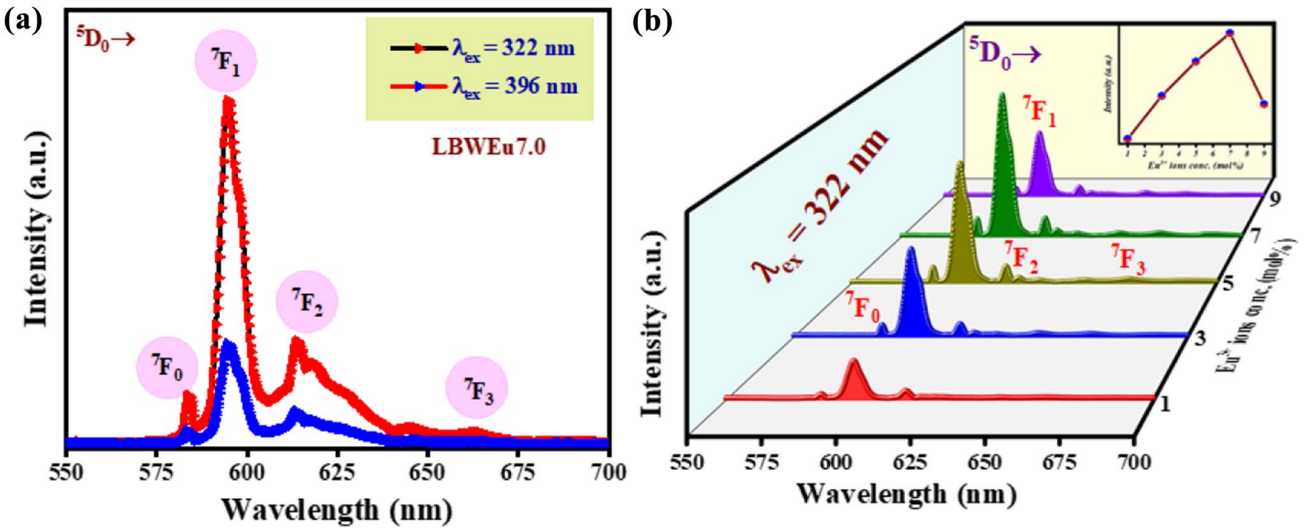


Fig. 9 **a** Excitation-wavelength-dependent photoluminescence analysis for LBWEu7.0 phosphor, **b** the emission spectra for LBW:xEu^{3+} ($1 \le x \le 9$) phosphors

(0.947Å) are lower than those of Ba^{2+} ions (1.35Å), surrounded by six atoms. The positions of the peak do not change as the concentration of Eu^{3+} ions increase, and the intensity of the emission peak increases as well. For Eu^{3+} ions-doped LBW phosphor, the maximum intensity is at 7 mol%, which is the optimum concentration. Due to the phenomenon of concentration quenching, the PL emission intensity decreases with increase in Eu^{3+} ions concentration [56–58]. The inset of Fig. 9b illustrates the intensity variations with doping concentration and confirms the concentration quenching at 7 mol% of Eu^{3+} ions in LBW host phosphor.

It has been observed that the concentration quenching is firmly related to the critical distance (R_c) between the Eu^{3+} ions. Blasse's theory is used to calculate R_c by applying the following formula [59]:

$$R_c = 2\left(\frac{3V}{4\pi N X_c}\right)^{\frac{1}{3}}, \tag{9}$$

where V and N are the volume and number of Eu^{3+} sites in the unit cell, respectively, and X_c is the critical concentration of Eu^{3+} ions. Earlier literature showed that the cause of concentration quenching depends on the value of R_c, i.e., if $R_c < 5$ Å, then the cause is the exchange; however, if $R_c > 5$ Å, then electric multipolar–multipolar interactions are the main cause for concentration quenching [60, 61]. Using Eq. (10), R_c is evaluated and found to be equal to 19.34Å, which is relatively greater than 5Å, confirming that the electric multipolar–multipolar interaction is the main reason for the concentration quenching in the as-synthesized LBW:xEu^{3+} phosphors.

According to the Dexter theory, dopant concentrations are related to luminescence emission intensity and therefore used to evaluate the multipolar–multipolar interactions, given as follows [62, 63]:

$$\frac{I}{x} = k\left[1 + \beta x^{\frac{Q}{3}}\right]^{-1}, \tag{10}$$

here, where x is the dopant ions concentration, k and β are the constants, I and Q are the luminescence intensity and strength of multipolar interaction, respectively. The value of Q suggests the type of interaction, i.e., Q = 6, 8, 10 for dipole–dipole, dipole–quadrupole, and quadrupole–quadrupole interaction, respectively.

Huang's theory gives a relation between the luminous intensity (I) and mole fraction (x) of the activator ions as follows [64]:

$$\log\left(\frac{I}{x}\right) = A - \frac{Q\log(x)}{3}, \tag{11}$$

where A is independent of the concentration of the dopant ions. Figure 10 represents the plot between $\log\left(\frac{I}{x}\right)$ and $\log(x)$ in LBW:xEu^{3+} phosphors. The experimental data were linearly fitted and the evaluated value of Q is around 6 representing that the

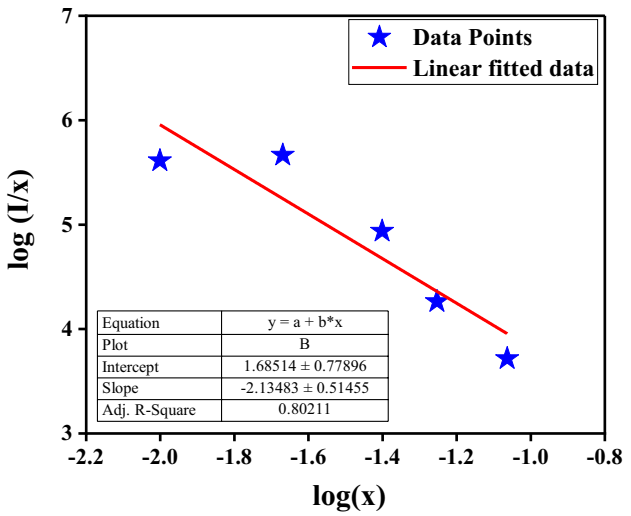


Fig. 10 Plot between log (*I*/*x*) versus log (*x*) in LBW:xEu^{3+} ($1 \le x \le 9$) phosphors

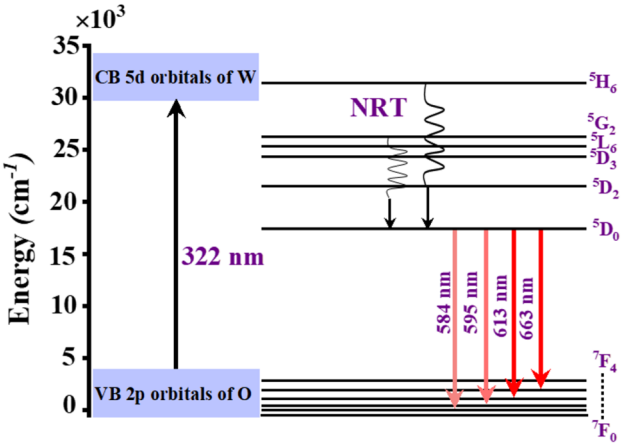


Fig. 11 Energy level diagram for Eu^{3+} ions in LBW host matrix

dipole–dipole interaction is responsible for the concentration quenching.

The emission, excitation, and non-radiative processes of Eu^{3+} ions in the LBW host matrix are represented in the energy level diagram, shown in Fig. 11. The oxygen 2p orbital electrons are excited by UV at 322 nm and move to the tungsten 5d orbital, corresponding to the valence band (VB) to conduction band (CB) transition. These electrons relax to lower energy states of tungsten through non-radiative transitions and then migrate their energies to excite Eu^{3+} ions to higher energy states. The multiphoton relaxation leads to the intensely excited electrons relaxing non-radiatively to a lower energy state and then radiatively from $^5D_0 \rightarrow {}^7F_J$ ($J = 0$ to 3).

3.5 CIE and CCT analysis

For the validation of the emission color of LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ mol%) phosphors, the CIE chromaticity coordinates were evaluated from the PL spectra with $\lambda_{ex} = 322$ nm using the formulae given in literature [65]. The evaluated values of CIE coordinates are (0.6265, 0.3729), (0.6263, 0.3732), (0.6272, 0.3723), (0.6274, 0.3721), and (0.6268, 0.3727) for LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ mol%) phosphors, respectively. Figure 12a represents the CIE coordinates for LBWEu7.0 phosphor sample and lies in the orange-red region.

The Correlated Color Temperature (CCT) is an additional factor, which helps in assessing the efficacy of a source of light and is measured using McCamy's polynomial equation described below [66]:

$$CCT = -449n^3 + 3525n^2 - 6823.3\,n + 5520.33, \quad (12)$$

where n is the anti-slope line and $n = \frac{(x-x_e)}{(y-y_e)}$, in which x_e (= 0.332) and y_e (= 0.186) are the chromaticity epicenters, x and y are the value of CIE coordinates. The "warmness" and "coolness" of the light are commonly examined by the CCT values. Light is classified as warm if the CCT is less than 3200 K and cold if the CCT is greater than 4200 K [67]. The computed CCT values for different Eu^{3+} ions-doped LBW phosphors are 1942, 1938, 1953, 1957, and 1946 K, respectively, and therefore, fall within the warm region. The formula mentioned in the literature is used to determine the color purity (CP) [68]. The evaluated values of CP are higher than 79% for all the as-prepared Eu^{3+} ions-doped LBW phosphors under the excitation wavelength of 322 nm, shown in Fig. 12b. The CIE, CCT, and CP values are listed in Table 3.

3.6 PL decay study

PL decay curves for LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ mol%) phosphors under $\lambda_{em} = 595$ nm and $\lambda_{ex} = 322$ nm are shown in Fig. 13a.

The double exponential equation is the best fit to the observed decay curves [67]:

$$I_t = I_0 + A \exp\left(-\frac{t}{\tau}\right), \quad (13)$$

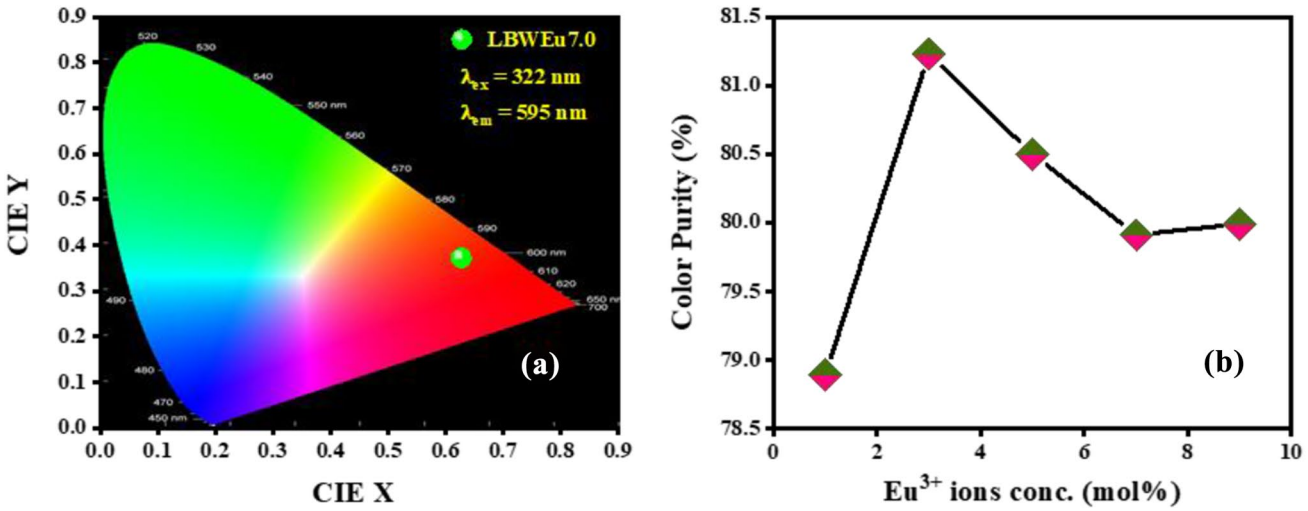


Fig. 12 **a** CIE chromaticity coordinates for LBWEu7.0 phosphor, **b** color purity for LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ mol%) phosphors

Table 3 CIE, CCT, color purity, and experimental lifetime (τ_{exp}) of the as-prepared LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ mol%) phosphors

System name	CIE Coordinates		CCT (K)	Color purity (%)	τ_{exp} (ms)	η (%)
	X	*Y*				
Eu1.0	0.626	0.373	1942	78.89	3.50	72.9
Eu3.0	0.626	0.373	1938	81.23	3.40	70.2
Eu5.0	0.627	0.372	1953	80.51	3.33	68.2
Eu7.0	0.627	0.372	1957	79.91	3.37	74.9
Eu9.0	0.626	0.373	1946	79.99	3.24	71.1

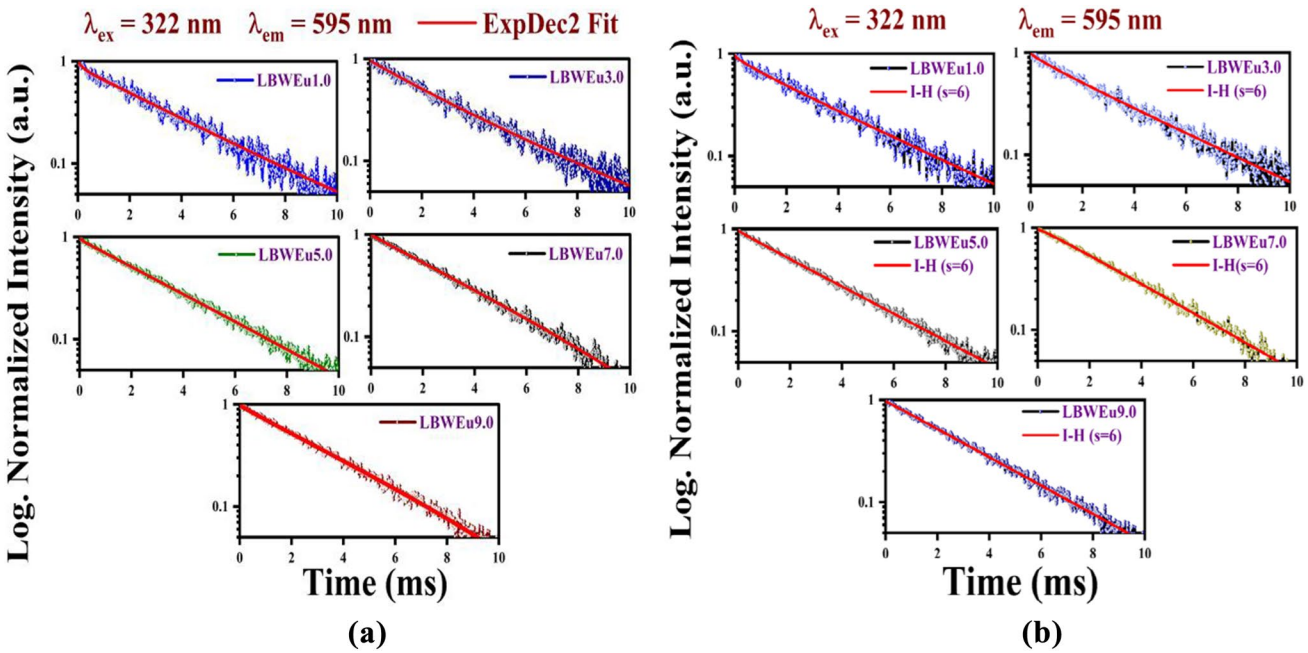


Fig. 13 **a** PL decay curves for LBW:xEu^{3+} ($1 \le x \le 9$) phosphors, **b** *I–H* model fitted curves for LBW:xEu^{3+} ($1 \le x \le 9$, $\Delta x = 2$ mol%) phosphors

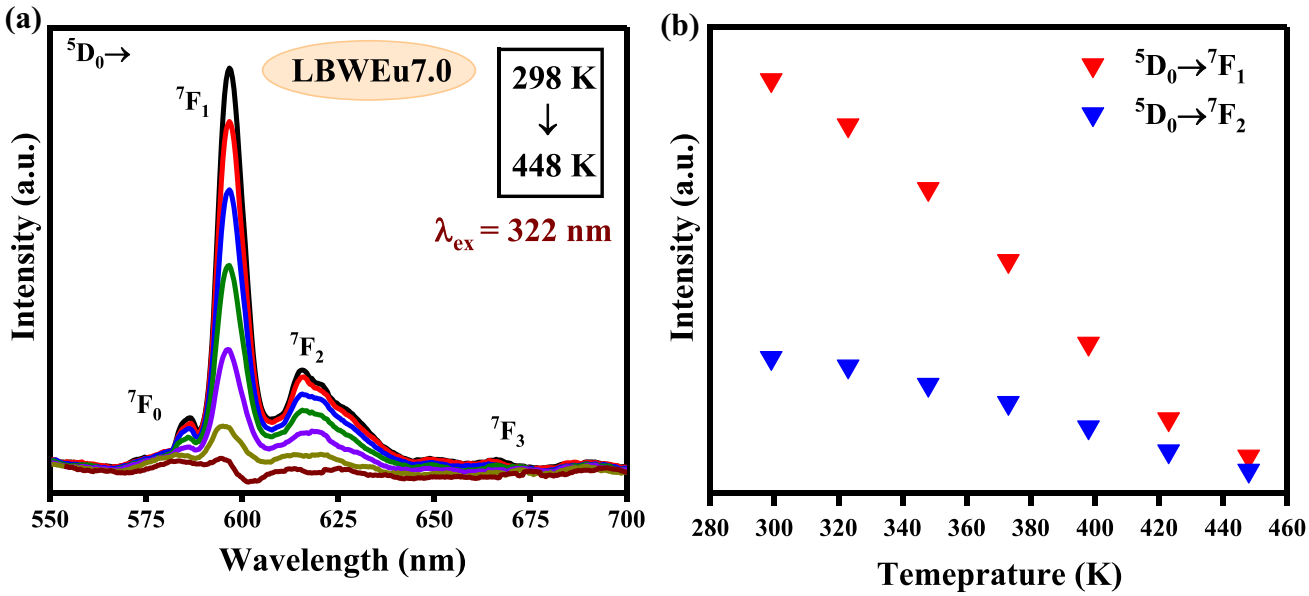


Fig. 14 **a** TD-PL spectra of the LBWEu7.0 phosphor at λ_{ex} = 322 nm, **b** variation of emission intensities of two TCLs of Eu^{3+} ions

where I_t and I_0 are the PL intensity at time t (in seconds) and at t = 0, respectively. A represents the fitting constant, τ denotes the decay lifetime. The decay lifetime values decrease with an increase in Eu^{3+} ions concentration in the range of (3.27–3.19) ms, presented in Table 3, due to the non-radiative transfer (NRT) of energy.

The luminous quantum efficiency η is defined as the ratio of the number of the photons emitted to the number of photons absorbed. Here we define this in terms of lifetime i.e., η is the ratio of average lifetime and radiative lifetime of the $^7F_1$ energy level and can be represented using the following equation [69]:

$$\eta = \frac{\tau}{\tau_{rad}} \times 100\%. \tag{14}$$

The quantum efficiencies of the as-prepared Eu^{3+} ions-doped LBW phosphors lie in the range 68–75% and presented in Table 3.

To identify the associated energy migration process, the I–H model has been used to the decay curves of all the as-prepared Eu^{3+} ions-doped LBW phosphors. The decay curves of LBW:xEu^{3+} ($1 \le x \le 9$, Δx = 2 mol%) have been best fitted for s = 6 using the equation given below [67] and shown in Fig. 13b:

$$I_t = I_0 \times \exp\left[-\frac{t}{t_0} - Q \times \left(\frac{t}{t_0}\right)^{\frac{3}{s}}\right], \tag{15}$$

where t_0 represents the donor lifetime and Q is the energy migration parameter. The value of "s" reflects the nature of the interaction between the activator ions in an individual host. The best I–H fit for the current samples was observed for s = 6, indicating that the Eu^{3+} ions' dominant interaction in LBW phosphor is d–d (dipole–dipole). Dexter's theory of energy migration yielded the same conclusion.

3.7 TD-PL studies

TD-PL spectra of the LBWEu7.0 phosphor at λ_{ex} = 322 nm are shown in Fig. 14a in the temperature range of 298–448 K. There was a drop in the emission strength of the spectra, but there was not a noticeable shift in the position of the emission spectra due to the temperature increase. The thermal quenching effect is responsible for the decrease in light intensity as temperature increases [70]. The intensity of PL emission varies differently for the peaks presents at 595 and 615 nm with temperature, represented in Fig. 14b for LBWEu7.0 phosphor sample.

Both the peak, at 595 and 615 nm, decreases with increase in temperature but the decrement in the intensity of 615 nm peak is steeper than at 595 nm and can be assumed as thermally coupled energy levels (TCLs). Temperature has an impact on the relative population of various emission levels, which follows a Boltzmann-type distribution. Using the temperature-dependent emission spectra of the

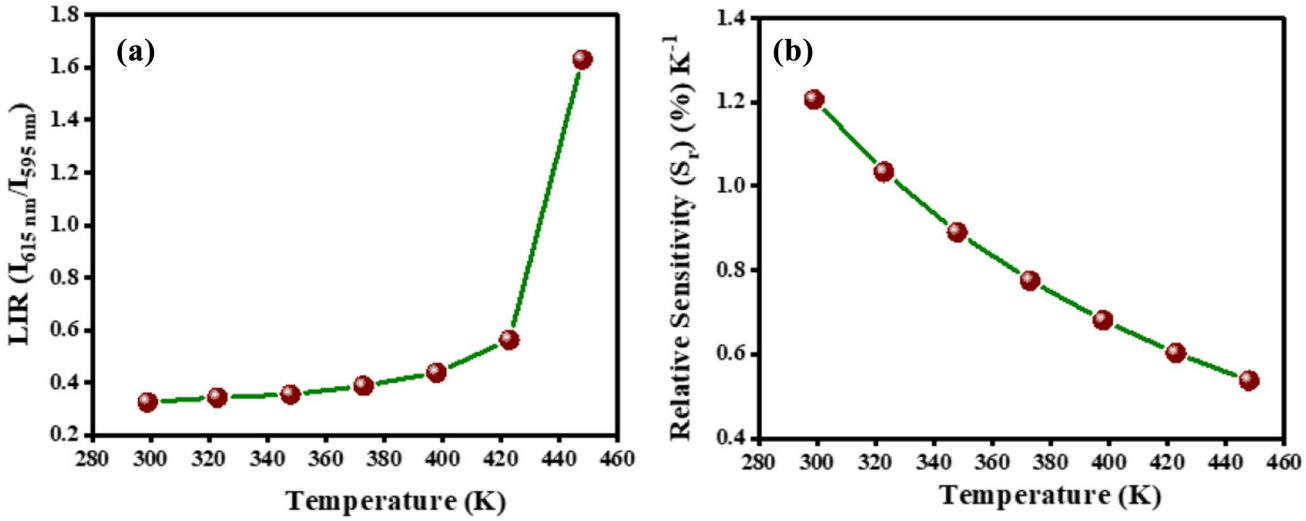


Fig. 15 **a** Dependency of LIR of the $^5D_0 \rightarrow {}^7F_1$ and $^5D_0 \rightarrow {}^7F_2$ transitions with the temperature, **b** the relative sensitivity (S_r) of LBWEu7.0 phosphor as a function of temperature

LBWEu7.0 phosphor sample (Fig. 14) under the 322 nm excitation, the luminescence intensity ratio (LIR) theory was applied to evaluate its performance as a luminescent thermometer. Due to its simplicity, equipment requirements, and, most importantly, the fact that it is ratiometric and self-referenced, the LIR has undergone the most research of any luminescent temperature read-out method. Excitation power variations have no impact on the temperature output. RE^{3+} ions have relatively strong emission lines and a range of spectral regions to choose from, ranging from ultraviolet (UV) to near-infrared (NIR), making them the most desirable materials for LIR thermometry. The temperature-dependent ratio of two emissions is referred to as LIR.

The ratio of the population between the two thermally coupled emission levels (TCLs), LIR, is determined by a population distribution of the Boltzmann type and is given by the following equation [71]:

$$\mathrm{LIR} = \frac{I_{595\mathrm{nm}}}{I_{615\mathrm{nm}}} = A \exp\left(\frac{-\Delta E}{KT}\right), \tag{16}$$

here K represents the Boltzmann constant, A denotes a proportional parameter, and ΔE is the energy separation of two TCLs ($^7F_1$ and $^7F_2$).

The measured plot of LIR vs. T (temperature) is nicely matched by Eq. (17), as seen in Fig. 15a. For temperature control purposes, it is crucial to know the rate of change in LIR with temperature. The sensor sensitivity, which illustrates the material's

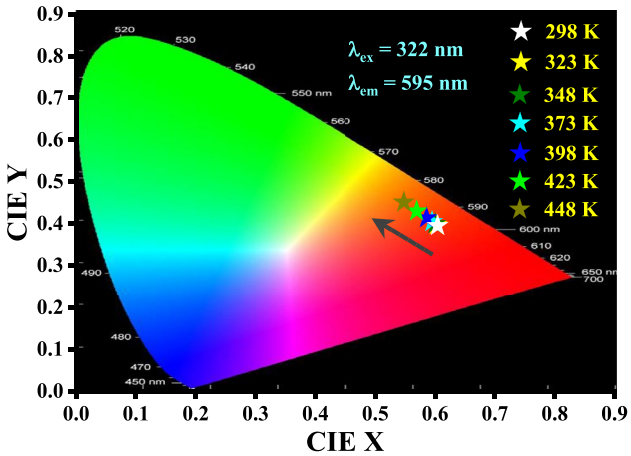


Fig. 16 The estimated CIE coordinate values for LBWEu7.0 at various temperatures

sensitivity, is the rate at which the LIR changes for a specific change in temperature. Relative Sensitivity (S_r) value, which can be computed using the calculation below, is used in practical applications to recognize a material's optical thermometric capability [72]:

$$S_r = \frac{1}{\mathrm{FIR}} \frac{\mathrm{d(LIR)}}{\mathrm{d}T} \times 100\% = \left(\frac{\Delta E}{KT^2}\right) \times 100\%. \tag{17}$$

The calculated relative sensitivity for TCLs is dependent on the absolute temperature, as shown in Fig. 15b. It is evident that the evaluated value of the S_r for prepared phosphors is 1.21% K^{-1} at 298 K and subsequently exhibits monotonic diminishing

Table 4 The variation of LIR, S_r, CIE, CP, and CCT of LBWEu7.0 with change in temperature (298–448 K)

Temperature (K)	LIR	S_r	CIE		CP (%)	CCT (K)
			X	*Y*		
298	0.33	1.21	0.621	0.379	79.68	1856
323	0.34	1.03	0.618	0.381	79.28	1833
348	0.36	0.89	0.616	0.383	80.14	1805
373	0.38	0.77	0.611	0.388	79.92	1763
398	0.44	0.68	0.601	0.398	80.35	1723
423	0.56	0.60	0.585	0.414	81.50	1750
448	1.63	0.53	0.563	0.435	81.90	2092

behavior as the temperature rises. This phosphor was found to have the highest observed sensor sensitivity for the low-temperature range as a result. As a result, LBWEu7.0 phosphor is a top choice for optical thermometry.

The estimated CIE coordinate values for LBWEu7.0 at various temperatures are shown in Fig. 16, confirm the tunability of the luminous coloring pattern from orange-red to yellowish-orange region. The color purity is also evaluated at various temperatures for the optimized phosphor sample. The values for all parameters LIR, S_r, CIE, CP, and CCT are presented in Table 4.

4 Conclusions

In conclusion, a series of Eu^{3+} ions-doped LBW phosphors has been synthesized using the traditional solid-state method and their structural, morphological, and PL properties have been investigated using various characterization techniques. The XRD profile verify the phase purity of the synthesized phosphors. The calculated values of average crystallite size using the Debye–Scherrer method and W–H plot were found to be nearly the same and equal to 83 and 84 nm, respectively. The SEM images display particles with irregular and non-uniform shapes, along with some mild agglomeration. The particle size is within the micro-range, which is suitable for the production of white LEDs. The diffuse reflectance spectrum reveals the presence of 4f transitions of Eu^{3+} ions. The band-gap values given by the linear fitting of the Tauc's plot and found to be in the range of 3.74–3.51 eV, which illustrates the uses of synthesized phosphors in the field of optical applications. The PL investigation suggests that the synthesized phosphors contain an intense band corresponding to $^5D_0 \rightarrow {}^7F_1$ transition at 595 nm when pumped with 322 nm wavelength. The concentration quenching takes place after 7 mol% concentration of Eu^{3+} ions in the host lattice. The Dexter theory as well as I–H model explain the dipole–dipole interaction as the energy transfer mechanism between Eu^{3+} ions. The CIE coordinates of the optimized sample at $\lambda_{ex} = 322$ nm are (0.6274, 0.3721), which lie in the orange-red region. The decay profile exhibits double exponential behavior with decay times lying in the millisecond range. Due to energy migration between Eu^{3+}–Eu^{3+} ions, the decay lifetime values decrease with an increase in Eu^{3+} ions concentration. LIR method was applied in order to examine the possibility of using the synthesized phsophor material in thermometry applications. PL measurements recorded from room temperature up to 448 K have shown that LBWEu7.0 phosphor possesses the maximum relative sensitivity of approximately 1.21% K^{-1} at a low temperature of 298 K, indicating its good temperature sensor ability within the studied temperature range and can be used in thermometry applications. Also, based on the optical investigations of Eu^{3+} ions-doped $Li_2Ba_5(WO_5)_3$ phosphor, it can be recognized as a suitable material for fabricating photonic devices like w-LEDs and non-contact thermometry.

Acknowledgements

One of the authors, 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) to her.

Author contributions

A: conceptualization, methodology, experimental results, writing the original draft—ASR: conceptualization, validation, correction of draft manuscript, and editing.

Funding

No funding was received to assist with the preparation of this manuscript. The authors declare they have no financial interest.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. The datasets are presented in the main manuscript.

Declarations

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

Research involving in human and animal participants No human participants and animals are involved in this research.

References

1. L.R. Đačanin, M.D. Dramićanin, S.R. Lukić-Petrović, D.M. Petrović, M.G. Nikolić, Radiat. Meas. **56**, 143 (2013)
2. A.H. Khalid, K. Kontis, 5673 (2008)
3. A. Ćirić, S. Stojadinović, M.D. Dramićanin, Ceram. Int. **46**, 23223 (2020)
4. Y. Zhu, Q. Meng, W. Sun, S. Lü, J. Alloys Compd. **784**, 456 (2019)
5. A. Siaï, P. Haro-González, K. Horchani, Naifer, M. Férid, Opt. Mater. (Amst). **76**, 34 (2018)
6. X. Huang, C. He, X. Zhu, C. Yang, Y. Liu, M. Wu, X. Min, R. Mi, Z. Huang, Ceram. Int. **47**, 32130 (2021)
7. S.K. Sharma, T. Köhler, J. Beyer, M. Fuchs, R. Gloaguen, J. Heitmann, Phys. Chem. Chem. Phys. **21**, 16329 (2019)
8. K. Li, R. Van Deun, J. Alloys Compd. **787**, 86 (2019)
9. A.M. Kaczmarek, J. Mater. Chem. C **6**, 5916 (2018)
10. K. Li, X. Liu, Y. Zhang, X. Li, H. Lian, J. Lin, Inrg. Chem. **54**, 323 (2015)
11. C. Zhang, L. Yang, J. Zhao, B. Liu, M.Y. Han, Z. Zhang, Angew. Chemie—Int. Ed **54**, 11531 (2015)
12. K. Siva Rama, K. Reddy, K. Swapna, S. Mahamuda, M. Venkateswarlu, A.S. Rao, G. Vijaya, Prakash, Opt. Mater. (Amst). **85**, 200 (2018)
13. S.K. Gupta, M. Sahu, P.S. Ghosh, D. Tyagi, M.K. Saxena, R.M. Kadam, Dalt Trans. **44**, 18957 (2015)
14. G.B. Nair, H.C. Swart, S.J. Dhoble, Prog Mater. Sci. **109**, 100622 (2020)
15. S. Ye, F. Xiao, Y.X. Pan, Y.Y. Ma, Q.Y. Zhang, Mater. Sci. Eng. R Reports. **71**, 1 (2010)
16. P.J. Yadav, C.P. Joshi, S.V. Moharil, J. Lumin. **136**, 1 (2013)
17. C. Sommer, P. Hartmann, P. Pachler, H. Hoschopf, F.P. Wenzl, J. Alloys Compd. **520**, 146 (2012)
18. Y.U. Quanmao, J. Rare Earths. **26**, 783 (2008)
19. P. Du, J.S. Yu, J. Alloys Compd. **653**, 468 (2015)
20. S. Som, P. Mitra, V. Kumar, V. Kumar, J.J. Terblans, H.C. Swart, S.K. Sharma, Dalt Trans. **43**, 9860 (2014)
21. H.P. Barbosa, I.G.N. Silva, M.C.F.C. Felinto, E.E.S. Teotonio, O.L. Malta, H.F. Brito, J. Alloys Compd. **696**, 820 (2017)
22. P. Chen, W. Hu, D. Yang, J. Zhu, J. Zhang, Y. Wu, Phys. B Condens. Matter. **530**, 127 (2018)
23. H. Nagabhushana, D.V. Sunitha, S.C. Sharma, B. Daruka Prasad, B.M. Nagabhushana, R.P.S. Chakradhar, J. Alloys Compd. **595**, 192 (2014)
24. Y. Li, X. Liu, Opt. Mater. (Amst). **42**, 303 (2015)
25. S. Sailaja, S.J. Dhoble, B. Sudhakar Reddy, J. Mol. Struct. **1003**, 115 (2011)
26. F.B. Dejene, J. Lumin. **199**, 433 (2018)
27. J. Liang, L. Sun, G. Annadurai, B. Devakumar, S. Wang, Q. Sun, J. Qiao, H. Guo, B. Li, X. Huang, RSC Adv. **8**, 32111 (2018)
28. Y. Zhang, Y. Li, Y. Yin, J. Alloys Compd. **400**, 222 (2005)
29. X. Huang, H. Guo, B. Li, J. Alloys Compd. **720**, 29 (2017)
30. B. Li, X. Huang, H. Guo, Y. Zeng, Dye. Pigment. **150**, 67 (2018)
31. Y.R. Parauha, D.K. Halwar, S.J. Dhoble, Displays. **75**, 102304 (2022)
32. S. Wang, C. Xu, X. Qiao, Ceram. Int. **47**, 1063 (2021)
33. W. Shi, A. Feng, H. Tang, Z. Ding, Y. Ma, M. Wu, G. Li, Opt. Mater. (Amst). **35**, 609 (2013)
34. J. Liao, B. Qiu, H. Lai, J. Lumin. **129**, 668 (2009)
35. S.A. Yan, Y.S. Chang, J.W. Wang, W.S. Hwang, Y.H. Chang, Mater. Res. Bull. **46**, 1231 (2011)

36. Y. Liu, Y. Wang, L. Wang, Y.-Y. Gu, S.-H. Yu, Z.-G. Lu, R. Sun, RSC Adv. **4**, 4754 (2014)
37. H. Zhu, C. Liang, W. Huang, Phys. B Condens. Matter. **582**, 411999 (2020)
38. B.S. Barros, A.C. de Lima, Z.R. da Silva, D.M.A. Melo, S. Alves-Jr, J. Phys. Chem. Solids. **73**, 635 (2012)
39. J.S. Thorp, E.A.E. Ammar, J. Mater. Sci. **11**, 1215 (1976)
40. S. Jana, A. Mondal, J. Manam, S. Das, J. Alloys Compd. **821**, 153342 (2020)
41. L. Alexander, H.P. Klug, J. Appl. Phys. **21**, 137 (1950)
42. P. Kumar, S. Singh, I. Gupta, K. Nehra, V. Kumar, D. Singh, Mater. Chem. Phys. (2022). https://doi.org/10.1016/j.matchemphys.2022.127035
43. A.K. Bedyal, A.K. Kunti, V. Kumar, H.C. Swart, J. Alloys Compd. **806**, 1127 (2019)
44. A.S. Rao, J. Fluoresc. (2023). https://doi.org/10.1007/s10895-023-03449-z
45. C. Kumari, A. Kumar, S.K. Sharma, J. Manam, Vacuum. **207**, 111599 (2023)
46. I. Tseng, M. Tsai, T. Hsu, C. Lin, S. Lin, C. Huang, Opt. Laser Technol. **157**, 108777 (2023)
47. S. Shisina, S. Das, S. Som, S. Ahmad, V. Vinduja, P. Merin, K.G. Nishanth, P. Kumari, M.K. Pandey, J. Alloys Compd. **802**, 723 (2019)
48. Z. Xia, Y. Zhang, M.S. Molokeev, V.V. Atuchin, J. Phys. Chem. C **117**, 20847 (2013)
49. T.T. Taru Chanu, N.R. Singh, J. Solid State Chem. **284**, 121190 (2020)
50. A. Ghosh, N. Kumari, S. Tewari, A. Bhattacharjee, Mater. Sci. Eng. B **196**, 7 (2015)
51. R. López, R. Gómez, J. Sol-Gel Sci. Technol. **61**, 1 (2012)
52. R. Köferstein, L. Jäger, S.G. Ebbinghaus, Solid State Ionics. **249–250**, 1 (2013)
53. B. Vasanthi, N. Gopakumar, P.S. Anjana, J. Lumin. **241**, 118486 (2022)
54. K.N. Shinde, S.J. Dhoble, A. Kumar, J. Rare Earths. **29**, 527 (2011)
55. V. Dubey, J. Kaur, S. Agrawal, N.S. Suryanarayana, K.V.R. Murthy, Superlattices Microstruct. **67**, 156 (2014)
56. S.R. Bargat, Y.R. Parauha, G.C. Mishra, S.J. Dhoble, J. Mol. Struct. **1221**, 128838 (2020)
57. Y.R. Parauha, V. Chopra, S.J. Dhoble, Mater. Res. Bull. **131**, 110971 (2020)
58. Y.R. Parauha, S.J. Dhoble, Luminescence. **36**, 1837 (2021)
59. Y. Hua, S.K. Hussain, J.S. Yu, Ceram. Int. **45**, 18604 (2019)
60. X. Geng, Y. Xie, Y. Ma, Y. Liu, J. Luo, J. Wang, R. Yu, B. Deng, W. Zhou, J. Alloys Compd. **847**, 156249 (2020)
61. G.V. Kanmani, V. Ponnusamy, G. Rajkumar, M.T. Jose, Opt. Mater. (Amst). **96**, 109350 (2019)
62. P. Rohilla, A.S. Rao, Mater. Res. Bull. **150**, 111753 (2022)
63. M. Jiao, Q. Xu, Y. Zhao, D. Wang, L. Liu, C. Yang, J. Lumin. **226**, 117388 (2020)
64. Y. Il Jeon, L. Krishna Bharat, J.S. Yu, J. Lumin. **166**, 93 (2015)
65. Anu, A. Kumar, Ravina, N. Deopa, A.S. Rao, J. Non Cryst. Solids. **616**, 122421 (2023)
66. C.S. McCamy, Color. Res. Appl. **17**, 142 (1992)
67. A.N. Deopa, A.S. Rao, Opt. Laser Technol. **154**, 108328 (2022)
68. G. Zhu, Z. Li, C. Wang, X. Wang, F. Zhou, M. Gao, S. Xin, Y. Wang, Dalt Trans. **48**, 1624 (2019)
69. R. Lohan, A. Kumar, M.K. Sahu, A. Mor, V. Kumar, N. Deopa, A.S. Rao, J. Mater. Sci. Mater. Electron. **34**, 694 (2023)
70. S.N. Menon, S.Y. Kadam, D.K. Koul, J. Lumin. **216**, 116716 (2019)
71. Z. Liang, F. Qin, Y. Zheng, Z. Zhang, W. Cao, Sens. Actuators Phys. **238**, 215 (2016)
72. Anu, A.S. Rao, Opt. Mater. (Amst). **145**, 114476 (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.

scientific reports

Check for updates

OPEN

Novel role of folate (vitamin B9) released by fermenting bacteria under Human Intestine like environment

Sharda Nara[1], Gulshan Parasher[3], Bansi Dhar Malhotra[1,2]✉ & Manmeet Rawat[3,4]✉

The anaerobic region of the gastrointestinal (GI) tract has been replicated in the anaerobic chamber of a microbial fuel cell (MFC). Electroactive biomolecules released by the facultative anaerobes (*Providencia rettgeri*) under anoxic conditions have been studied for their potential role for redox balance. MALDI study reveals the presence of vitamin B9 (folate), 6-methylpterin, para-aminobenzoic acid (PABA) and pteroic acid called pterin pool. ATR-FTIR studies further confirm the presence of the aromatic ring and side chains of folate, 6-methylpterin and PABA groups. The photoluminescence spectra of the pool exhibit the maximum emission at 420, 425, 440, and 445 nm when excited by 310, 325, 350, and 365 nm wavelengths (day 20 sample) highlighting the presence of tunable bands. The cyclic voltammetric studies indicate the active participation of pterin pool molecules in the transfer of electrons with redox potentials at − 0.2 V and − 0.4 V for p-aminobenzoate and pterin groups, respectively. In addition, it is observed that under prolonged conditions of continuous oxidative stress (> 20 days), quinonoid tetrahydrofolate is formed, leading to temporary storage of charge. The results of the present study may potentially be useful in designing effective therapeutic strategies for the management of various GI diseases by promoting or blocking folate receptors.

Under normal physiological conditions, a large gradient of oxygen is known to exist in different regions of the intestine and within the epithelial cells of the large intestine[1]. The bacterial population of the gastrointestinal (GI) tract has been reported to exhibit different sensitivities to different oxygen levels, and some of the bacteria are unable to survive in the presence of oxygen[2]. In the colon portion of the intestine, oxygen levels have been recorded to be below 1 mm Hg[3]. As a result, in the human gut, the anaerobic bacteria are 100 to 1000 times more populous than aerobic bacteria[4,5]. This means that their genetic make-up allows them to thrive in these conditions, allowing for a sustained release of redox mediators, for example, vitamins which may be necessary to support the gut epithelial cell in oxygen-deprived conditions[6,7].

In addition, the correlation of colorectal cancer has been found with intracellular redox imbalance in the body[8] with generation of higher levels of oxidants such as of reactive oxygen species (ROS) and reactive nitrogen species (RNS) than antioxidants that can oxidize macromolecules of proteins, DNA, lipid and increased cytotoxicity[9]. In the case of colon cancer, the response of the cancer cells has been shown to be regulated by oxygen partial pressure[2].

According to an estimate, around 1.2 million patients are diagnosed with colorectal cancer every year, and around 0.6 million die from it. The cases are reported to be higher in men as compared to women[8,10,11]. Globally, in 2018, more than 1.8 million people were diagnosed with colorectal cancer; in 2019, about 150,000 new cases were recorded, while 52,000 deaths occurred due to colorectal cancer in the United States alone[12,13]. Compared to developing countries, colorectal cancer is the leading cause of death in industrialized western countries[14].

Current treatment strategies are based on a combination of different modalities, including chemotherapy, surgical resection, and immunomodulatory therapy. However it has been reported that the health of 40% of

[1]Nanobioelectronics Laboratory, Department of Biotechnology, Delhi Technological University, Delhi 110042, India. [2]Environmental Sciences & Biomedical Metrology, CSIR-National Physical Laboratory, Dr K.S. Krishnan Road, New Delhi 110012, India. [3]Division of Gastroenterology and Hepatology, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA. [4]Division of Gastroenterology and Hepatology, Department of Medicine, The Penn State University College of Medicine, Penn State University, 500 University Drive, Hershey, PA 17033, USA. ✉email: bansi.malhotra@gmail.com; manmeetrawart@gmail.com; mrawat@pennstatehealth.psu.edu

cancer patients eventually declines with disease recurrence, making treatment effective for less than 15% of total patients, even as they also suffered from drug resistance and other side effects. Effective target selectivity coupled with knowledge of precise scientific mechanisms may help us to treat colorectal cancer more effectively with fewer side effects[15,16].

Several designs have previously attempted to create an oxygen gradient for the co-culture of microbes and intestinal cells to study their behavior. For example, Raehyun Kim et al. have designed a device consisting of materials with differential oxygen permeability to provide a gradient of oxygen to intestinal epithelial cells and create an oxygen-deficient lumen for co-culture of the anaerobes[17]. Pranjul Shah et al. have designed a microfluidic based model for co-culture of intestinal anaerobes and human intestinal cells[18].

Furthermore, tumor cells show greater adaptability to persistent oxidative stress, therefore more information is needed to manipulate oxidative stress and related transcription factors to target cancer cells specifically. Otherwise, the adaptability of tumor cells makes them more drug resistant, thereby hindering cancer treatment[13].

Thus, there is a need to study, in more detail, the gut bacterial biomolecules involved in maintaining the redox balance and the mechanism of charge transfer to prevent the generation of free radicals. This may not only prevent pathological conditions but also help to develop therapeutic options for GI diseases.

To study the nature of these biomolecules in vitro, the anodic chamber of a microbial fuel cell (MFC) can be used because it is relatively more user-friendly and allows for the release of biomolecules participating in the redox reactions. Normally, an MFC operates in the range of 0 to 20% oxygen partial pressure[19], similar to the anoxic lumen of the intestine. In the anodic chamber of MFCs, bacteria that survive under anoxic conditions have previously reported releasing vitamins that typically act as electron acceptors and/or redox mediators in the absence of oxygen[20]. Hanne Wiessenkens et al. have studied the intestinal cell culture in an anaerobic chamber ($< 0.1\%$ O_2). This study showed that anoxic conditions for intestinal cells reduce oxidative stress with higher oxidized glutathione levels[2]. Therefore, by providing gut-like oxygen-deprived conditions to facultative anaerobic bacteria, the released biomolecules participating in redox reactions can be isolated and studied for their potential role in preventing oxidative stress to the biological cell.

In the present study, we have taken *Providencia rettgeri*, a facultative anaerobe that is classified in the order *Enterobacteriales*[21], a large population of which resides in the intestinal part of the human body [22]. The fluorescence, cyclic voltammetry and circular dichroism analysis of folate molecules and folate synthesis intermediate molecules including para-aminobenzoic acid (PABA). 6-methyl pterin and pteroic acid called pterin pool released in the fuel cell anaerobic chamber has been studied to implicate the mechanism of charge transfer through these molecules to prevent oxidative stress. As it has been previously reported that with the progression of the pathological state of the tissue especially in case of neoplastic changes, the relevant physiochemical and biochemical changes occur at the tissue and cellular level could be perceived through changes in the fluorescence emission of endogenous fluorophores such as collagen, NADH and FAD. Also, the fluorescence measurement along with the measurement of redox ratio in the different tissue types, informs about the cellular metabolic rate[23–25].

Although cyclic voltammetry and circular dichroism have different roles and purposes, they both can be applied to analyze the folate molecules in various contexts. The electrochemical activities and the redox properties of the folate and other pterin based molecules could be studied using cyclic voltammetric analysis while circular dichroism can be helpful in providing information related to the presence of chiral molecules as well as their interactions with other molecules. Both of these techniques could complement each other in finding the multifaceted properties of folate for its therapeutical use[26].

The role of folate has previously been mainly emphasized in the synthesis of nucleic acids and amino acids[27], while current study throws light on the new role of folate and intermediate molecules of the folate synthesis pathway including PABA, 6-methylpterin and pteroic acid secreted by fermenting bacteria in maintaining redox balance under anaerobic conditions.

PABA is an aromatic compound with a carboxylic and amino group attached to a benzene ring. It participates in the electron transfer during the synthesis of dihydropteroic acid[28].

PABA has also been reported to be essential for various human pathogens but not indispensable for humans, thus it could be modified and used as an antimicrobial agent. Martin Krátký et al.[28] have suggested that hybridization of PABA and aromatic aldehydes resulting in Schiff bases, this derivatization procedure actually converts the non-toxic PABA into an effective antimicrobial agent and even shown to exhibit anticancer activity towards HepG2 cell line. While 6-methylpterin, a derivative of pterin with an attached methyl group, is a heterocyclic molecule present in various biomolecules. It involves the charge transfer process in biochemical reactions where it forms a cofactor and coenzyme, such as tetrahydrobiopterin participate in many enzymatic processes including the synthesis of neurotransmitter[29].

Another intermediate molecule is pteroic acid which is formed of a pterin ring and a moiety of PABA; its charge transfer ability mainly depends on its role as a precursor molecule in biosynthesis and other metabolic reactions[30].

The folate molecule contains a pterin ring, a PABA moiety and a polyglutamate tail. And, it is mainly involved in the synthesis of nucleic acid and repair. Its charge transfer ability mainly occurs during the interconversion of its various forms including, dihydrofolate and tetrahydrofolate. Further, the charge transfer property of the folate is crucial when it acts as a coenzyme in case of one-carbon metabolism and participates in methyl group transfer in various biochemical reactions[31].

Additionally, folate has been found to be correlated with the regulation of oxidative–nitrosative stress and it also acts as a scavenger of hydroxy ion radicals[32]. Also, protein acting as ROS sinker has been previously shown to decrease with low levels of folate molecules. Similarly, the mitochondrial glutathione level has been reported to be suppressed with the low levels of folate[29,32,33].

The present study attempts to provide an understanding of the underlying mechanisms of the individual and synergistic role of these biomolecules (pterin pool) through a charge transfer to balance the oxidative stress

caused by the absence of oxygen. Since oxidative stress caused by redox imbalance generates reactive oxygen species (ROS) which has been found to be significantly associated with colorectal cancer[8,13]. Therefore, the given analysis may be helpful in developing therapy to treat colorectal cancer.

Experimental section

Anaerobic set-up

Oxidative stress has been applied to fermenting bacteria of *Providencia rettgeri* by culturing them in an anaerobic chamber wherein oxygen was removed from the bacterial media by purging the nitrogen gas. The chamber comprises of a three-electrode system: graphite electrode as working electrode, Ag/AgCl as reference electrode, and platinum as control electrode[20]. 1 ml of *Providencia rettgeri* broth was added to 20 ml of acetate growth media. The biomolecules released in the bacterial spent media were extracted on days 2, 10, and 20 via a solvent-enriched method, following the biomolecule isolation process as given in our previous study[20]. Briefly, the bacterial spent media was centrifuged for 20 min at 8000 rpm, 4 °C. Then, the supernatant was collected and further purified by organic solvents of chloroform and methanol. Here, the chloroform and methanol were used in a 2:1 ratio (volume), and approximately its equal volume was added dropwise to the centrifuged supernatant. The solution was mixed thoroughly and left undisturbed until the aqueous and organ phases were separated. Biomolecules were gradually migrated into the organic phase because of their solvent affinity. This organic phase with biomolecules was collected gently using an auto pipet[34].

MALDI-TOF mass spectrometry

Detailed information regarding MALDI-TOF mass spectroscopy has been given in our previous publication[20]. MALDI-TOF mass spectrometric measurements were conducted using BRUKER Microflex (Bruker 168 Daltonics, Billerica, MA) with a 2,5-dihydroxybenzoic acid (DHB) matrix. Here, the sample of 1 µL was kept at the parafilm following addition of 1µL of DHB, the solutions were mixed thoroughly with a pipette and the mixture was kept in MSP 96 target ground steel plate. The reduced pressure was applied to dry the sample. Subsequently, the plate containing the sample was set for the mass spectroscopic measurements. Afterward, the data was obtained from 0 to 800 m/z at 60 Hz laser frequency, with 500 laser shots. The subsequent mass spectrum was binned at increments of 0.5 m/z and inspected manually for masses of interest. Masses were further imagined by false coloring using an optical image of the original sample plate.

Photoluminescence study

The photoluminescence (PL) spectra was obtained using a spectrofluorometer (Flourolog-3, Horiba Jobin Yvon Inc.) having a 450 W xenon lamp. The excitation and emission spectra were recorded using a quartz cuvette whereas the forward geometry stage was used to remove the internal filter effect.

Electrochemical analysis

Cyclic voltammetry (CV) was performed with top and bottom peak potentials of + 0.8 V and − 0.8 V at scan rates of 10 mV/s using Autolab AUT 85279 Potentiostat Galvanostat (Metrohm, Netherlands). NOVA 1.11 software was used to examine the recorded data while Origin 8.0 software was used for plotting the data obtained from NOVA 1.11. A three-electrode system was used to analyze the electrochemical activity in which a graphite electrode was employed as a working electrode, Ag/AgCl as the reference electrode and platinum as the counter electrode. The study was carried out to analyze the involvement of the released molecules in the redox reactions. The spent bacterial media at 2 day, 10 day and 20 day were used as an electrolyte.

ATR-FTIR

The attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectroscopy was conducted with all samples in liquid form using PerkinElmer (USA) instrument. All spectra were measured in the transmission mode at a data acquisition rate of 5 cm^{-1} per point in the 4000 to 500 cm^{-1} range with 50 consecutive scans per sample, followed by stacked spectra analysis for relative comparison.

UV–Vis absorption spectroscopy

The UV–Vis spectroscopic studies were performed using a UV/Vis spectrophotometer, Lambda 950, (PerkinElmer, USA) to evaluate the behavior of released biomolecules in the bacterial spent media. Purification via the solvent extraction method has been previously described[20]. Briefly the bacterial spent media was centrifuged at 8000 rpm for 20 min at 4 °C. After that the supernatant was collected and purified by organic solvents of chloroform and methanol in 2:1 ratio (volume).

Circular dichroism

Circular Dichroism (CD) spectral analysis of 2, 10 and 20-day-old bacterial spent media was conducted using AVIV CD spectrometer (Biomedical Inc., NJ, USA). A quartz cell of 1 mm path length was used and the bandwidth was taken to be 1.0 nm. The CD spectral measurements were based on 10 consecutive scans with a multi-scan wait of 0.50 s. The response time was 1.0 s and the settling time was 0.3 s. All measurements were performed at 25 °C.

Results

In the present study, the anaerobic condition of the colon part of our intestine (Fig. 1) was maintained in the anaerobic chamber of the MFC. The biomolecules isolated from bacterial spent media at day 2, 10 and 20 in our previous study[20] were identified here in more detail using Mass spectroscopy (Fig. 2) to understand the ability of intermediate molecules to transfer charge and maintain redox equilibrium.

MALDI-TOF–MS

Mass spectroscopy was conducted to identify the released biomolecules by the bacteria under anerobic conditions. Since the relative molecular mass of folate or pteroylglutamic acid (vitamin B9) is reported to be 441.4 g/mol[35,36], hence the mass peak $[M^+ + H^+]$ appeared at 441.57, 441.42 and 441.30 m/z in day 2, 10 and 20 day old spent bacterial media, respectively (Fig. 2), indicating the release of folate molecules. Further, mass peaks at 136.60 and 136.47 m/z indicated the release of PABA[37] in day 2 and day 20 samples (Fig. 2a,c), respectively. PABA in addition to be a precursor of folate molecules, has been previously reported to reveal the antioxidant ability and has also been commercially used for UV blocking action in sunscreen[38]. It appears that PABA perhaps plays an important role in protecting the bacterial cell from the antioxidants produced in the provided anoxic condition. While the mass peak found at 176.63 m/z in day 10 (Fig. 2b) bacterial spent media indicates the release of 6-methylpterin[39] and the mass peak seen at 310.01 and 310.14 m/z indicates the presence of the oxidized form of pteroic acid[40] in day 2 and day 10 samples of bacterial spent media (Fig. 2a,b). Thus, the mass spectroscopy results reveal the presence of deprotonated folate and folate synthesis intermediates (PABA, 6-methylpterin, pteroic acid) called pterin pool.

As the redox imbalance can be further traced to linkage with the unregulated quantity of vitamins B, folate in particular that has been reported to reduce the oxidative stress[9,41,42]. Therefore, the release of folate molecules in all samples, indicates the molecule's crucial role in balancing the oxidative stress by temporarily storing or transferring the charge[43]. In order to further understand the activity of the moieties and to -determine whether they are affecting each other at the molecular electronic level in oxygen-deprived conditions, circular dichroism (CD) study was conducted.

Circular dichroism (CD) study

The CD spectra of the folate and pterin pool appear to be in the UV visible region. As these molecules have a pteridine ring that comprises of the conjugated double bonds, the spectral peaks found at 280 nm in the CD spectra of the pterin pool indicate characteristic bands of pteridine, which is due to the pi-electron transitions at the molecular level whereas absence of the 370 nm peak reveals that the moiety of PABA does not largely contribute to the CD signal[44,45]. Here the PABA possibly is not in a conformation that could significantly contribute to the CD signal. Furthermore, pteridine derivatives such as 6-methylpterin have been shown to exhibit CD signals at 280 nm and 320 nm (Fig. 3), which are attributed to the interaction of the pteridine ring with

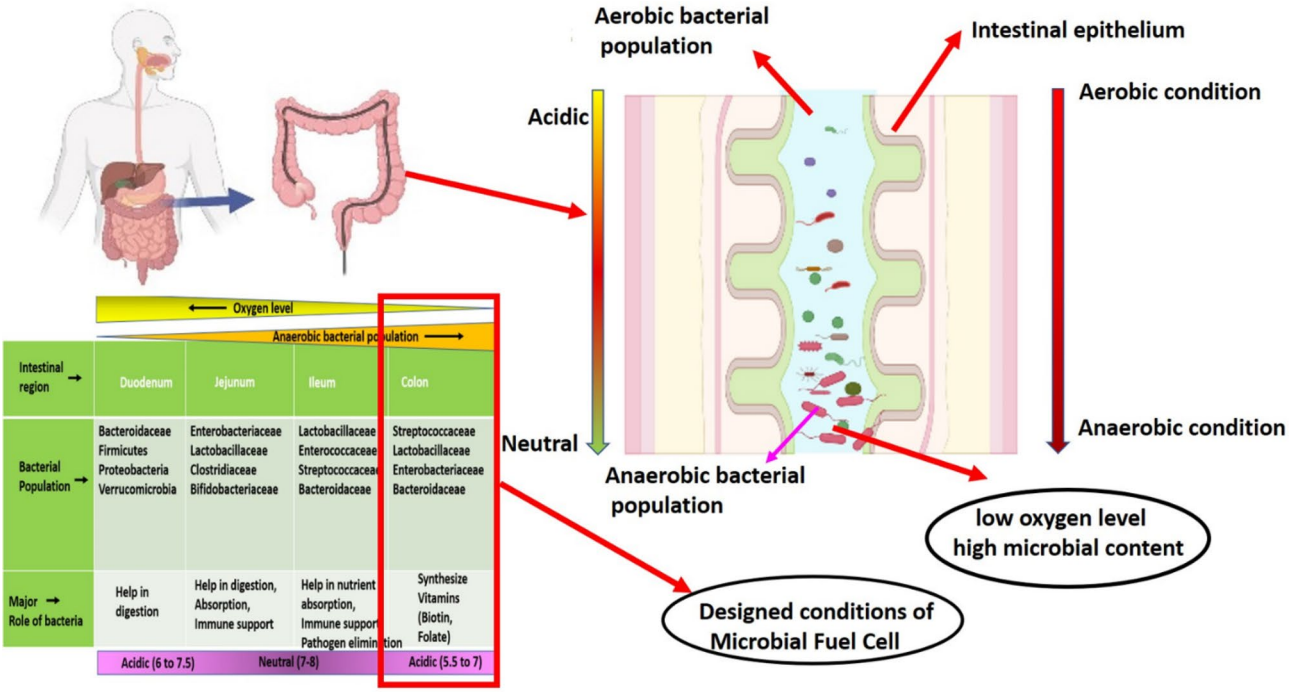


Figure 1. Schematic picture (image created with the help of BioRender software) showing the similarity between the conditions of low oxygen level and high microbial content in the colon region of the gastrointestinal tract and the anodic chamber of a microbial fuel cell.

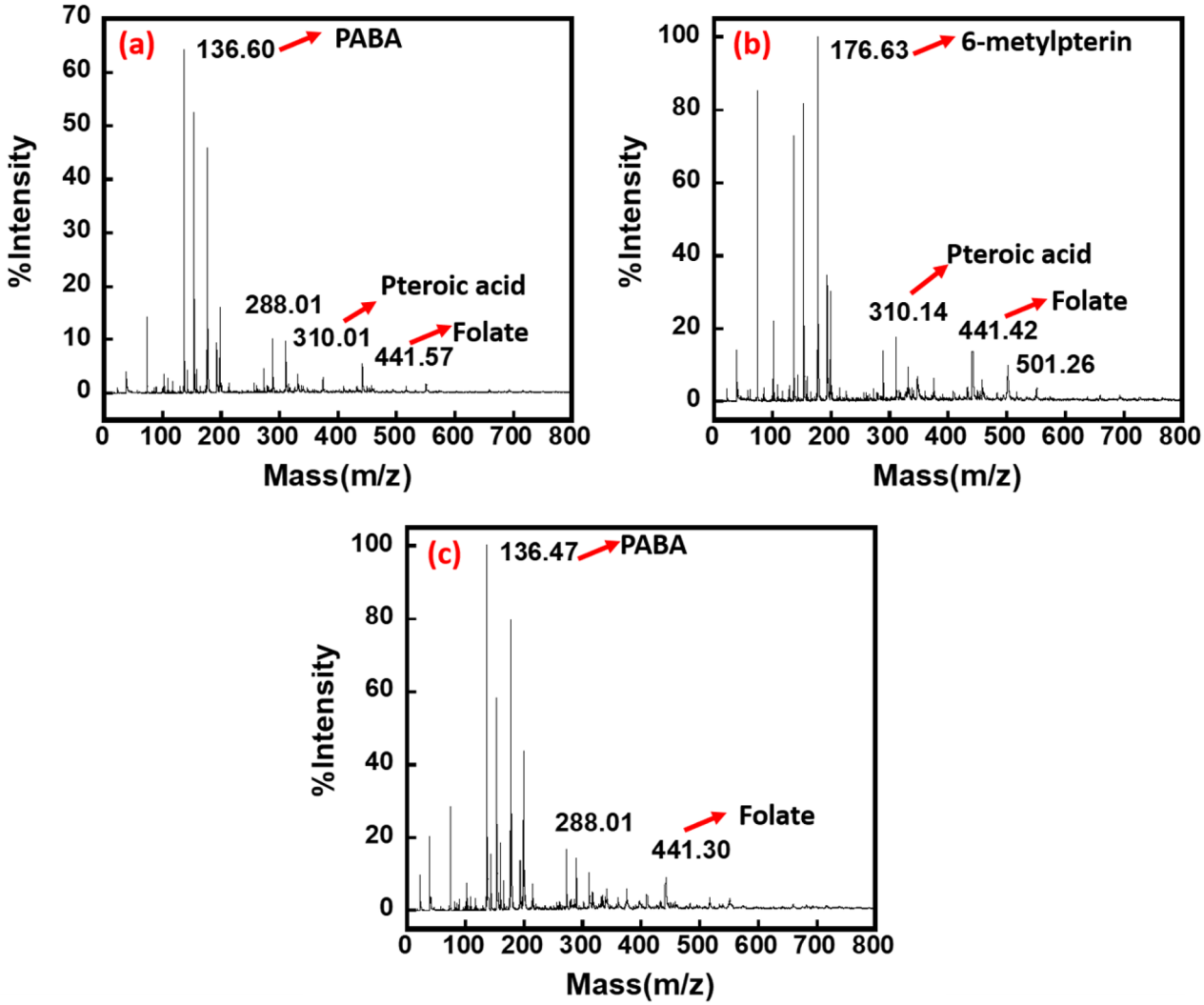


Figure 2. Mass spectroscopic analysis of day 2 (**a**), day 10 (**b**) and day 20 (**c**) samples of bacterial spent media, indicated the release of folate, pteroic acid, 6-methylpterin and para-aminobenzoic acid (PABA) biomolecules.

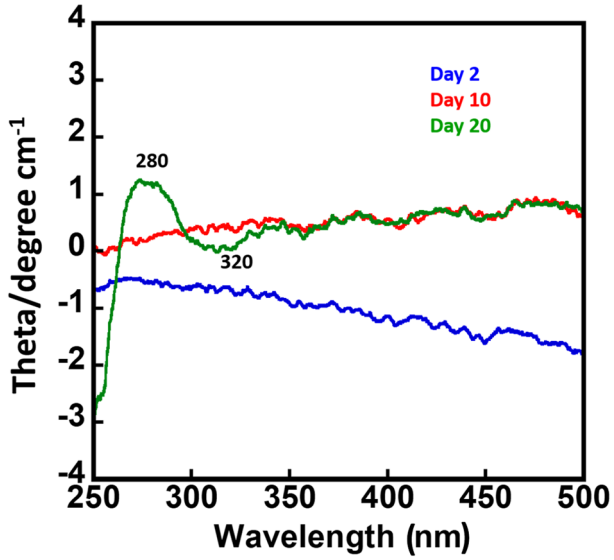


Figure 3. Circular Dichroism of day 2, day 10 and day 20 bacterial spent media. Spectral peaks appeared at 280 and 320 nm in day 20 sample confirm the presence of pteridine ring.

neighboring aromatic rings and the overall molecular conformation, respectively[46]. The CD spectral peaks are not clearly visible in day 2 and 10 samples indicating the predominant presence of precursors of PABA and pteroic acid may hinder the chiral activity of the folate and pteridine derivatives of 6-methylpterin. The release of 6-methylpterin in day 10 of bacterial growth (Fig. 2b) and its chiral activity in the CD spectra (Fig. 3) anticipate that the molecule may actively participate in molecular electronic transitions due to its favorable conformational arrangement. While the role of PABA and pteroic acid may be corelated to the free radical scavenging action to relieve the oxidative stress generated due to the anoxic condition. Molecular electronic transitions were further analyzed using UV–vis absorption spectroscopy.

UV–Vis absorption spectroscopy

In the UV spectroscopy study, the molecules of PABA that contribute to the absorption bands at 250 nm in the UV spectrum[47] were predominantly found in day 2 and day 10 samples (Fig. 4). This UV spectrum band is due to the pi–pi* electronic transition of the PABA moieties[48]. Further, a spectral band seen at 325 nm which is considerably higher at day 20 (Fig. 4) attributed to the n–pi* transition of the amino group of the pteridine ring system of the 6-methylpterin and folate[49]. Although the peak at 325 nm can also be seen on day 2 and day 10, however, the intensity of the peak on day 2 is much lower, which may be due to the low expression of folate and 6-methylpterin molecules in this time period (Fig. 8). Further, at day 20, the UV spectral peak at 280 nm can be assigned to the pi–pi* transition of the pteridine ring representing the main chromophore of the folate molecule[50,51]. Overall, the UV spectral study of the molecules exhibits the peak corresponding to the pteridine ring, glutamic acid, and PABA, and no spectrum pertaining to individual molecules of the folate, 6-methylpterin and pteroic acid are seen (Fig. 4), indicating that molecules interact with each other. Previously, it has been reported that the peaks at 255 nm, 286 nm, and 366 nm are characteristic of the chromophores present in the folate molecule[52].While in the present study, the peaks around 250, 280 and 325 nm reveal the inter and intramolecular interaction of the moieties of PABA, 6-methylpterin, pteroic acid and folate molecules to balance the absence of oxygen[43]. The functional group interactions were further investigated by an ATR-FTIR spectroscopy study.

ATR-FTIR study

As the isolated molecules comprise of folate, 6-methylpterin, pteroic acid, and PABA which are all pterin derivatives and also have some common functional groups, there are differences in the side chains attached to the pterin ring. Consequently, their ATR-FTIR spectra may be related to the presence or absence of peaks pertaining to specific functional groups. The spectral peaks seen at 1221 cm^{-1} and 1367 cm^{-1} may be due to the C–N stretching vibrations while those at 1743 cm^{-1} and 1642 cm^{-1} may be due to the stretching vibrations of the carbonyl group (C=O) and carbon–carbon double bond (C=C) in the aromatic ring, respectively[53].

There is an evident change in the intensity of the spectral peaks especially in the 10 day sample indicating fermi resonance in the spectral peaks at 1221 cm^{-1} and 1367 cm^{-1} as well as 1743 and 1642 cm^{-1}[54]. The peak at 1435 cm^{-1} is possibly due to C=C stretching vibrations present in the aromatic ring of pterin while 1221 cm^{-1} peak could be due to C–N stretching vibrations of the functional groups present in the folate molecule[55,56].

In the case of the day 2 sample, a sharp peak at 1011 cm^{-1} can be observed, this peak may be assigned to the C–N stretching vibration of the amino group present at the para position of the benzene ring of PABA. While no clear peak at this position has been observed at day 10 and day 20 samples, probably due to the presence of

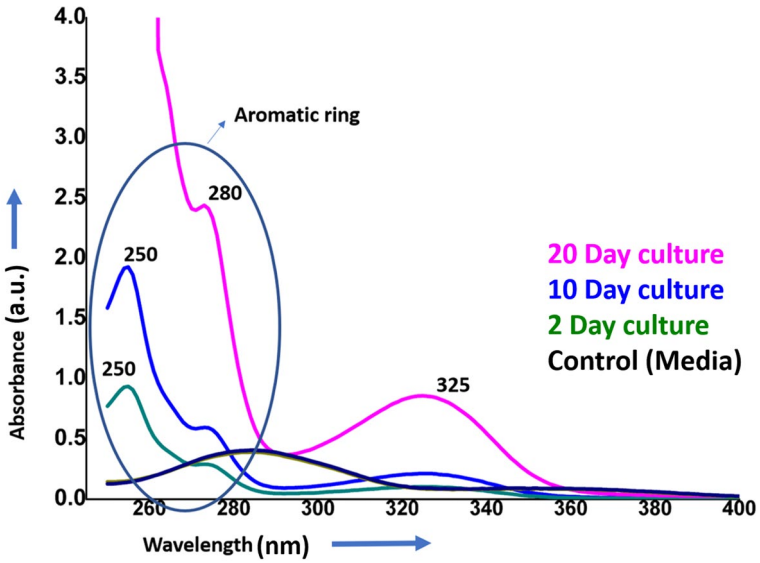


Figure 4. UV spectroscopy studies of folate and folate intermediates released into spent media of bacteria at 2, 10 and 20 days of bacterial growth showing spectral absorption due to para-aminobenzoic acid (PABA) and pteridine ring system.

other biomolecules of the pterin pool (pteroic acid, 6-methylpterin and folate) influenced the intensity and peak position of the functional groups of PABA[57].

In the day 20 sample, the spectral peaks observed at 3324 cm^{-1} and 1642 cm^{-1} (Fig. 5) are due to the N–H stretching and C=O stretching vibrations of the glutamic acid side chain, respectively[58]. The spectral peaks seen at 650 cm^{-1}, 556 cm^{-1}, and 457 cm^{-1} in the day 20 sample (Fig. 5) may be attributed to the out-of-plane bending of the C–H bonds and N–H bonds in the pterin moiety of the molecule[59].

In the present study the 3324 cm^{-1} peak (Fig. 5) is attributed to the N–H stretching vibration of the amide group (–CONH–) in the glutamic acid side chain of folate. This has perhaps partly overlapped due to the O–H stretching vibration present in molecules of water. And, the peak at 1642 cm^{-1} (Fig. 5) could be due to C=O stretching vibration of the amide group (–CONH–) in the glutamic acid side chain of folate[60]. The peak at 650 cm^{-1} (Fig. 5) can be associated with the out-of-plane bending of the C–H bonds in the aromatic ring of the pterin moiety of folate. While the peak at 556 cm^{-1} is perhaps due to the in-plane bending of the C-H bonds in the aromatic ring of the pterin moiety of folate and 457 cm^{-1} peak can be associated with the out-of-plane bending of the N–H bonds in the pterin moiety of folate[61–63].

ATR-FTIR spectroscopy results show that pi–pi complexation of aromatic rings is due to charge donor and acceptor interactions which also induce the protonation process in folate molecules upon strengthening the electron-accepting ability of pterin rings. These aromatic rings have both the electron acceptors and the pi donors[64]. In order to further understand which moieties of the identified biomolecules are active in the charge transfer, the electroactivity was analyzed via cyclic voltammetry.

Cyclic voltammetry

The cyclic voltammetry study has been conducted to identify the reduction potential and may be related electrochemical reactions occurring in the anaerobic chamber. The results of cyclic voltammetry, reveal that the biomolecules released at different time periods have a difference in reduction potential. The reduction potential at day 2, 5 and 10 was observed at − 0.4 V with a shift towards more negative value indicating the greater tendencies to accept the electrons. While two clear peaks could be observed at day 20 only at − 0.2 V and − 0.4 V, it indicates that there are two electroactive moieties in these molecules, one moiety has more tendency to accept the electrons. The difference in reduction potential indicates thermodynamic favorability with the electron transfer from one species to another species each has a different reduction potential. However, as different electroactive species co-exist in a pterin pool indicates that there is a multiple electron transfer process. The two redox peaks found at − 0.4 and − 0.2 V (Fig. 6) can be associated with the reduction potentials of the pterin moiety and p-aminobenzoate groups, respectively of the pterin pool[65]. The peak observed at − 0.4 V depicts a comparatively more negative reduction potential, indicating that the pterin moiety perhaps easily gain the electrons and more easily undergo the reduction compared to the p-aminobenzoate group[66,67].

It appears that intra-molecular electron transfer occurs with the transfer of an electron between the pterin and PABA groups of the same molecule while inter-molecular electron transfer occurs with the electron transfer from the pterin moiety of the one molecule and PABA of another molecules. These redox properties play an essential role in the cellular metabolic pathways involving the synthesis of DNA and methyl group transfer reactions.

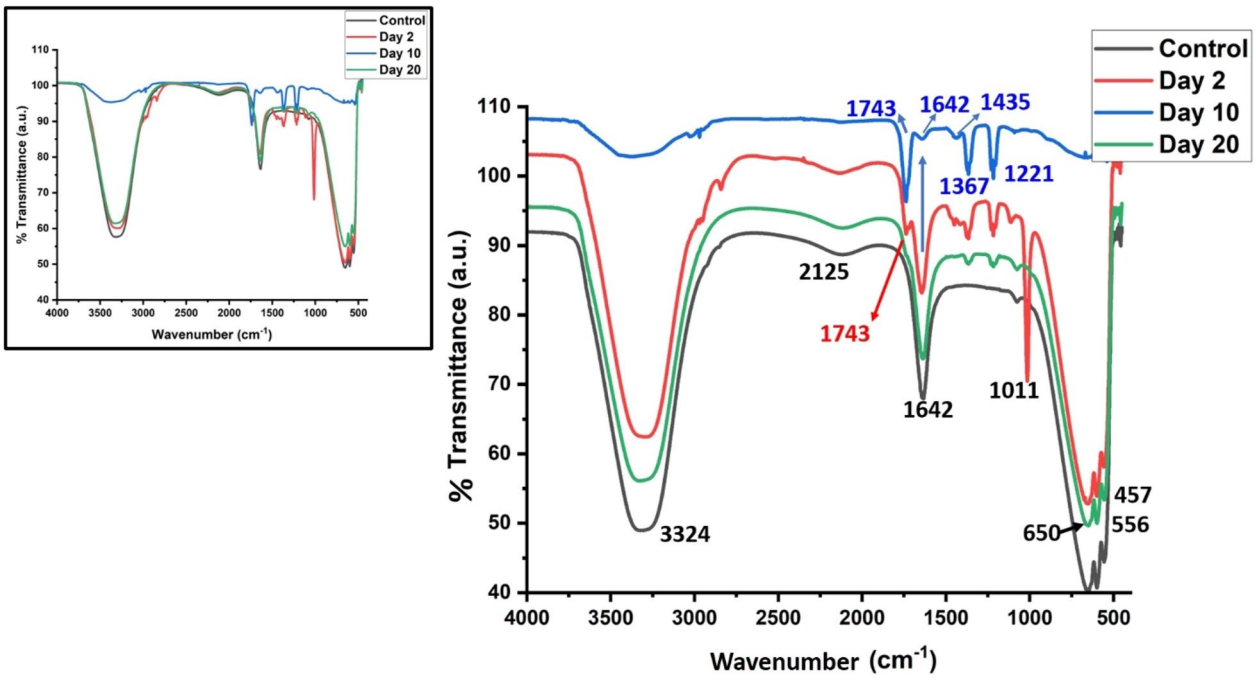


Figure 5. Stacked ATR-FTIR spectra of bacterial spent media at different time period of growth (Inset showing FTIR spectra of the same without stacking).

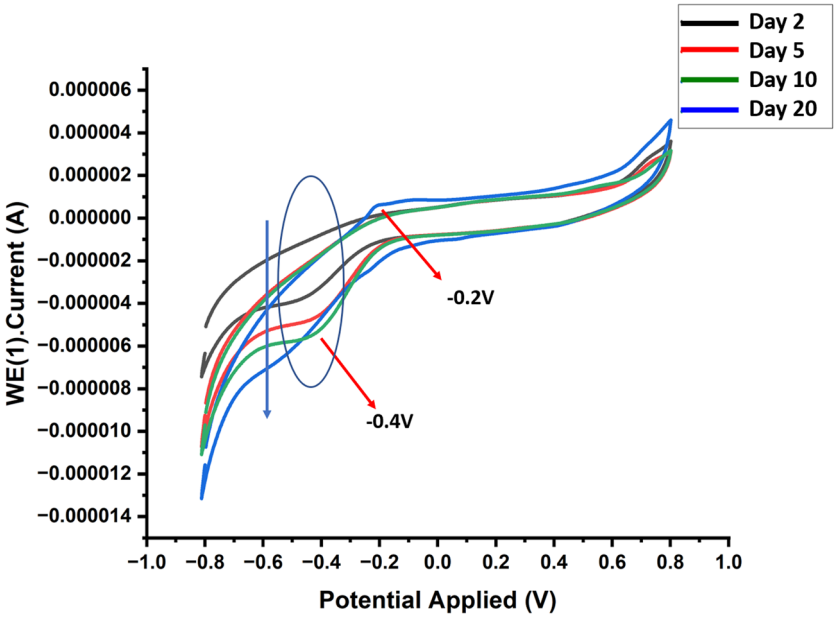


Figure 6. The cyclic voltammetry pattern changes with bacterial growth under anaerobic condition. The dip in curve depicting the high redox ability of the released molecules of folate and folate intermediates especially of pterin moiety and p-aminobenzoate groups.

Thus, based on the results of cyclic voltammetry, the presence of two redox-active moieties can be confirmed with different reduction potentials, and that the pterin moiety has a lower reduction potential than the PABA group. Therefore, it could be deduced that the inter and intra molecular electron transfer occurs through the PABA and pterin moieties in the absence of oxygen. Furthermore, the electrochemical behavior of the pterin pool present in a solution, depends on their relative concentrations and redox properties, determining the transfer of charge between the two molecules under the applied electrode potential. There is also a high probability of transfer of charge from 6-methylpterin to pteroic acid through an oxidation–reduction reaction, where the molecules of 6-methylpterin can easily donate the electrons to the molecules of pteroic acid. As a result, the radical anion of 6-methylpterin ions and a radical cation of pteroic acid are formed in the solution. Various factors, such as kinetics of the charge transfer and the stability of the resulting radical species could further control the efficiency of this charge transfer[68]. The probability of maintaining the redox equilibrium by charge transfer within various available energy levels has been further confirmed with the photoluminescence study.

Photoluminescence study

The photoluminescence (PL) spectra was obtained using excitation wavelengths of 310, 325, 350 and 365 nm and the emission was obtained at 410, 416, 429 and 440 nm on day 2: at 419, 419, 431 and 446 nm on day 10; at 440, 445, 425 and 420 nm on day 20; and at 410, 440, 422 and 436 nm on day 30. There is a clear shift in emission wavelength when excited by 350 nm from 440 (day 20) to 429 nm and 431 nm at day 2 and day 10 (Fig. 7), respectively. The maximum emission occurs at 440 nm when excited by 350 nm, indicating aromatic rings of folate molecules[69] (Fig. 7).

Pterins are aromatic compounds with heterocyclic rings and are involved in various photobiological processes[70]. The change in emission spectra with the change in excitation wavelength indicates that there are various tuneable optical band gaps for the transfer of electrons that may be created through the splitting of energy levels[71]. As a result, intra and inter-molecular electron transfer within the folate and pterin based molecules perhaps maintain the redox balance. One possible explanation for this observation is that the excitation energy is perhaps transferred between the molecules of 6-methylpterin, folate, and pteroic acid; these may interact through electron transfer, proton transfer, or dipole–dipole interactions[64]. These interactions could result in the transfer of energy between the molecules and changes in their photoluminescence emission spectra which could have important implications for their biological functions and properties[72].

Role of folate and folate intermediates in redox activities

Here it is hypothesized that initially, pterin based intermediates which are actively involved in the folate synthesis are present in reduced form viz. dihydropterins and tetrahydropterins, Pterin dihydroform may participate in the synthesis of folate while tetrahydro form acts as pterin cofactor. Later, they may both undergo auto-oxidation forming fully oxidized aromatic pterins called biopterin and, these aromatic pterins in oxidized form perhaps do not participate in the synthesis of folate[73]. This is supported by MALDI and cyclic voltammetry study, MALDI data (Fig. 2) depict the presence of 6-methylpterin at day 10 in the exponential phase of the growth curve and

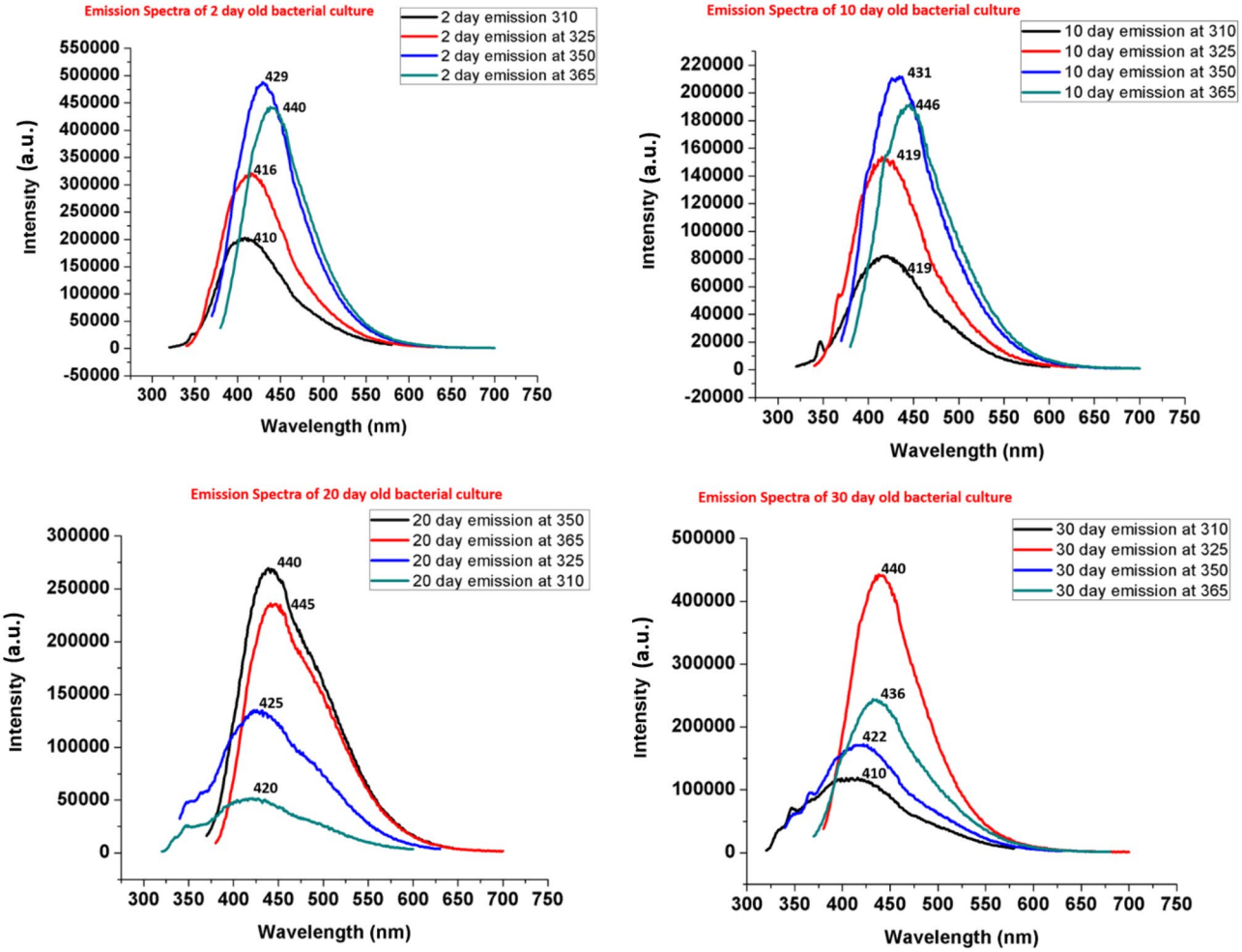


Figure 7. Photoluminescence emission spectra of bacterial spent media at 2 day (**a**), 10 day (**b**), 20 day (**c**) and 30 day (**d**) showing the release of folate and intermediates at different growth phase.

cyclic voltammetry (Fig. 6) indicating the active role of pterin moiety as redox mediators to transfer the charge to the graphite electrode.

While studying the behavior of pterin based molecules under anaerobic condition in terms of time, at day 10, the solution turns pale yellow indicating biopterin that may be formed through auto-oxidation with carboxyl and hydroxyl functional group of dihydropterin[74,75] as can be seen in Fig. 8, inset (B), this is further supported by the MALDI results (Fig. 2). At day 20, biopterin undergoes further oxidation and forms quinoid tetrahydrobiopterin[75] with two keto groups and one hydroxyl group that are observed to be blue in color (Fig. 8C). This pterin pool exhibits photoluminescence, which possibly is due to the presence of a pteridine aromatic ring with the conjugated system of double bonds and delocalized electrons resulting in the formation of excited energy states. The critical information in the pterin pool is the splitting of energy levels where electrons could switch from one energy level to another. Furthermore, the formation of quinoid tetrahydrobiopterin possibly hinders the transfer of charge, and may cause redox imbalance. Therefore, the presence of quinoid tetrahydrobiopterin reveals the level of oxidative stress in the gut region and can be used as a biomarker. Moreover, the photoluminescence exhibited by the quinoid tetrahydrobiopterin can be used as a diagnostic tool to analyze oxidative stress and ROS level.

Discussion

The results of this study suggest the release of folate and intermediates of folate synthesis including PABA, 6-methylpterin, and pteroic acid, by the facultative bacteria under the given anoxic condition. Thus, bacteria have been observed to transfer charge on graphite electrodes in the absence of oxygen via folate and its intermediates (please see the details in our previous research paper[20]).

The bacterial growth curve and its electron transfer ability show a sigmoid curve (Fig. 8), suggesting the transfer of the electrons differentially at different period of time according to the released biomolecules in the media. These biomolecules may have differential redox properties. The confirmation of PABA at day 2 in MALDI analysis (Fig. 2), suggest that initially (lag phase of the bacterial growth curve) PABA biomolecules are perhaps released to protect the bacterial cell from oxidative stress. There are various by-products which are produced during the anaerobic metabolism of the bacterial cells especially within enteric bacteria, such as nitric oxide,

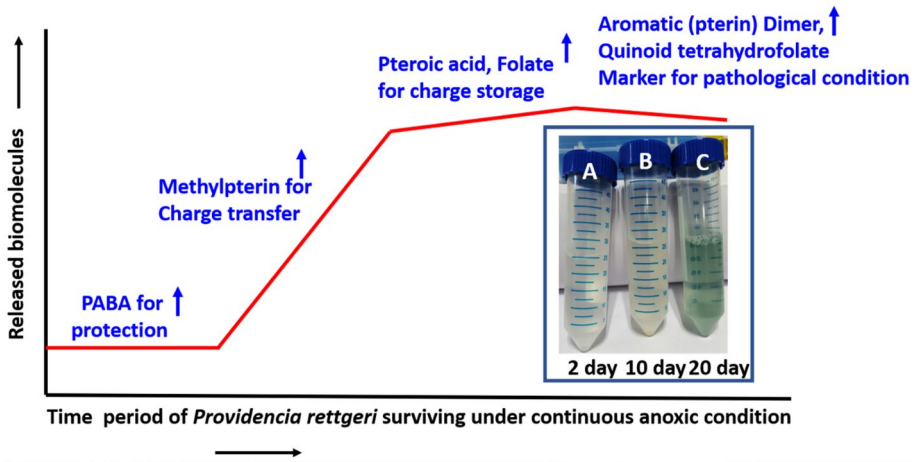


Figure 8. Schematic graph for illustrating the role of released folate, folate intermediates (PABA, 6-methylpterin, pteroic acid) and pterin dimer at the different time period of anoxic environment leading towards pathological conditions. Inset showing change in color as transparent (day 2), yellow (day 10) and blue-green (day 20).

which resembles the superoxide generated during the natural process of respiratory system[76,77]. Thus, PABA presence may protect against the DNA damage from the nitric oxide like by-products[78]. As it has been established previously that PABA apart from being used as a precursor of folate molecules, can protect the intracellular environment of the bacterial cells against free radicals, including reactive oxygen species (ROS) and reactive nitrogen species (RNS) by directly interacting with radicals of hydroxyl ions which prevent deoxyribose oxidation [47,79]. Furthermore, Svetlana V Vasilieva et al. have reported that the total ROS generated during the redox reactions have been found to be decreased with the increase in PABA cellular level[78]. Thus, PABA effectively engrosses ROS such as hydroxyl radicals, singlet oxygen and hypochlorous acid and protects the DNA from the oxidative damage[78].

In addition to PABA, the presence of folate and 6-methylpterin has been confirmed (Fig. 2). While folate and 6-methylpterin have antioxidant properties, their mechanisms of action are different. Folate acts primarily as a hydrogen donor to neutralize free radicals, whereas 6-methylpterin directly scavenges free radicals[80,81]. Folate also has other functions in the body beyond its antioxidant activity, such as its role in DNA synthesis and repair, and its impact on methylation pathways[82]. On the other hand, 6-methylpterin is a specialized antioxidant compound that is specifically designed to scavenge free radicals[83]. Overall, both folate and 6-methylpterin have important roles in maintaining antioxidant defenses and protecting against oxidative damage in the body.

Herein, the oxidation state of the pterin ring is perhaps the deciding factor assigning its bifunctional role because the reduced form of pterin such as dihydro or tetrahydro form of pterin moiety may perhaps participate in folate synthesis whereas the fully oxidized form may perhaps not undergo folate synthesis and exist as a pterin pool creating various energy levels to donate and accept the charge (Fig. 7) because a cell may need them for their redox balance[73]. Further, the oxidized pterin (biopterin) may help to prevent from oxidative damage as it could efficiently act as an antioxidant and scavenge the ROS and RNS produced in the oxygen deprived condition. Additionally, it could donate the electrons to the thioredoxin reductase and glutathione peroxidase another antioxidant systems[84].

In the present study, folate ions present at low concentration (< 10 mm) on day 2 possibly exist as unfolded and extended conformation. However, with the increased concentration of folate ions (day 10 and 20), the folate ions may undergo the intermolecular interactions and resulting in vertical stacking of this biomolecule[85,86]. Intramolecular interactions may involve the interactions between pteridinyl and PABA moieties and intermolecular interactions revealing that one pteridinyl moiety perhaps interacts with PABA of other molecule[87]. Here, it is possible that charge transfer (proton transfer) may occur through extended hydrogen bonding in an aqueous solution[88].

Further, recently, it has been reported that decreased folate levels in oxidative stress is correlated with a high probability of colorectal cancer. Stef Jose et al. have conducted the study using the model gut bacterium Enterococcus durans (MTCC 3031) under oxidative stress and have shown that oxidative stress results in decreased folate synthesis by the gut bacteria which is not found to be favorable for the growth of the host cell and deficiency may result in colorectal cancer[33]. Furthermore, as the oxidized folate molecules more easily undergo dimerization relative to the reduced form of folate ions, the presence of dimer under prolonged oxygen deprived condition (Fig. 8) possibly indicates a diseased cellular state. Therefore, in normal physiological conditions bacterial cells have been reported to exhibit oscillatory expression[33]. However here we have observed that in the hypoxic condition they exhibit sigmoid behavior (Fig. 8). In addition, the bacterial biomass has been found to be increased with an increase in redox ratio while under the oxidative stress if the redox ratio falls, then there is a decrease in the folate and bacterial growth.

Overall, the absence of oxygen is balanced by various energy levels created by folate and pterin derivatives and redox imbalance may be due to deprivation of various energy states due to dimerization of pterin derivatives.

Moreover, depending on the surrounding conditions, bacterial secreted molecules may undergo dimerization or persist as discrete units or synthesize larger biomolecules such as folate. This information may be helpful in understanding how the bacterial cells behave in the gut and how the pathological condition progresses. Therefore, it can be suggested that strategies of cancer treatment should focus on the redox related causes[89].

Conclusions

It can be inferred from the present study that the released folate and pterin derivatives by fermenting bacteria (*Providencia rettgeri*) in the anaerobic chamber of MFC undergo inter as well as intra-molecular electron transfer forming different intermediate states for the long-range transfer of electrons. Analysis of the electrochemical and photoluminescence properties of folate molecules and pterin pool has provided a mechanistic understanding of charge transfer across split energy levels required for maintaining cellular redox balance, with promising applications in the management of GI diseases. Under prolonged anoxic conditions, the dimer of the aromatic ring present in pterin derivatives perhaps acts as a temporary storage of charge and is unable to transfer the charge, potentially inducing a diseased state by redox imbalance. This information can further be used for analyzing the progression of the disease.

Data availability

The data that support the findings of this study are available from the corresponding author on reasonable request.

Received: 3 July 2023; Accepted: 10 November 2023
Published online: 18 November 2023

References

1. Chen, Y. *et al.* Bioengineered 3D tissue model of intestine epithelium with oxygen gradients to sustain human gut microbiome. *Adv. Healthc. Mater.* **11**, 2200447 (2022).
2. Vissenaekens, H., Grootaert, C., Rajkovic, A., Van De Wiele, T. & Calatayud, M. The response of five intestinal cell lines to anoxic conditions in vitro. *Biol. Cell* **111**, 232–244 (2019).
3. Albenberg, L. *et al.* Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. *Gastroenterology* **147**, 1055–1063 (2014) (**e1058**).
4. Loesche, W. J. Oxygen sensitivity of various anaerobic bacteria. *Appl. Microbiol.* **18**, 723–727 (1969).
5. Macfarlane, G. & Gibson, G. R. Metabolic activities of the normal colonic flora. in *Human Health: The Contribution of Microorganisms*, 17–52 (1994).
6. Rolfe, R. D., Hentges, D. J., Campbell, B. J. & Barrett, J. T. Factors related to the oxygen tolerance of anaerobic bacteria. *Appl. Environ. Microbiol.* **36**, 306–313 (1978).
7. Bellali, S., Lagier, J.-C., Raoult, D. & Bou Khalil, J. Among live and dead bacteria, the optimization of sample collection and processing remains essential in recovering gut microbiota components. *Front. Microbiol.* **10**, 1606 (2019).
8. Liu, H. *et al.* Redox imbalance in the development of colorectal cancer. *J. Cancer* **8**, 1586 (2017).
9. Kok, D. E. *et al.* Bacterial folate biosynthesis and colorectal cancer risk: more than just a gut feeling. *Crit. Rev. Food Sci. Nutr.* **60**, 244–256 (2020).
10. Ginting, A. & Sueta, M. A. D. Characteristics of lower distal colorectal cancer treated with ultra-low anterior resection (ULAR) in Sanglah general hospital, Denpasar between January 1st 2014–December 31th 2016.
11. Schreuders, E. H. *et al.* Colorectal cancer screening: A global overview of existing programmes. *Gut* **64**, 1637–1649 (2015).
12. Lichtenstern, C. R., Ngu, R. K., Shalapour, S. & Karin, M. Immunotherapy, inflammation and colorectal cancer. *Cells* **9**, 618 (2020).
13. Basak, D., Uddin, M. N. & Hancock, J. The role of oxidative stress and its counteractive utility in colorectal cancer (CRC). *Cancers* **12**, 3336 (2020).
14. Patil, P. S. *et al.* Colorectal cancer in India: An audit from a tertiary center in a low prevalence area. *Indian J. Surg. Oncol.* **8**, 484–490 (2017).
15. Muz, B., de la Puente, P., Azab, F. & Azab, A. K. The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia* **3**, 83 (2015).
16. Narayan, T. *et al.* Protein functionalised self assembled monolayer based biosensor for colon cancer detection. *Talanta* **201**, 465–473 (2019).
17. Kim, R. *et al.* An in vitro intestinal platform with a self-sustaining oxygen gradient to study the human gut/microbiome interface. *Biofabrication* **12**, 015006 (2019).
18. Shah, P. *et al.* A microfluidics-based in vitro model of the gastrointestinal human–microbe interface. *Nat. Commun.* **7**, 11535 (2016).
19. Yang, J., Cheng, S., Li, P., Huang, H. & Cen, K. Sensitivity to oxygen in microbial electrochemical systems biofilms. *Science* **13**, 163–172 (2019).
20. Nara, S. *et al.* Exploring Providencia rettgeri for application to eco-friendly paper based microbial fuel cell. *Biosens. Bioelectron.* **165**, 112323 (2020).
21. Li, R. *et al.* Deciphering the structural diversity and classification of the mobile tigecycline resistance gene tet (X)-bearing plasmidome among bacteria. *Msystems* **5**, e00134 (2020).
22. Jandhyala, S. M. *et al.* Role of the normal gut microbiota. *World J. Gastroenterol.* **21**, 8787 (2015).
23. Sethupathi, R., Gurushankar, K. & Krishnakumar, N. Optical redox ratio differentiates early tissue transformations in DMBA-induced hamster oral carcinogenesis based on autofluorescence spectroscopy coupled with multivariate analysis. *Laser Phys.* **26**, 116202 (2016).
24. Gurushankar, K., Rimac, H., Potemkin, V. & Grishina, M. Investigation of the newly characterized baimantuoluoamide a and baimantuoluoamide b alkaloids as potential cyclin-dependent kinase 4 (CDK4) inhibitors using molecular docking and molecular dynamics simulations. *J. Mol. Struct.* **1230**, 129925 (2021).
25. Viswanathan, K., Soumya, K., Gurusankar, K. & Jeyavijayan, S. Raman spectroscopic analysis of ovarian cancer tissues and normal ovarian tissues. *Laser Phys.* **29**, 045701 (2019).
26. Rastakhiz, N., Beitollahi, H., Kariminik, A. & Karimi, F. Voltammetric determination of carbidopa in the presence of uric acid and folic acid using a modified carbon nanotube paste electrode. *J. Mol. Liq.* **172**, 66–70 (2012).
27. Abbasi, I. H. R. *et al.* Folate promotes S-adenosyl methionine reactions and the microbial methylation cycle and boosts ruminants production and reproduction. *Amb. Express* **8**, 1–10 (2018).

28. Krátký, M. *et al.* 4-Aminobenzoic acid derivatives: Converting folate precursor to antimicrobial and cytotoxic agents. *Biomolecules* **10**, 9 (2019).
29. Armarego, W. L. & Schou, H. Pterins. Part 2. Stereochemistry of catalytic reduction of 6-methyl-and 6, 7-dimethyl-pterin and of 2, 4-diamino-6-methylpteridine. *J. Chem. Soci. Perkin Trans.* **1**, 2529–2536 (1977).
30. Scott, J. M. *Methods in Enzymology* Vol. 66, 657–660 (Elsevier, 1980).
31. Sudimack, J. & Lee, R. J. Targeted drug delivery via the folate receptor. *Adv. Drug Deliv. Rev.* **41**, 147–162 (2000).
32. Lai, K.-G. *et al.* Novel roles of folic acid as redox regulator: Modulation of reactive oxygen species sinker protein expression and maintenance of mitochondrial redox homeostasis on hepatocellular carcinoma. *Tumor Biol.* **39**, 1010428317702649 (2017).
33. Jose, S., Bhalla, P. & Suraishkumar, G. Oxidative stress decreases the redox ratio and folate content in the gut microbe, *Enterococcus durans* (MTCC 3031). *Sci. Rep.* **8**, 12138 (2018).
34. Kumar, T., Singh, M., Purohit, H. & Kalia, V. Potential of *Bacillus* sp. to produce polyhydroxybutyrate from biowaste. *J. Appl. Microbiol.* **106**, 2017–2023 (2009).
35. Nelson, B. C. The expanding role of mass spectrometry in folate research. *Curr. Anal. Chem.* **3**, 219–231 (2007).
36. Stokes, P. & Webb, K. Analysis of some folate monoglutamates by high-performance liquid chromatography–mass spectrometry. I.. *J. Chromatogr. A* **864**, 59–67 (1999).
37. Xia, H., Zhang, Y., Pavlov, J., Jariwala, F. B. & Attygalle, A. B. Competitive homolytic and heterolytic decomposition pathways of gas-phase negative ions generated from aminobenzoate esters. *J. Mass Spectrom.* **51**, 245–253 (2016).
38. Li, Y., Tang, L.-P., Zhou, W. & Wang, X.-R. Fabrication of intercalated p-aminobenzoic acid into Zn-Ti layered double hydroxide and its application as UV absorbent. *Chin. Chem. Lett.* **27**, 1495–1499 (2016).
39. Hettich, R. L. & Jacobson, K. B. Structural characterization of naturally occurring pteridines by laser desorption Fourier transform mass spectrometry. *J. Mass Spectrom.* **30**, 872–884 (1995).
40. Gentili, A. *et al.* Simultaneous determination of water-soluble vitamins in selected food matrices by liquid chromatography/electrospray ionization tandem mass spectrometry. *Rapid Commun. Mass Spectr.* **22**, 2029–2043 (2008).
41. Kim, Y.-I. Folic acid supplementation and cancer risk: Point. *Cancer Epidemiol. Biomark. Prev.* **17**, 2220–2225 (2008).
42. Mason, J. B. *et al.* A temporal association between folic acid fortification and an increase in colorectal cancer rates may be illuminating important biological principles: A hypothesis. *Cancer Epidemiol. Biomark. Prev.* **16**, 1325–1329 (2007).
43. Li, G., Magana, D. & Dyer, R. B. Photoinduced electron transfer in folic acid investigated by ultrafast infrared spectroscopy. *J. Phys. Chem. B* **116**, 3467–3475 (2012).
44. Raghu, M. *et al.* Synthesis, characterization, antimicrobial and interaction studies of pteridines with human serum albumin: A combined multi-spectroscopic and computational study. *J. Mol. Struct.* **1250**, 131857 (2022).
45. Fang, Z., Hu, L., Gao, Y., Li, J. & Liang, L. Protection of resveratrol against the photodecomposition of folic acid and photodecomposition-induced structural change of beta-lactoglobulin. *Food Res. Int.* **102**, 435–444 (2017).
46. Jha, N. S. & Kishore, N. Thermodynamic studies on the interaction of folic acid with bovine serum albumin. *J. Chem. Thermodyn.* **43**, 814–821 (2011).
47. Hu, M.-L., Chen, Y.-K., Chen, L.-C. & Sano, M. Para-aminobenzoic acid scavenges reactive oxygen species and protects DNA against UV and free radical damage. *J. Nutr. Biochem.* **6**, 504–508 (1995).
48. Baibarac, M., Smaranda, I., Nila, A. & Serbschi, C. Optical properties of folic acid in phosphate buffer solutions: The influence of pH and UV irradiation on the UV-VIS absorption spectra and photoluminescence. *Sci. Rep.* **9**, 14278 (2019).
49. Shinkai, S. *et al.* Coenzyme models: 4.0 Spectral and reactivity studies of roseoflavin analogs: Correlation between reactivity and spectral parameters. *Bioorg. Chem.* **14**, 119–133 (1986).
50. Jaimes-Aguirre, L. *et al.* Biodegradable poly (D, L-lactide-co-glycolide)/poly (L-γ-glutamic acid) nanoparticles conjugated to folic acid for targeted delivery of doxorubicin. *Mater. Sci. Eng. C* **76**, 743–751 (2017).
51. Du, Z. *et al.* Synthesis, photophysical and cellular characterisation of folate and methotrexate labelled luminescent lanthanide complexes. *J. Inorg. Biochem.* **178**, 32–42 (2018).
52. Off, M. K. *et al.* Ultraviolet photodegradation of folic acid. *J. Photochem. Photobiol. B* **80**, 47–55 (2005).
53. Bertel Garay, L., Mendez Sanchez, S. C. & Martinez Ortega, F. Use in vitro of gold nanoparticles functionalized with folic acid as a photothermal agent on treatment of HeLa cells. *J. Mexican Chem. Soc.* **62**, 385 (2018).
54. Asundi, R. & Padhye, M. Fermi resonance in benzene. *Nature* **163**, 638–638 (1949).
55. Ullah, S. *et al.* 5-fluorouracil-loaded folic-acid-fabricated chitosan nanoparticles for site-targeted drug delivery cargo. *Polymers* **14**, 2010 (2022).
56. Jalilian, A. *et al.* Preparation and biological evaluation of radiolabeled-folate embedded superparamagnetic nanoparticles in wild-type rats. *J. Radioanal. Nuclear Chem.* **287**, 119–127 (2011).
57. Lieber, E., Rao, C., Pillai, C., Ramachandran, J. & Hites, R. D. The infrared spectra of 4-substituted-thiosemicarbazides and diazotization products. *Can. J. Chem.* **36**, 801–809 (1958).
58. Khristoforova, Y., Bratchenko, L. & Bratchenko, I. Combination of Raman spectroscopy and chemometrics: A review of recent studies. *Spectrochim. Acta A*. arXiv:2210.10051 (2022).
59. Konings, R. *et al.* The infrared spectrum and molecular structure of gaseous UF4. *J. Mol. Struct.* **378**, 121–131 (1996).
60. Bourassa, P., Hasni, I. & Tajmir-Riahi, H. Folic acid complexes with human and bovine serum albumins. *Food Chem.* **129**, 1148–1155 (2011).
61. Thomas, D. B., McGoverin, C. M., McGraw, K. J., James, H. F. & Madden, O. Vibrational spectroscopic analyses of unique yellow feather pigments (spheniscins) in penguins. *J. R. Soc. Interface* **10**, 20121065 (2013).
62. He, Y., Wang, X., Jin, P., Zhao, B. & Fan, X. Complexation of anthracene with folic acid studied by FTIR and UV spectroscopies. *Spectrochim. Acta A* **72**, 876–879 (2009).
63. Vora, A., Riga, A., Dollimore, D. & Alexander, K. S. Thermal stability of folic acid. *Thermochim. Acta* **392**, 209–220 (2002).
64. Liu, L., Yang, D. & Li, P. PH-related and site-specific excited-state proton transfer from pterin to acetate. *J. Phys. Chem. B* **118**, 11707–11714 (2014).
65. Zhang, J. *et al.* The folic acid/β-casein complex: Characteristics and physicochemical implications. *Food Res. Int.* **57**, 162–167 (2014).
66. Raghavan, R. & Dryhurst, G. Redox chemistry of reduced pterin species. *J. Electroanal. Chem. Interf. Electrochem.* **129**, 189–212 (1981).
67. Sari, W., Qudus, H. I. & Hadi, S. The chemical reactivity study of organotin (IV) 4-aminobenzoates using cyclic voltammetry and antioxidant activity test by the DPPH method. *Rev. Chim.* **71**, 28–37 (2020).
68. Le Gall, A.-C. & van den Berg, C. M. Determination of folic acid in sea water using adsorptive cathodic stripping voltammetry. *Anal. Chim. Acta* **282**, 459–470 (1993).
69. Wang, M., Jiao, Y., Cheng, C., Hua, J. & Yang, Y. Nitrogen-doped carbon quantum dots as a fluorescence probe combined with magnetic solid-phase extraction purification for analysis of folic acid in human serum. *Anal. Bioanal. Chem.* **409**, 7063–7075 (2017).
70. Thomas, A. H. *et al.* Fluorescence of pterin, 6-formylpterin, 6-carboxypterin and folic acid in aqueous solution: pH effects. *Photochem. Photobiol. Sci.* **1**, 421–426 (2002).

71. Liu, S. *et al.* Rational design of metallophosphors with tunable aggregation-induced phosphorescent emission and their promising applications in time-resolved luminescence assay and targeted luminescence imaging of cancer cells. *J. Mater. Chem.* **22**, 22167–22173 (2012).
72. Uyeda, K. & Rabinowitz, J. C. Fluorescence properties of tetrahydrofolate and related compounds. *Anal. Biochem.* **6**, 100–108 (1963).
73. Noiriel, A., Naponelli, V., Gregory, J. F. III. & Hanson, A. D. Pterin and folate salvage. Plants and *Escherichia coli* lack capacity to reduce oxidized pterins. *Plant Physiol.* **143**, 1101–1109 (2007).
74. Jung-Klawitter, S. & Kuseyri Hübschmann, O. Analysis of catecholamines and pterins in inborn errors of monoamine neurotransmitter metabolism: From past to future. *Cells* **8**, 867 (2019).
75. Davis, M. D., Kaufman, S. & Milstien, S. The auto-oxidation of tetrahydrobiopterin. *Eur. J. Biochem.* **173**, 345–351 (1988).
76. Kaur, T. & Ganguly, N. K. Modulation of gut physiology through enteric toxins. *Mol. Cell. Biochem.* **253**, 15–19 (2003).
77. Tiso, M. & Schechter, A. N. Nitrate reduction to nitrite, nitric oxide and ammonia by gut bacteria under physiological conditions. *PLoS ONE* **10**, e0119712 (2015).
78. Vasilieva, S., Petrishcheva, M., Gusarova, E. & Osipov, A. Vitamin para-aminobenzoic acid (PABA) controls generation of nitric oxide (NO) in vitro and its biological functions in the bacterial cells. *Adv. Tech. Biol. Med.* **4**, 195–195 (2016).
79. Shih, M. K. & Hu, M. L. UVA-potentiated damage to calf thymus DNA by Fenton reaction system and protection by para-aminobenzoic acid. *Photochem. Photobiol.* **63**, 286–291 (1996).
80. Rusche, K. M. & Marletta, M. A. Reconstitution of pterin-free inducible nitric-oxide synthase. *J. Biol. Chem.* **276**, 421–427 (2001).
81. Vassie, J. A., Whitelock, J. M. & Lord, M. S. Targeted delivery and redox activity of folic acid-functionalized nanoceria in tumor cells. *Mol. Pharm.* **15**, 994–1004 (2018).
82. Xiu, Y. & Field, M. S. The roles of mitochondrial folate metabolism in supporting mitochondrial DNA synthesis, oxidative phosphorylation, and cellular function. *Curr. Dev. Nutr.* **4**, 153 (2020).
83. Shen, R.-S. & Zhang, Y. Reduced pterins as scavengers for reactive oxygen species. in *Chemistry and Biology of Pteridines and Folates*, 351–354 (1993).
84. Schallreuter, K. U. & Wood, J. M. Thioredoxin reductase: Its role in epidermal redox status. *J. Photochem. Photobiol. B* **64**, 179–184 (2001).
85. Lam, Y.-F. & Kotowycz, G. Self association of folic acid in aqueous solution by proton magnetic resonance. *Can. J. Chem.* **50**, 2357–2363 (1972).
86. Moorthy, P. & Hayon, E. One-electron redox reactions of water-soluble vitamins. II. Pterin and folic acid. *J. Org. Chem.* **41**, 1607–1613 (1976).
87. Khaled, M. A. & Krumdieck, C. L. Association of folate molecules as determined by proton NMR: Implications on enzyme binding. *Biochem. Biophys. Res. Commun.* **130**, 1273–1280 (1985).
88. TzankoI Doukov, H. H., Drennan, C. L. & Ragsdale, S. W. Structural and kinetic evidence for an extended hydrogen-bonding network in catalysis of methyl group transfer. *J. Biol. Chem* **2007**, 6609–6618 (2006).
89. Ragsdale, S. W. Catalysis of methyl group transfers involving tetrahydrofolate and B12. *Vitamins Horm.* **79**, 293–324 (2008).

Author contributions

Conceptualization, S.N., M.R.; data curation, S.N., G.P., B.D. and M.R.; formal analysis, S.N., G.P., and M.R.; funding acquisition, M.R.; investigation, S.N., M.R.; methodology, S.N., B.D. and M.R; supervision, B.D. and M.R.; visualization, S.N., M.R.; writing-original draft, S.N., M.R.; writing-review and editing, S.N., G.P., B.D. and M.R.

Funding

This research has been supported in part by the startup Grant-120000003048 from Penn State-COM (to M. Rawat).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.D.M. or M.R.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Heliyon 9 (2023) e22540

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Numerical simulation of free and submerged hydraulic jump over trapezoidal and triangular macroroughness

Harshit Kumar Jayant *, Bharat Jhamnani

Delhi Technological University, Department of Civil Engineering, Delhi 110042, India

ARTICLE INFO

Keywords:
Computational fluid dynamics
Hydraulic jump
Macroroughness

ABSTRACT

Hydraulic jump is a phenomenon that occurs in open channel flow and has wide utility in hydraulic engineering. The present study considered macroroughness features in the bed to improve the characteristics of free and submerged hydraulic jumps. For that, the most often utilized triangular macroroughness elements are contrasted with a distinctive trapezoidal macroroughness. This study implied a numerical model based on computational fluid dynamics (CFD) to investigate the hydraulic jump characteristics over the macroroughness and represented the important features of hydraulic jump such as sequent depth ratio, tail-water depth ratio, longitudinal velocity profile, flow pattern of velocity vector, and contours for turbulent kinetic energy over smooth and macrorough beds. It is seen that for a specific Froude number, triangular macroroughness outperforms trapezoidal macroroughness. The average value of the length of the roller for trapezoidal and triangular macroroughness is found to be 23 % and 46.06 % shorter than the smooth bed in the case of a free jump, whereas 17.85 % and 36.33 % shorter than the smooth bed in the case of a submerged jump, respectively. To check the accuracy of the numerical model, the basic parameters of a hydraulic jump, such as sequent depth and tail-water depth ratio, are compared with the experimental findings of the present work, and the mean error between the CFD and experimental models is reported to be within 6 %, which confirms the reliability of the CFD model and can be applied to real hydraulic engineering.

1. Introduction

A hydraulic jump is a phenomenon that occurs in open-channel flow, typically involving the flow of water in rivers, channels, or spillways. This phenomenon is characterized by a rapid decrease in flow velocity and an abrupt increase in water depth over a relatively short distance. Classical hydraulic jumps are the most common type and are associated with abrupt energy dissipation and the formation of turbulent rollers. Classical hydraulic jumps were studied in depth by Refs. [1–6]. Further hydraulic jumps are classified into two categories free and submerged hydraulic jumps. A submerged hydraulic jump takes place below the water's surface, while a free hydraulic jump happens above it [7]. Fig. 1 depicts the definition sketch of hydraulic jump and their important parameters where, a sluice gate is provided at the boundary of the upstream and downstream sections in order to illustrate the free and submerged hydraulic jump phenomenon. The sluice gate functions as a hydraulic structure to release extra water from the reservoir, which transforms potential energy into kinetic energy, upstream portion contains reservoir conditions. To reduce the risk of damaging to the downstream bed and hydraulic system, this extra energy must be treated close to it. The hydraulic jump is a common occurrence that

* Corresponding author.
E-mail address: harshit_2k19phdce04@dtu.ac.in (H.K. Jayant).

https://doi.org/10.1016/j.heliyon.2023.e22540
Received 31 August 2023; Received in revised form 14 November 2023; Accepted 14 November 2023
Available online 21 November 2023
2405-8440/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Index of notations and abbreviations

CFD	computation fluid dynamics
RANS	Reynolds-averaged Navier-Stokes
RNG	re-normalization group
SST	shear stress transport
VOF	volume of fluid
TKE	turbulent kinetic energy
k	turbulent kinetic energy
ε	dissipation rate of turbulent kinetic energy
ω	specific dissipation rate of turbulent kinetic energy
μ	dynamic viscosity of the fluid
υ	kinematic viscosity
g	gravitational acceleration
u	time-averaged longitudinal velocity at any section
U	the maximum longitudinal velocity in the horizontal direction at any section
Q	discharge
ρ	fluid density
y_0	opening of sluice gate
y_1	initial depth of water in case of hydraulic jump
y_2	tail-water depth of water in case of a free hydraulic jump
y_3	submerged depth of water in case of a submerged hydraulic jump
y_4	tail-water depth of water in case of a submerged hydraulic jump
y_2/y_1	sequent depth ration in case of free hydraulic jump
y_4/y_1	tail-water depth ratio in case of submerged hydraulic jump
L_r/y_1	relative roller length
E_{y_2/y_1}	relative error in prediction of sequent depth ratio
H	height of macroroughness elements
B	the base width of macroroughness elements
Fr_1	Froude number of the inlet section of the hydraulic jump
R_e	Reynolds number
C_r	Courant number
L_r	length of roller
L_{j1}	length of hydraulic jump in case of free hydraulic jump
L_{j2}	length of hydraulic jump in case of submerged hydraulic jump
E_1	specific energy beginning of the jump
E_2	specific energy after the jump
R^2	coefficient of correlation
u_i	velocity vector in three-dimensional space, where i represents the cartesian coordinates (x, y, z)
x_i	position vector in three-dimensional space, where i represents the cartesian coordinates (x, y, z)
t	time variable
p	pressure of the fluid
ρ	density of the fluid
x*	mesh element size
y^+	dimensionless wall distance
t_{ji}	viscous stress tensor
x_i	co-ordinate in the i-axis of a cartesian reference system
μ_t	turbulent eddy dynamic viscosity
G_k	production term for turbulent kinetic energy
G_b	buoyancy effect term
Y_M	dilatation oscillation effect term
S_k	source terms for turbulent kinetic energy
S_ε	source terms for dissipation rate of turbulent kinetic energy

has been exploited in the past to control this surplus.

The characteristics of the jump may be sequent depth, submerged depth, length of the jump, and roller length of jump depending upon the jump type. Equation (1) depicts the sequent depth ratio (y_2/y_1) which is a well-known Belanger equation [8], it is a fundamental parameter of hydraulic engineering that helps in characterizing and quantifying the changes in water depth that occur in hydraulic jumps. The tail-water depth ratio (y_4/y_1) also provides information about the turbulence within the hydraulic jump which

was discussed by Ref. [7]. Other crucial jump metrics include length of jump (L_j) and roller length of jump (L_r), which measure the distance between the toe and the surface stationary region [9]. Length of jump (L_j), the distance between the toe and the location where the water surface almost levels off and the mean surface elevation reaches its maximum, is used to define the length of the jump, however, this definition is ambiguous [1]. According to Ref. [10], length of roller (L_j) can be defined as the distance over which the mean free surface level increased monotonically. Further, research of [11] shows the clear identification of L_r where it was the distance between the toe of jump and the location where longitudinal velocity at a section start gaining the only positive values.

$$\frac{y_2}{y_1}=\frac{1}{2}\left(\sqrt{1+8Fr_1^2}-1\right) \tag{1}$$

In equation (1), Fr_1 is the dimensionless Froude number which has important role in open channel flows and can be defined as the ratio of inertial force to gravity force $\left(Fr_1=\frac{u_1}{\sqrt{gy_1}}\right)$. In order to minimizing the excess kinetic energy downstream of the hydraulic structure, several articles were published in the past to improve the energy dissipation using hydraulic jump over smooth and rough beds. Few of them are discussed here.

[12]investigated hydraulic jump characteristics experimentally over two types of bed roughness first strip corrugation and second using gravel bed in a rectangular channel, they confirmed that the channel roughness decreases the sequent depth, and length of the jump, also the characteristics of the jump are the function of the initial Froud number and channel roughness.

[13]studied hydraulic jump experimentally over a sinusoidal corrugated bed and found that the corrugated bed is effective in maximizing the energy dissipation and reducing the sequent depth, and length of jump by enhancing the bed shear stress at the bottom of the bed. The integrated bed shear stress in the case of a corrugated bed was approximately ten times that smooth bed.

[14]experimentally investigated the hydraulic jump's characteristics including velocity, shear stress distribution, and free surface location over six different types of corrugated bed, also examined the effect of roughness height and length of roughness on the characteristics of the hydraulic jump, they found that the energy dissipation over the corrugated bed is much higher than that of smooth bed, the bed shear stress over the corrugated bed is almost ten times higher than the smooth bed.

[15]investigated the effects of corrugations and prismatic roughness elements in the case of a forced hydraulic jump in a rectangular channel and stated that the length of the jump is reduced by about 35 %, 40 %, and up to 55 % by corrugations, strip

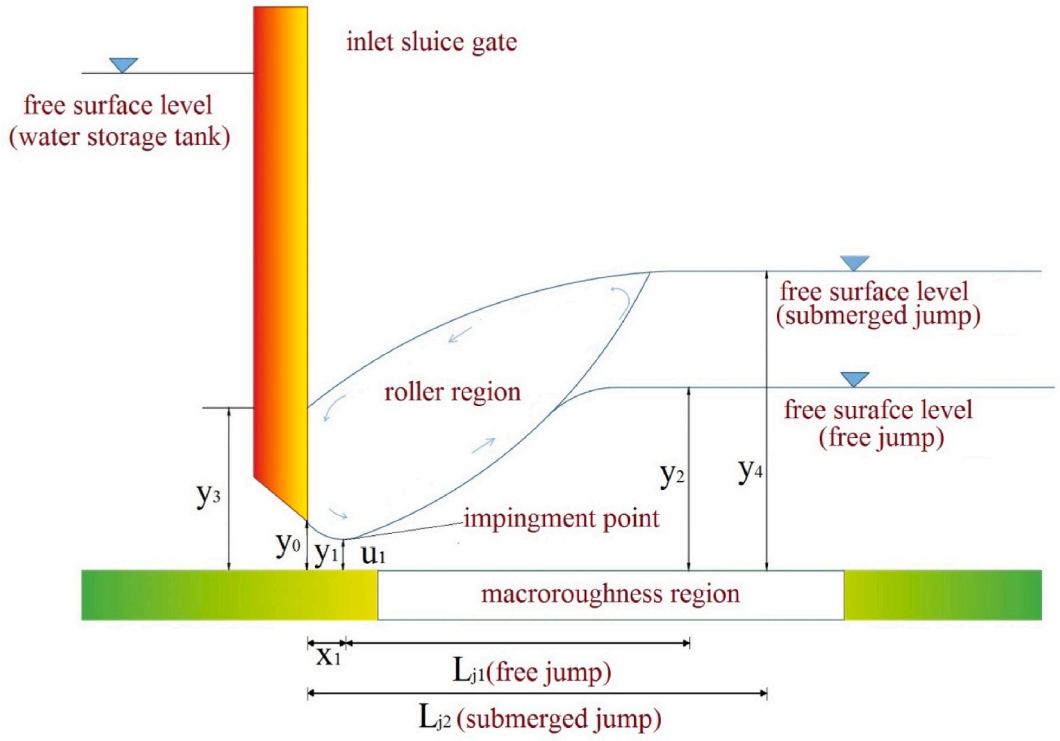


Fig. 1. Definition sketch of free and submerged hydraulic jump showing relevant parameters associated with the phenomenon. x_1 is longitudinal jump toe position; u_1 is the depth-averaged velocity at that section.

roughness, and staggered roughness respectively whereas tail-water depth reduces to 20 %, up to 13 %, up to 15 % by corrugations, strip roughness, and staggered roughness respectively.

[16]investigated the free surface location and energy dissipation in the case of a hydraulic jump using a machine learning technique and found that the parameter that affects free surface location and energy dissipation is the function of the initial Froud number.

[17]effect of triangular corrugation on the characteristics of submerged hydraulic jump was studied and found that the sequent depth and jump length were reduced by average values 15.14 % and 21.03 % than smooth bed whereas improvement in jump efficiency was achieved up to 50 % at optimum spaced corrugation.

[18]studied hydraulic jump experimentally over discontinuous roughness elements of lozenge shape in a rectangular channel with adverse slope, they found that the hydraulic jump characteristics are unstable on an averaged slope but the staggered roughness of bed is well able to stabilize the jump.

[19]experimentally investigated the air-water flow pattern and basic distributions of air-water flow properties over uniform macroroughness in the case of hydraulic jump. They found that an increase in bubble count rate and void fraction in the region close to the jump to over macrorough bed as compared to smooth bed.

[20]experimentally studied the hydraulic jump over the rough bed in an abrupt asymmetric channel and found that the sequent depth, length of jump, and roller considerably decreased over the rough bed as compared to the smooth bed of the prismatic channel. Also, the energy dissipation improved over a rough bed as compared to classical jump.

[21]numerically investigated the hydraulic jump over a rough bed. They found that the presence of roughness element near the bed increases the bed shear, and reduces the velocity. Also, the tail-water depth and length of jump reduce considerably over a rough bed as compared to a smooth bed.

[22]studied the hydraulic jump to analyse the effect of bed roughness and abrupt negative drops on hydraulic jump characteristics, they confirmed that the bed roughness causes a reduction in sequent depth of 16.6 % and a reduction in relative length of jump of 20.7 %. Also, they found that by increasing the step height, the sequent depth ratio and relative length of jump value increased by 6.5 % and 7 % respectively.

[23]tested two different heights of roughness to investigate the hydraulic jump properties, roughness was discontinuously arranged over the horizontal and adverse beds, they found that the bed shear coefficient in case of adverse bed condition with roughness is 12.4 times higher than smooth bed.

[24]studied two-phased flow properties in hydraulic jumps on a pebbled rough bed, they found that the turbulence intensity in the roller region was higher over a rough bed as compared to a smooth bed and higher energy dissipation occurred on a rough babbled bed. In continuation, a triple decomposition analysis was performed and confirmed that the slow fluctuation associated with flow instability correlated with the turbulent kinetic energy of the flow.

Based on the referred literatures it seems that rough beds are more effective than smooth beds for the energy dissipation of excess kinetic energy of flow downstream of the hydraulic structure. It is found that the rough beds in the form of strip corrugation have been more utilized in past studies, but the exploration of strip corrugation with some new shapes is still unknown. Further, it is also noted that the experimental studies are adequate on hydraulic jump analysis but numerical studies are less in numbers which needed more studies. For that this work preferred to discussed the numerical simulation technique for hydraulic jump analysis.

The objective of this study is to explore the new shape of corrugation for the analysis of free and submerged hydraulic jumps using numerical modelling. The present study used a novel distinctive trapezoidal shape in contrast to the most commonly used triangular shape of corrugation to plot the variation of flow profile, the flow pattern in cavity region, longitudinal velocity profile, and kinetic energy distribution numerically. The numerical modelling of hydraulic jump of the present study may provide better insights to the real-life hydraulic designer for designing the energy dissipator structure and they may take the benefit of numerical modelling software to analyse a complex fluid phenomenon.

2. Methodology

This work is original research work which was numerical simulation investigated and validated d using the experimental findings of the present research as well as by comparison with earlier research. The initial step involved performing experiments in the hydraulics lab and varying the same parameters in the numerical simulation to obtain the results. Due to capability of predicting the real world phenomenon with acceptable percentage of error, numerical simulation techniques were utilized in the past by Refs. [11,21,25] to performed the fluid flow related studies. The only limitation of this study may be the errors in showing the findings of experimental and numerical model which may be due to the two reasons, first roughness constant for macrorough beds in numerical model was not same as of physical macrorough materials and second the influence of mesh, turbulence sensitivity however, these parameters were checked

to achieve the required accuracy.

2.1. Experimental background

In the present work, tests were carried out in an open channel rectangular flume at Delhi Technological University's hydraulics lab that was 10.0 m long, 0.5 m deep, and 0.5 m wide. Fig. 2 shows the physical macroroughness prepared with iron material that were positioned across the whole width of the flume were used to create bed macroroughness. For every round of the experiment, the bed macroroughness length remains constant. Single bed roughness arrangement was taken into consideration in order to explore the various jump parameters. On a mild bed slope with varying Fr_1 numbers ranging from 3.5 to 7.0, forty-two runs were done, with seven runs performed for each type of bed and jump conditions. Discharge and velocity were measured with the help of digital discharge meter and digital pitot tube (L-type pitot tube of 6 mm diameter and digital manometer with an accuracy of ±0.5% full scale deflection) respectively. Other relevant measurements such as water depth were done with the help of rail mounted pointer gauges with an accuracy of 0.2 mm. Table 1 depicts the parametric variations considered in the present study.

2.2. Numerical modelling

The practice of utilizing computer algorithms to simulate the behaviour of a system or process in the real world usually takes the form of mathematical models known as numerical modelling. When air and water come together, the flow zone creates turbulence, which leads to the hydraulic jump phenomena. A potent method for examining the properties of hydraulic jump is computational fluid dynamics (CFD). The advancement of both numerical techniques and digital machines with growing processing capacity has enabled the widespread application of CFD approaches to various fields of fluid dynamics in recent decades [26]. Numerical simulation has been employed over the past twenty years in a variety of study fields to investigate the new demanding parameters. Because CFD produces more trustworthy results in the field of fluid flow, several studies have used it to simplify the solution of challenging issues. The flow must be modelled using the suitable numerical techniques in order to carry out a CFD examination of hydraulic jump characteristics. Reynolds-Averaged Navier-Stokes (RANS) equation is based on time-averaged or Reynolds-averaging of Navier Stokes equation, where the instantaneous values of velocity and pressure is assumed to be the sum of a time-averaged value and a fluctuating component [25]. In the past [11,27,28] employed CFD to track the free surface of the flow using a complex method called the volume of fluid (VOF) model. Here, numerical simulations are done using the well-known licensed program Ansys Fluent 2022 R1. For this study, both two-phase gases and liquids are taken into account. Here, air is one type of gas and water is one type of liquid.

Equation (2) and Equation (3) represents the continuity and Navier-Stokes equation, respectively which is a basic equation for CFD. Volume of fluid equation is written by Equation (4).

(a) Continuity:

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{2}$$

(b) Navier-Stokes equations:

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial t_{ji}}{\partial x_j} \tag{3}$$

(c) Volume of fluid tracking equations:

$$< listaend > \frac{\partial F}{\partial t} + \frac{1}{V_f}\left[\frac{\partial (FA_x u_x)}{\partial x} + \frac{\partial \left(FA_y u_y\right)}{\partial y} + \frac{\partial (FA_z u_z)}{\partial z}\right] = 0 \tag{4}$$

The fraction function F in this case indicates whether a cell is filled or empty. It is position-dependent and indicates the percentage of the cell that is occupied by a fluid. F = 0 if a cell is dry. This shows that the fluid's percentage of the cell volume is zero. F = 1 is the

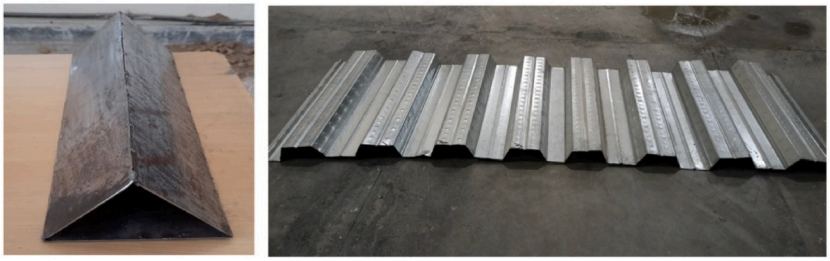


Fig. 2. Physical macroroughness elements.

Table 1
Parametric variations of flow for the present study (numerical and experimental).

Bed Type	Q (l/s)	y_o (cm)	y_1 (cm)	u (m/s)	Fr_1
Smooth	15–25	2.5–3.0	1.1–1.5	1.24–2.23	3.7–7.0
Rough	15–25	2.5–3.0	1.1–1.5	1.24–2.23	3.7–7.0

case if a cell is filled [29,30]. This shows that the fluid occupies a fraction of one, or a fully filled cell, of the cell's volume.

The fraction function F is used to locate the free surface and map out the fluid distribution inside a computational domain. A cell's intermediate fluid concentrations can be identified by giving 'F' various values, which enables the free surface position to be identified. Typically, the fraction function's intermediate value, F = 0.5, is used to determine the free surface. This indicates that at the free surface, fluid occupies half of the cell volume and the remaining half is empty.

2.3. Modelling domain

Fig. 3 depicts the general information regarding the geometry of numerical simulation, here geometry was created considering the same dimension as the experimental flume except the length of flume, to minimize the computation effort and time the length of flume was reduced to 5 m. The mesh analysis that was done in a meshing setup with the various faces given the appropriate names. Inflation layers were also included at the flume's bed to improve the flow behaviour.

2.4. Turbulence model

The software allows for the use of several turbulent models; nevertheless, the precise and useable model is dependent on the phenomenon under investigation. In the past [31] performed the mesh and turbulence sensitivity analysis using Ansys Fluent software for a nuclear fuel bundle placed within a standard pressure flume, they checked the mesh and turbulence sensitivity for predicting the velocity profile and confirmed that the choice of mesh and turbulence models may alter the results [28]. investigated classical hydraulic jump using numerical simulation and found RNG k- ε, followed by SST k- ε and standard k- ε, is the most accurate model for hydraulic jump analysis [11]. investigated the hydraulic jump over macrorough beds numerically and said that RNG k-ε is similar to the standard k- ε model, even if it contains some refinement [21]. numerically performed hydraulic jump over rough beds and demonstrated how well the RNG k-***ε***model predicts the hydraulic jump over a corrugated bed. To obtain the averaged equations for the two turbulence parameters turbulent kinetic energy (k) and its dissipation rate (ε), the RNG k- ε method uses statistical approaches. This model has the advantage that, as compared to the standard k- ε, it often produces better outcomes when modelling whirling flows [28,29]. For the present study authors tried to test the sensitivity of standard k-***ε***, RNG k- ***ε***, Realizable k-***ε***, SST k-***ω*** turbulence model for study the flow parameters of the free hydraulic jump in the sensitivity analysis section.

The formulation for turbulence kinetic energy (k) and kinetic energy dissipation (***&epsi***) is represented in Equations (5) and (6) respectively. Whereas Equation (7) represent the formulation of (μ_t) for equations (5) and (6).

$$\frac{\partial}{\partial t}(\rho\varepsilon)+\frac{\partial}{\partial x_i}(\rho\varepsilon u_i)=\frac{\partial}{\partial x_j}\left[\left(\mu+\frac{\mu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_i}\right]+G_k+G_b-\rho\varepsilon-Y_M+S_k \tag{5}$$

$$\frac{\partial}{\partial t}(\rho\varepsilon)+\frac{\partial}{\partial x_i}(\rho\varepsilon u_i)=\frac{\partial}{\partial x_j}\left[\left(\mu+\frac{\mu_t}{\sigma_\varepsilon}\right)\frac{\partial \varepsilon}{\partial x_i}\right]+C_{1\varepsilon}\frac{\varepsilon}{k}(G_k+C_{3\varepsilon}G_b)-C_{2\varepsilon}\frac{\varepsilon^2}{k}+S_\varepsilon \tag{6}$$

$$\mu_t=C_\mu\frac{\rho k^2}{\varepsilon} \tag{7}$$

In order to control the behaviour and the accuracy of the turbulence model, certain parameters have precise values allocated to them. In the RNG k-ε model, the precise values for $C_\mu, C_{1_\varepsilon}, C_{2_\varepsilon}, C_{3_\varepsilon}, \sigma_k, \sigma_\varepsilon$ are 0.084, 1.42, 1.68, 1.0, 0.7194, and 0.7194, respectively [29].

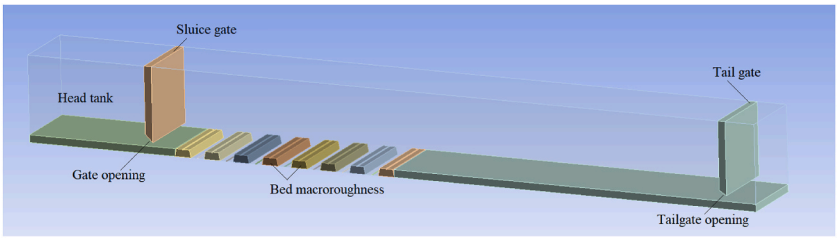


Fig. 3. Model geometry of simulation setup showing relevant parameters associated with the flume.

2.5. Boundary conditions

Depending on the particular hydraulic jump being emulated and the amount of accuracy required, the right boundary conditions should be chosen. In order to ensure that the simulation truly depicts the hydraulic jump that occurs in reality, it is crucial to carefully choose the boundary conditions. Ansys Fluent features a toggle key for open channel flow, which has been used for the simulation's purposes. Pressure inlet has been chosen to cause a jump near the flume, whereas gauge pressure has been chosen for the flume's top surface and outflow. The wall is thought of as a no-slip wall, with a 0.5 wall roughness constant and an assumed zero flow velocity at the wall. The initial settings at the inlet and outflow are made to resemble experimental circumstances.

2.6. Near wall treatment

Wall functions are empirical correlations that use data from the next cell to estimate flow variables along the wall. These functions save processing time by avoiding the resolution of the near-wall region. The wall function resolves the flow along the wall, over-refinement of the mesh could lead to less accurate findings [28]. An important factor in wall functions is the non-dimensional distance between the wall and the first node away from the wall, which is represented by y^+ [29]. Laminar sub-layer, transition layer and turbulence are the three portions that make up the turbulent boundary layer with limits ($y^+ < 5$), ($5 < y^+ < 30$) and ($y^+ > 30$), respectively. Normal wall function, non-equilibrium wall function, improved wall treatment, and other options for near wall treatment are available in Ansys Fluent for the k-εturbulence models. Because it offers reasonably accurate predictions for the majority of high Reynolds-number (Re), wall-bounded flows and reduces computational effort by doing away with the resolution of the near-wall region, the standard wall function was used in the present study.

2.7. Discretization scheme

In the case of numerical simulation, discretization has a greater impact on the parameters that determine the outcome. Because upwind schemes in the RANS model show reduced instability, according to Ref. [28], the upwind schemes were chosen in the present investigation. In numerical simulations, a dimensionless measurement termed as the Courant number (C_r) is used to indicate the relationship between the time step and the spacing between the computing grid points. It is used to gauge the simulation's stability because a high Courant number might cause numerical instability. The highest C_r employed in the current simulation of the submerged hydraulic jump is 0.40, which is within the safe range for such a form of simulation. The pressure-velocity coupling method as SIMPLE (Semi-Implicit Method for Pressure-Linked Equations), a popular algorithm for resolving fluid flow issues, is employed. Using the first and second-order upwind techniques, respectively, the density and momentum equations are discretized.

2.8. Stability & convergence criteria

The residuals are set to 10^{-6} in order to examine the convergence criteria in this study. Given that the hydraulic jump is abrupt in nature, maintaining the stability of the numerical model for the hydraulic jump is somewhat difficult. Here, the hit-and-trial method is used to stabilize the jump, by measuring the discharge equilibrium curve for the inflow and outflow. After 11 s of time steps, the jump's stable conditions are apparent.

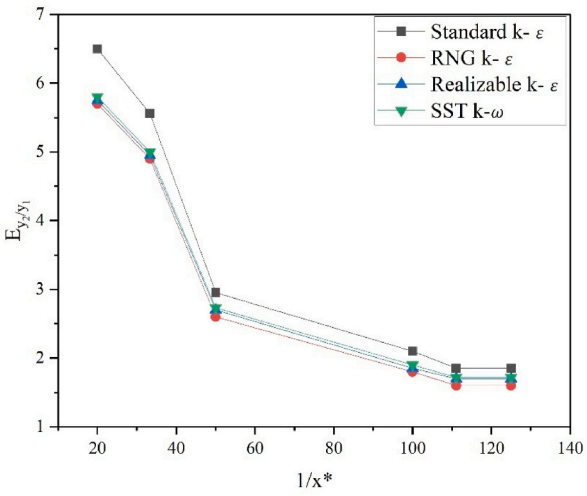


Fig. 4. Mesh and turbulence model sensitivity analysis on y_2/y_1 prediction.

2.9. Sensitivity analysis

Grid sensitivity analysis is a statistical method for assessing how sensitive a model's output is to changes in the parameters or other inputs. In order to identify which, input elements, have the greatest influence on the output, the process entails adjusting one or more input factors throughout a predetermined range of values, running the model for each set of input parameters, and analysing the resulting output. Creating a grid of input parameter values and running the model for each combination of values in the grid are typical steps in a grid sensitivity analysis. The sensitivity of the output to each input factor is then determined by statistically analysing the output from each model run.

For the numerical simulation, sensitivity analysis is required to obtain the best combination of mesh and turbulence to achieve the accuracy in results at an optimal computational cost. This study had done the sensitivity analysis on grid independency and on turbulence models. A brief description is provided. This study considered the sequent depth ratio (y_2/y_1) as a primary parameter to perform the grid and turbulence model sensitivity analysis as procedure adopted by Refs. [27,28]. Here six different grid size of course to fine were considered for the sensitivity analysis. Fig. 4 shows the variation of relative error in predicting the sequent depth ratio for different turbulence model at different mesh element size. It was found that all the different models show very small variation in predicting sequent depth ratio. RNG k- ε model yielded superior results with minimum error followed by realizable k- ε, SST k-ω and standard k- ε model. Fig. 4 confirmed that the relative error in predicting the sequent depth ratio reduces with reducing the mesh element size and a constant value comes after the 9 mm mesh element size which signifies that the further reduction in size is not required to avoid the computational cost. The maximum error at 9 mm size with RNG k- ε was reported below 2 %.

Further Fig. 5 shows the grid sensitivity analysis on longitudinal velocity of free hydraulic jump on smooth bed using RNG k- ε turbulence model and it is performed on the basis of procedure provided by Ref. [32]. Fig. 5 shows the result of grid convergence on the longitudinal velocity which confirms that the fluctuation in longitudinal velocity profile is minimum for GC-5 & GC-6, further reduction in grid size is not required to save the cost of computation.

Hence, the present study considered the 9 mm mesh size and RNG k- ε turbulence model to discussed the hydraulic jump over smooth and macrorough beds numerically at an affordable computational cost.

3. Results and discussion

3.1. Validation of numerical model

Numerical simulation validation is the process of confirming that computer simulation findings accurately reflect the processes being examined in reality. To determine the simulation's accuracy and dependability, experimental data are compared to the results of the simulation. Quantifying the disparities between simulation results and experimental data and locating potential areas for simulation improvement can help to determine the simulation's accuracy and reliability [33]. studied the scale effect while analysing the hydraulic jump in two different channel dimensions, they found that some scale effect the void fraction and bubble count rate, to overcome such issues the present numerical study considered the dimensions of flume and roughness elements identical to experimental setup except the length of flume, length of the flume was reduced to avoid the computational cost.

In Fig. 6, which compares numerical and experimental findings on fundamental jump parameters including the sequent depth ratio (y_2/y_1) and tail-water depth ratio (y_4/y_1), there is good agreement between the two models, with a standard error of less than 6 %. This

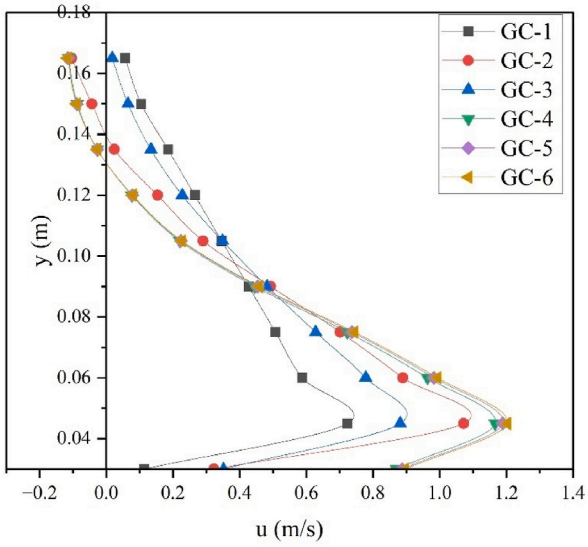
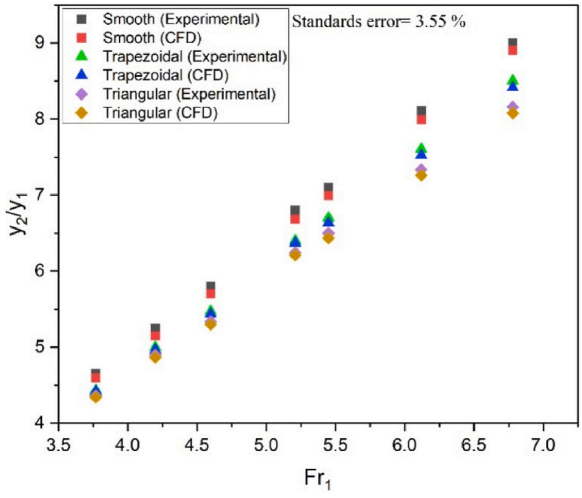
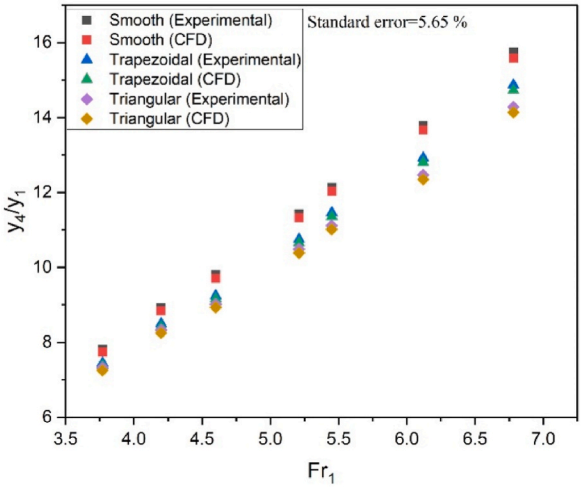


Fig. 5. Sensitivity analysis of mesh size for horizontal component of velocity profile.



(a)



(b)

Fig. 6. Comparison of a numerical model with experimental model for basic parameters of hydraulic jump (a) y_2/y_1(b) y_4/y_1.

degree of agreement demonstrates that CFD modelling may faithfully simulate complex fluid dynamics issues like hydraulic jumps. Equations (8) and (9) are derived for sequent depth ratio in case of smooth and macrorough beds, respectively. Equations (10) and (11) represents for tail-water depth ratio in case of smooth and macrorough beds, respectively.

$$y_2 / y_1 = 1.4569\ F_{r1}\text{-}0.8975R^2 = 0.99\ \text{(smooth bed)} \tag{8}$$

$$y_2 / y_1 = 1.3042\ F_{r1}\text{-}0.547R^2 = 0.99\ \text{(macrorough bed)} \tag{9}$$

$$y_4 / y_1 = 2.6059\ F_{r1}\text{-}2.1254R^2 = 0.99\ \text{(smooth bed)} \tag{10}$$

$$y_4 / y_1 = 2.4229\ Fr_1\text{-}1.8121R^2 = 0.99\ \text{(macrorough bed)} \tag{11}$$

3.2. Velocity variation

According to Ref. [34], there is an inverse relationship between velocity magnitude and the upstream Froude number (Fr_1). The

general velocity profile in the hydraulic jump is shown in Fig. 7 [1]. Maximum velocity is observed at a vertical distance of y* from the bed level, and at a vertical distance of b', velocity drops to approximately half of the maximum velocity. This is a typical velocity fluctuation in a fully developed roller zone; just above point b', reverse flow happens due to velocity stagnation.

The standard pictorial representation of the numerically obtained free & submerged hydraulic jump's velocity profile on smooth and macrorough beds are shown in Fig. 8 & Fig. 9 respectively.

By examining the results of Figs. 8 and 9, it can be deduced that the mean flow velocity falls at the topmost region of the gate opening because this is where it is highest. Following the development of the jump, the jump velocity in the roller zone reverses with a stagnation point in a backward direction. Such a description is valid for both macrorough and smooth beds. Velocity distribution confirms the past velocity distribution of [1,11,13].

Fig. 10 represents the dimensionless velocity distribution of submerged hydraulic jump on smooth and macroroughness arrangements for $Fr_1 = 6.12$. Here, the vertical axis represents the ratio of y*/b', where y* is the distance from the bed and b' is the length scale at which velocity is half of maximum velocity. The horizontal axis represents the ratio of u'/U, where u is the time-averaged longitudinal velocity at any section and U is the maximum longitudinal velocity at that section.

The dimensionless velocity distribution graph is typically plotted for the point where the velocity is only half the maximum velocity, which is at a distance of b' from the jump toe. This is due to the downstream region's generally flat velocity distribution, which does not exhibit much variance. Additional computations are made to normalize the velocity values depending on the mean velocity and the height of the jump in order to provide the dimensionless velocity distribution. For validation purposes, previous studies [1,11,35] are used as references and presented in comparison with the present findings in Fig. 10. It is discovered that the trend of dimensionless velocity profiles is more or less similar to the previous research and that the maximum horizontal forward velocity occurred at a shallower flow depth than in the [1,35] studies.

3.3. Velocity vector

The velocity vector variation for the free and submerged hydraulic jumps on macrorough beds is shown in Fig. 11 & Fig. 12, respectively. The results shows that the flow over macrorough beds is in circular motion in the empty space zone near the bed, which may be the cause of the considerable reduction in excess kinetic energy, while there is no circular motion near the bed in case of smooth surface is seen. Additionally, the configuration of the macrorough bed affects the direction of velocity vector circulation. Additionally, it is interesting to note that the velocity in the empty space region between two successive macroroughness element is quite lesser than the mean velocity, which suggests that the flow in this region behaves differently from the rest of the flow. Lastly, it seems that the presence of empty space between two successive elements near the bed may affect the behaviour of the flow, especially for higher initial Froude numbers.

3.4. Relative roller length for hydraulic jump

This section deals with L_r/y_1, which is also an important parameter to understand the reduction in turbulence of flowing water, previously such parameters was deeply investigated by Refs. [13,17]. Figs. 8 and 9 represents the approximate values of roller length parameter for free and submerged jumps over various bed conditions accordingly addition calculation was made to present the dimensionless parameter relative roller length of jump. From Fig. 13, it can state that the relative roller length decreases with the provision of macroroughness to the bed and the more reduction in such parameter was observed in case of free hydraulic jump. Several equations for the relative roller length were derived and presented in Equations (12)–(15). Equations (12) and (13) are depicts for smooth and macrorough beds, respectively in case of free hydraulic jump. Equations (14) and (15) depicts for smooth and macrorough beds, respectively in case of submerged hydraulic jump.

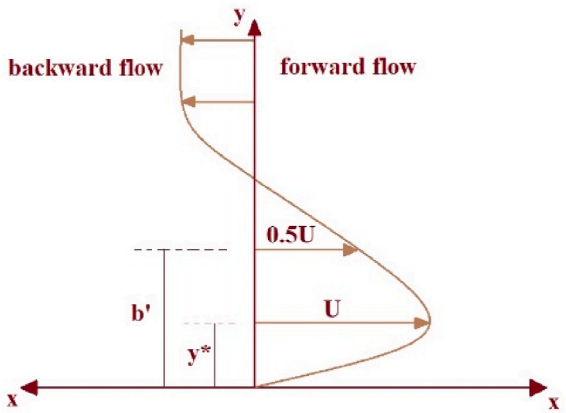


Fig. 7. Typical longitudinal velocity profile for a hydraulic jump.

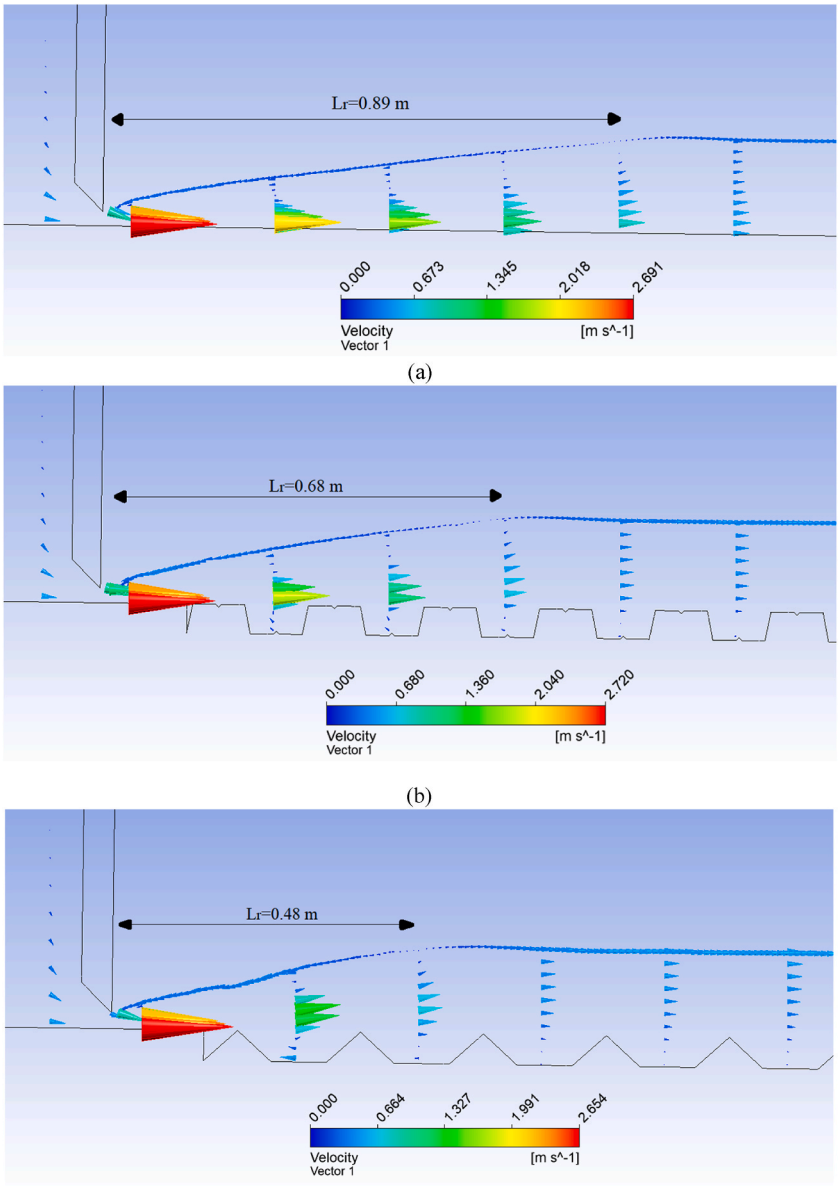


Fig. 8. Pictorial representation of velocity profile in free hydraulic jump(a) smooth bed (b) trapezoidal macroroughness (c) triangular macroroughness ($Fr_1 = 5.45$).

$$L_r / y_1 = 4.0858F_{r1} + 0.6186, R^2 = 0.99 (\text{smooth bed, free jump}) \quad (12)$$

$$L_r / y_1 = 2.6865F_{r1} + 4.9481, R^2 = 0.96 (\text{macrorough bed, free jump}) \quad (13)$$

$$L_r / y_1 = 5.4912\ F_{r1} - 3.778, R^2 = 0.99\ (\text{smooth bed, submerged jump}) \quad (14)$$

$$L_r / y_1 = 4.7148\ F_{r1} - 2.2652, R^2 = 0.98\ (\text{macrorough bed, submerged jump}) \quad (15)$$

3.5. Turbulent kinetic energy

To investigate the energy loss in a two-channel section, the turbulent kinetic energy (TKE) is a crucial metric. By averaging the velocities in all possible directions, the TKE may be calculated. Fig. 14 compares the TKE over flume sections with various macroroughness setups to the smooth bed. Fig. 14 shows that the TKE drops across the section, also shows that the TKE is greatest when the bed is smooth and reduces when the bed is assigned macroroughness. Further it is noted that the free jump dissipates more kinetic

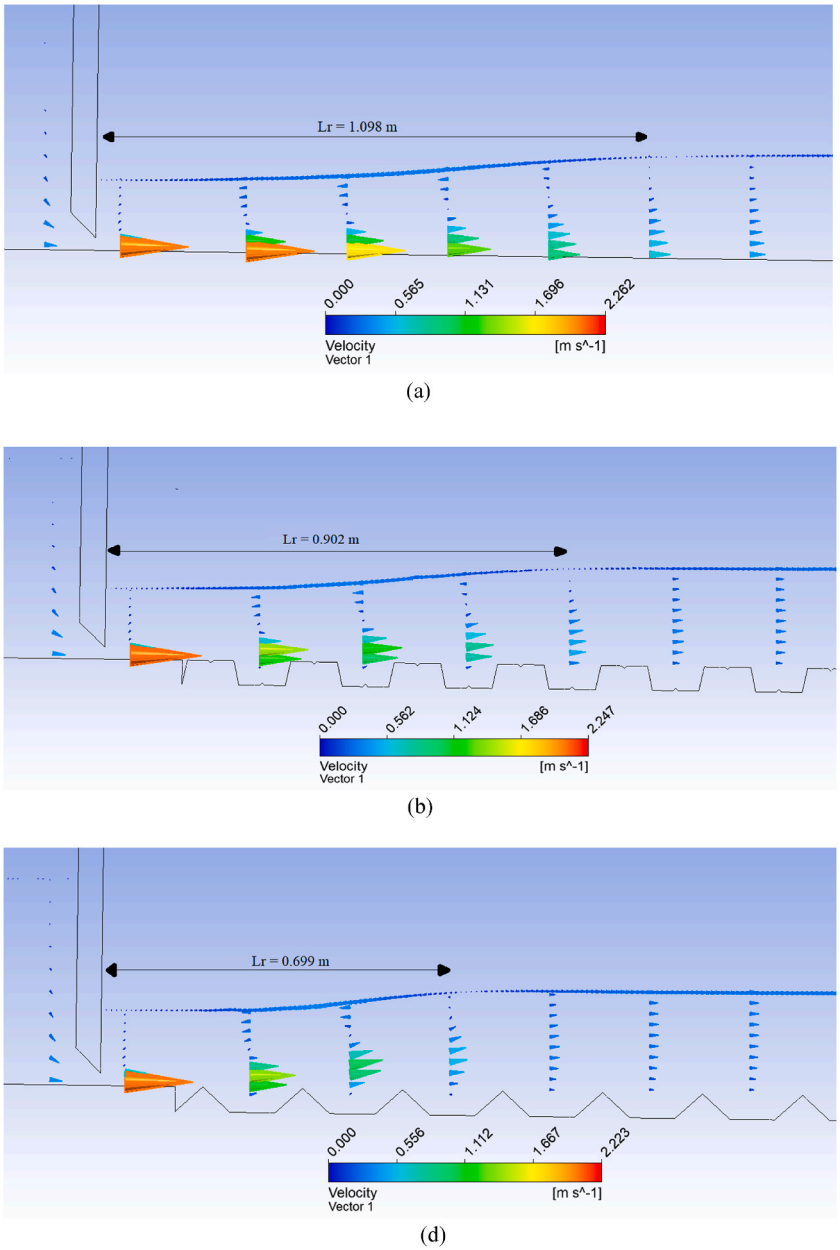


Fig. 9. Pictorial representation of velocity profile in submerged hydraulic jump (a) smooth bed (b) trapezoidal macroroughness (c) triangular macroroughness ($Fr_1 = 6.12$).

energy as compare to submerged jump which confirm the findings of [11].

4. Conclusion

This research aimed to enhance the understanding of free and submerged hydraulic jumps by exploring various flow characteristics and their dependence on different shapes of macroroughness with unique arrangements. This work used a novel distinctive trapezoidal shape in contrast to the widely used triangular shape of macroroughness to numerically analyse the flow pattern near the bed which was completely ignored in past studies. The several findings of the present work are presented here.

1. It is found that the CFD is fairly good at predicting the free &submerged hydraulic jump on smooth and macrorough beds.
2. The turbulence model, RNG k-$\boldsymbol{\epsilon}$ showed better accuracy in predicting the hydraulic jump characteristics followed by realizable k-$\boldsymbol{\epsilon}$, SST k-$\boldsymbol{\omega}$, and standard k-$\boldsymbol{\epsilon}$.

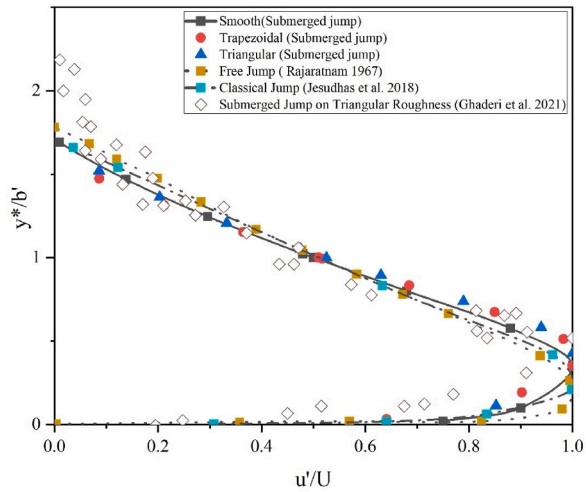


Fig. 10. Combined dimensionless flow velocity distribution of submerged hydraulic jump for smooth and macroroughness arrangements at Fr1 = 6.12.

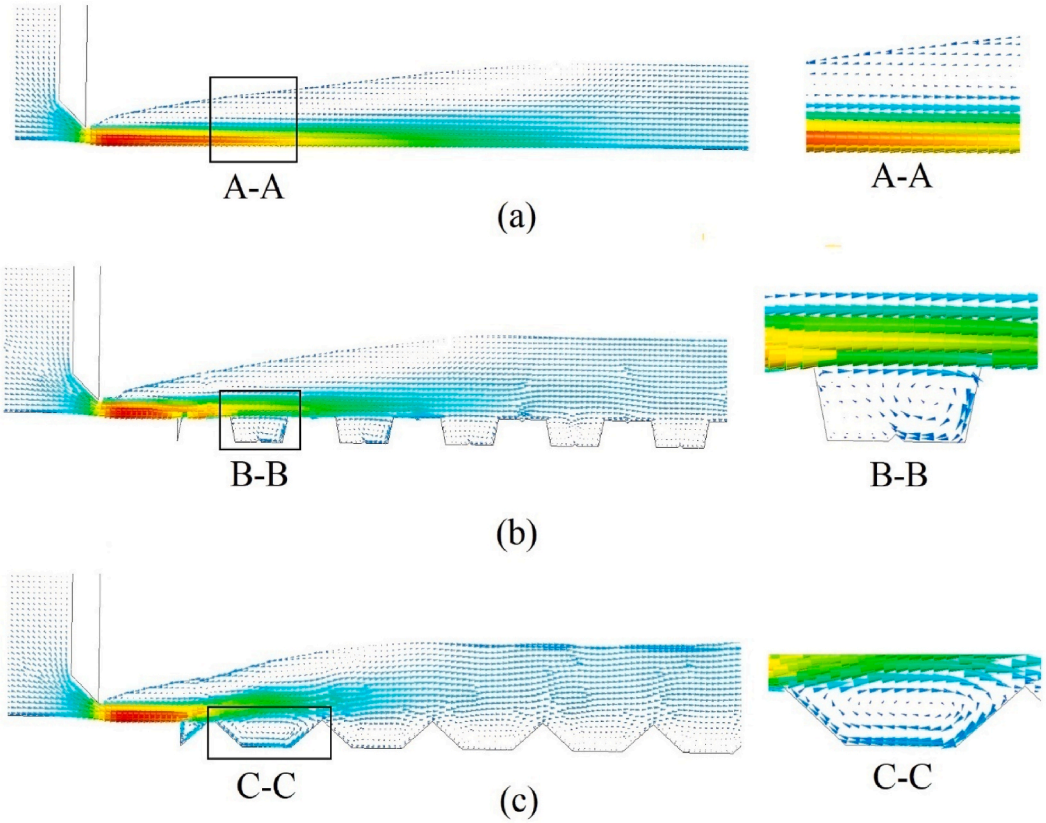


Fig. 11. Velocity vector field of free hydraulic jump (a) smooth bed (b) trapezoidal macroroughness (c) triangular macroroughness.

3. The trend of velocity profiles obtained for free &submerged jumps over smooth and macrorough beds are in good agreement with past studies.
4. It is reported that the empty space between two successive macrorough elements near the bed improves the jump characteristics.
5. No circular motion of velocity vector near the bed is seen in case of smooth bed.

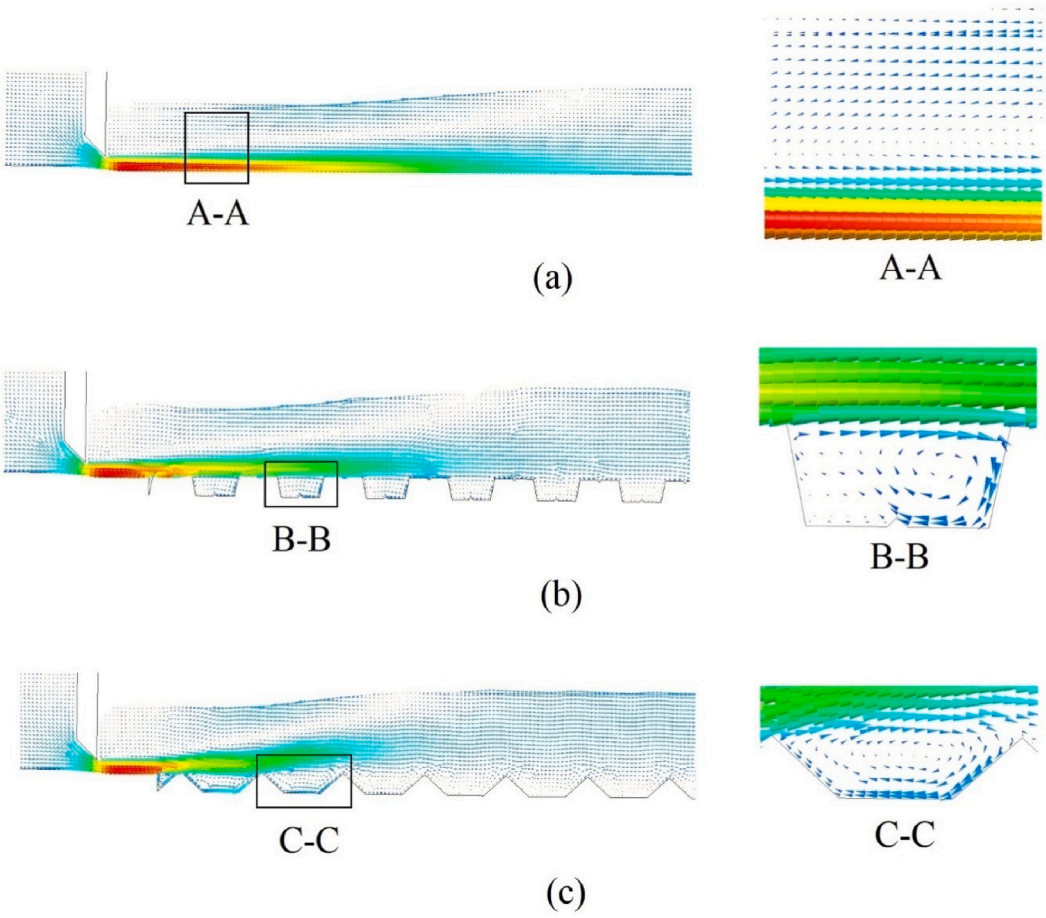


Fig. 12. Velocity vector field of submerged hydraulic jump (a) smooth bed(b) trapezoidal macroroughness (c) triangular macroroughness.

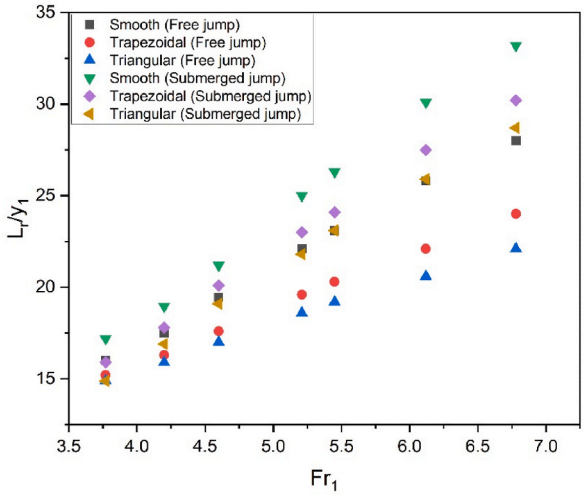


Fig. 13. Relative roller jump length vs Froude number.

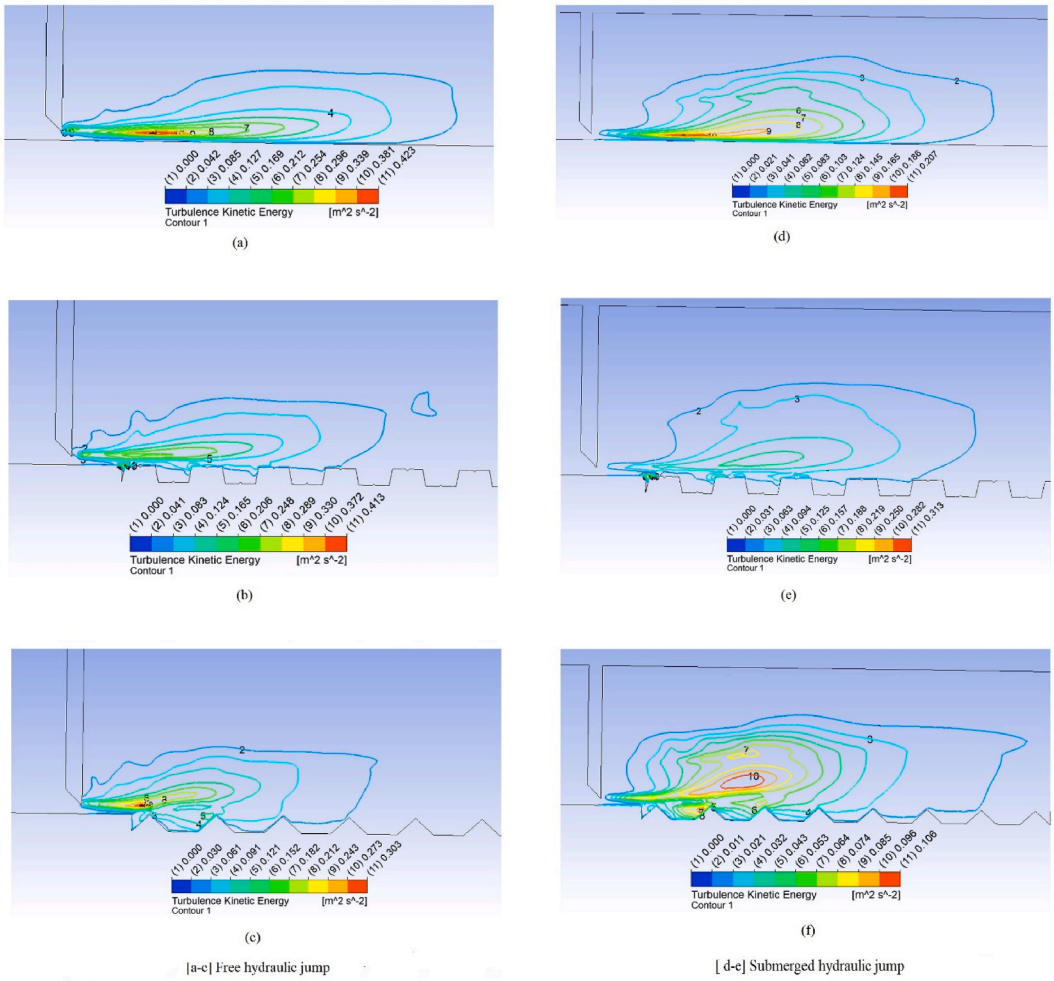


Fig. 14. Turbulent kinetic energy contours for different bed condition in free (a–c) and submerged (d–e) hydraulic jump.

6. Clockwise circular motion of velocity vector near the bed is observed in case of macrorough beds for both free &submerged hydraulic jumps.
7. The average value of the length of the roller for trapezoidal and triangular macro roughness is found to be 23 % and 46.06 % shorter than the smooth bed in the case of free jump.
8. The average value of the length of the roller for trapezoidal and triangular macro roughness is found to be 17.85 % and 36.33 % shorter than the smooth bed in the case of submerged jump.
9. Free hydraulic jump has a better impression over the submerged jump for improvement of characteristics of jump.
10. It is found that the triangular macroroughness has a better impression over trapezoidal macroroughness for improvement of characteristics of the jump.
11. The relative mean error for basic parameters of free & submerged hydraulic jump, such as sequent depth ratio and tail-water depth ratio, obtained numerically is reported within 6 % of the experimental result. Hence, supporting the reliability of the CFD approach in predicting hydraulic jump characteristics over macrorough beds.

As a final remark for future research, the new shape of macroroughness and their unique orientation, size, and stability may be investigated numerically at an affordable cost. Furthermore, this work may provide real-world hydraulic designers with improved insights for constructing the energy dissipator structure, and they may benefit from numerical modelling tools to analyse a complex fluid phenomenon.

Funding

The authors did not receive any specific funding for this research..

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Harshit Kumar Jayant: Conceptualization, Investigation, Methodology, Validation, Writing – original draft. **Bharat Jhamnani:** Supervision.

Declaration of competing interest

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

Acknowledgements

The authors want to express their sincere gratitude to the anonymous reviewers for contributing their time to this article and to DTU for providing research facilities.

References

[1] N. Rajaratnam, Hydraulic Jumps, ACADEMIC PRESS INC., 1967, https://doi.org/10.1016/b978-1-4831-9935-1.50011-2.
[2] W.H. Hager, Ressaut hydraulique du type b dans un canal a forte pente, J. Hydraul. Res. 26 (1988) 539–558, https://doi.org/10.1080/00221688809499192.
[3] W.H. Hager, R. Bremen, Sequent depths: le ressaut hydraulique classique: étude des hauteurs conjuguées, J. Hydraul. Res. 27 (1989) 565–585, https://doi.org/10.1080/00221688909499111.
[4] W.H. Hager, N.V. Bretz, Ressauts hydrauliques par-dessus des marches positives et négatives, J. Hydraul. Res. 24 (1986) 237–253, https://doi.org/10.1080/00221688609499303.
[5] V. Yakhot, S.A. Orszag, Renormalization group analysis of turbulence. I. Basic theory, J. Sci. Comput. 1 (1986) 3–51, https://doi.org/10.1007/BF01061452.
[6] N. Rajaratnam, Energy Dissipators and Hydraulic Jump, 1995, https://doi.org/10.1139/l95-075.
[7] A. Ghaderi, M. Dasineh, F. Aristodemo, A. Ghahramanzadeh, Characteristics of free and submerged hydraulic jumps over different macroroughnesses, J. Hydroinf. 22 (2020) 1554–1572, https://doi.org/10.2166/HYDRO.2020.298.
[8] N.D. Tokyay, Effect of channel bed corrugations on hydraulic jumps, World Water Congr, Impacts Glob. Clim. Chang. - Proc. 2005 World Water Environ. Resour. Congr 2005 (2005) 408, https://doi.org/10.1061/40792(173)408.
[9] W.H. Hager, R. Bremen, N. Kawagoshi, Length of rollerRessaut hydraulique classique: longueur du rouleau, J. Hydraul. Res. 28 (1990) 591–608, https://doi.org/10.1080/00221689009499048.
[10] F. Murzyn, H. Chanson, Free-surface fluctuations in hydraulic jumps: experimental observations, Exp. Therm. Fluid Sci. 33 (2009) 1055–1064, https://doi.org/10.1016/j.expthermflusci.2009.06.003.
[11] A. Ghaderi, M. Dasineh, F. Aristodemo, C. Aricò, Numerical simulations of the flow field of a submerged hydraulic jump over triangular macroroughnesses, Water (Switzerland) (2021) 13, https://doi.org/10.3390/w13050674.
[12] W.C. Hughes, J.E. Flack, Hydraulic jump properties over a rough bed, J. Hydraul. Eng. 110 (1984) 1755–1771, https://doi.org/10.1061/(asce)0733-9429, 1984)110:12(1755.
[13] S.A. Ead, N. Rajaratnam, Hydraulic jumps on corrugated beds, J. Hydraul. Eng. 128 (2002) 656–663, https://doi.org/10.1061/(asce)0733-9429, 2002)128:7(656.
[14] A. Abbaspour, A. Hosseinzadeh Dalir, D. Farsadizadeh, A.A. Sadraddini, Effect of sinusoidal corrugated bed on hydraulic jump characteristics, J. Hydro-Environment Res. 3 (2009) 109–117, https://doi.org/10.1016/j.jher.2009.05.003.
[15] N.D. Tokyay, D. Velioğlu, Effects of Bed Roughness on the Characteristics of Hydraulic Jumps, 2012, pp. 17–19.
[16] A. Abbaspour, D. Farsadizadeh, M.A. Ghorbani, Estimation of hydraulic jump on corrugated bed using artificial neural networks and genetic programming, Water Sci. Eng. 6 (2013) 189–198, https://doi.org/10.3882/j.issn.1674-2370.2013.02.007.
[17] H.M.A. Ahmed, M. El Gendy, A.M.H. Mirdan, A.A.M. Ali, F.S.F.A. Haleem, Effect of corrugated beds on characteristics of submerged hydraulic jump, Ain Shams Eng. J. 5 (2014) 1033–1042, https://doi.org/10.1016/j.asej.2014.06.006.
[18] P. Parsamehr, D. Farsadizadeh, A. Hosseinzadeh Dalir, A. Abbaspour, M.J. Nasr Esfahani, Characteristics of hydraulic jump on rough bed with adverse slope, ISH J. Hydraul. Eng. 23 (2017) 301–307, https://doi.org/10.1080/09715010.2017.1313143.
[19] S. Felder, H. Chanson, Air–water flow patterns of hydraulic jumps on uniform beds macroroughness, J. Hydraul. Eng. 144 (2018), 04017068, https://doi.org/10.1061/(asce)hy.1943-7900.0001402.
[20] N. Torkamanzad, A.H. Dalir, F. Salmasi, A. Abbaspour, Hydraulic jump below abrupt asymmetric expanding stilling basin on rough bed, Water (Switzerland) 11 (2019) 1–29, https://doi.org/10.3390/w11091756.
[21] S. Nikmehr, Y. Aminpour, Numerical simulation of hydraulic jump over rough beds, Period. Polytech. Civ. Eng. 64 (2020) 396–407, https://doi.org/10.3311/PPci.15292.
[22] K. Sayyadi, M. Heidarpour, Z. Ghadampour, Effect of bed roughness and negative step on characteristics of hydraulic jump in rectangular stilling basin, Shock Vib. 2022 (2022), https://doi.org/10.1155/2022/1722065.
[23] P. Parsamehr, A. Kuriqi, D. Farsadizadeh, A.H. Dalir, R. Daneshfaraz, R.M.L. Ferreira, Hydraulic jump over an adverse slope controlled by different roughness elements, Water Resour. Manag. 36 (2022) 5729–5749, https://doi.org/10.1007/s11269-022-03330-x.
[24] F. Bahmanpouri, C. Gualtieri, H. Chanson, Experiments on two-phase flow in hydraulic jump on pebbled rough bed: Part 1–Turbulence properties and particle chord time and length, Water Sci. Eng. (2023) 1–11, https://doi.org/10.1016/j.wse.2023.05.002.
[25] C. Gualtieri, H. Chanson, Physical and numerical modelling of air-water flows: an Introductory Overview, Environ. Model. Software 143 (2021), 105109, https://doi.org/10.1016/j.envsoft.2021.105109.
[26] B. Blocken, C. Gualtieri, Ten iterative steps for model development and evaluation applied to computational fluid dynamics for environmental fluid mechanics, Environ. Model. Software 33 (2012) 1–22, https://doi.org/10.1016/j.envsoft.2012.02.001.
[27] M. Dasineh, A. Ghaderi, M. Bagherzadeh, M. Ahmadi, A. Kuriqi, Prediction of hydraulic jumps on a triangular bed roughness using numerical modeling and soft computing methods, Mathematics 9 (2021), https://doi.org/10.3390/math9233135.
[28] A. Bayon-Barrachina, P.A. Lopez-Jimenez, Numerical analysis of hydraulic jumps using OpenFOAM, J. Hydroinf. 17 (2015) 662–678, https://doi.org/10.2166/hydro.2015.041.

[29] T.D. Canonsburg, ANSYS Fluent Theory Guide, vol. 15317, ANSYS Inc., USA, 2013, p. 814. http://www.afs.enea.it/project/neptunius/docs/fluent/html/th/main_pre.htm.
[30] L. Choufu, S. Abbasi, H. Pourshahbaz, P. Taghvaei, S. Tfwala, Investigation of flow, erosion, and sedimentation pattern around varied groynes under different hydraulic and geometric conditions: a numerical study, Water (Switzerland) (2019) 11, https://doi.org/10.3390/w11020235.
[31] Z. Lu, M.H.A. Piro, M.A. Christon, Mesh and turbulence model sensitivity analyses of computational fluid dynamic simulations of a 37M CANDU fuel bundle, Nucl. Eng. Technol. 54 (2022) 4296–4309, https://doi.org/10.1016/j.net.2022.06.004.
[32] F.E. Division, E.P. Statement, N. Accuracy, A.J.F. Eng, F. Engineering, E. Policy, N. Accuracy, A.J.F. Eng, Procedure for Estimation and Reporting of Uncertainty Due to Discretization in CFD Applications 130 (2008) 1–4, https://doi.org/10.1115/1.2960953.
[33] H. Chanson, C. Gualtieri, Similitude and scale effects of air entrainment in hydraulic jumps, J. Hydraul. Res. 46 (2008) 35–44, https://doi.org/10.1080/00221686.2008.9521841.
[34] H. Wang, H. Chanson, Self-similarity and scale effects in physical modelling of hydraulic jump roller dynamics, air entrainment and turbulent scales, Environ. Fluid Mech. 16 (2016) 1087–1110, https://doi.org/10.1007/s10652-016-9466-z.
[35] V. Jesudhas, R. Balachandar, V. Roussinova, R. Barron, Turbulence characteristics of classical hydraulic jump using DES, J. Hydraul. Eng. 144 (2018), 04018022, https://doi.org/10.1061/(asce)hy.1943-7900.0001427.

Uncorrected Proof

Journal of Hydroinformatics

Check for updates

© 2023 The Authors

Journal of Hydroinformatics Vol 00 No 0, 1 doi: 10.2166/hydro.2023.302

Numerical study of submerged hydraulic jumps over triangular macroroughnesses

Harshit Kumar Jayant * and Bharat Jhamnani

Department of Civil Engineering, Delhi Technological University, Delhi, India
*Corresponding author. E-mail: harshit_2k19phdce04@dtu.ac.in

HKJ, 0009-0002-3493-2925

ABSTRACT

The hydraulic jump is a phenomenon that occurs in open channels. In past studies, hydraulic jumps over smooth and macrorough beds have been investigated to enhance energy dissipation, but triangular macroroughness, specifically the right-angled triangular macroroughness, has not been dealt with. The objective of this article is to numerically investigate submerged hydraulic jumps over right angle and isosceles triangular macrorough beds. To achieve this, a numerical model based on computational fluid dynamics (CFD) has been utilized. Numerically obtained jump characteristics such as submerged depth ratio, tailwater depth ratio, longitudinal velocity profile, flow pattern in the cavity region, and energy dissipation have been presented in detail. In particular, initial Froude number reduction in both tailwater and submerged depth ratios as well as an increase in the energy dissipation of submerged hydraulic jumps have been noticed on isosceles triangular macroroughness with different arrangements, as compared to smooth beds. The present numerical model has been validated with the experimental model, and the mean error between the two for submerged depth and tailwater depth ratios was found to be below 6%. This confirms the adequacy of the present CFD model in predicting relevant submerged hydraulic jump characteristics over macrorough beds.

Key words: ANSYS fluent, computational fluid dynamics, submerged hydraulic jump, triangular macroroughness

HIGHLIGHTS

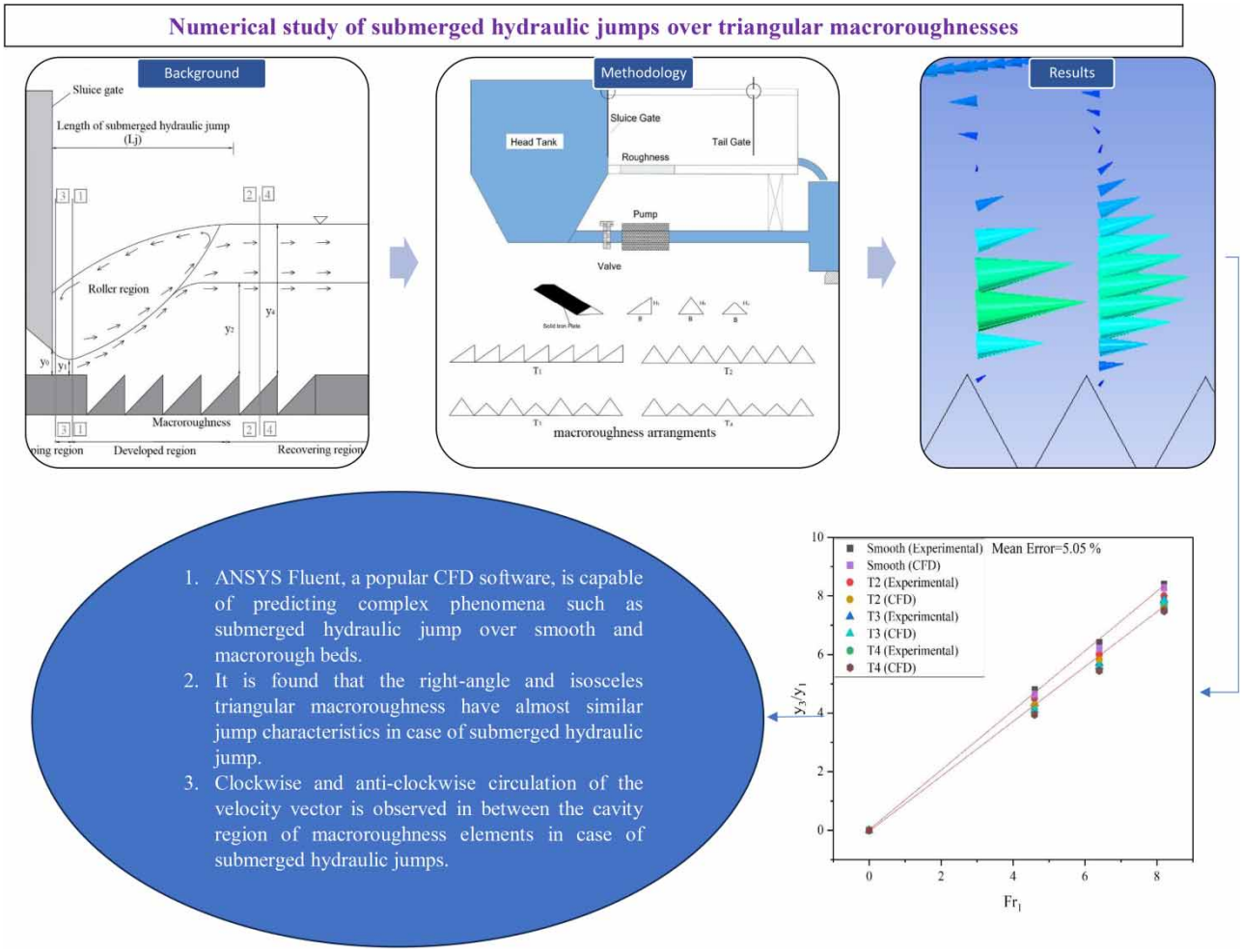
- The authors found that the isosceles triangular shape are more commonly used and right-angled shape has not been addressed in the past study.
- Right-angled triangular macro roughness is added as a new shape of triangular configuration.
- There is also a need to reinvestigating different macroroughness arrangements to enhance the jump characteristics, which had not been extensively explored in past studies.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

GRAPHICAL ABSTRACT



NOTATION

CFD	computation fluid dynamics
RANS	Reynolds-averaged Navier–Stokes
RNG	re-normalization group
VOF	volume of fluid
k	turbulent kinetic energy
ε	dissipation rate of turbulent kinetic energy
μ	dynamic viscosity
υ	kinematic viscosity
g	gravitational acceleration
v	depth-averaged velocity at a section where jump starts
u	time-averaged longitudinal velocity at any section
U	maximum velocity in the horizontal direction at any section
Q	discharge
ρ	fluid density
y_0	opening of sluice gate
y_1	initial depth of water in case of submerged hydraulic jump
y_2	tailwater depth of water in case of a free hydraulic jump
y_3	submerged depth of water in case of a submerged hydraulic jump
y_4	tailwater depth of water in case of a submerged hydraulic jump
H_1, H_2	height of macroroughness elements

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

B base width of macroroughness elements
Fr_1 Froude number of the inlet section of the hydraulic jump
C_r Courant number
L_j length of submerged hydraulic jump
E_1 specific energy at the upstream end of the jump
E_2 specific energy at the downstream end of the jump
R^2 coefficient of determination

INTRODUCTION

Jean–Baptiste Belanger developed the Belanger equation in 1838, which relates the pre- and post-jump depths, also known as sequent depths defined in Equation (1). This equation has been used extensively in open-channel hydraulics and has led to the development of various correlations between flow parameters upstream and downstream of the jump.

The length of the jump and the length of the roller of the jump are other important jump characteristics, and it measures the separation between the toe and the surface standstill region (Hager *et al.* 1990). The definition of the length of the jump is ambiguous, where L_j is the distance between the toe and the point at which the water surface nearly levels out and the mean surface elevation reaches its maximum (Rajaratnam 1967). The ratio of the sequent depths continues smaller than Belanger's equation because of the frictional forces on the wall (Hager & Bremen 1989).

$$\frac{y_2}{y_1} = \frac{1}{2}\left(\sqrt{1 + 8\mathrm{Fr}_1^2} - 1\right) \tag{1}$$

where y_1 and y_2 are the flow depths at the upstream of the jump and downstream of the jump, respectively, and Fr_1 and Fr_2 are the upstream and downstream Froude numbers, respectively. As the hydraulic jump is a phenomenon with non-deterministic characteristics, the characteristics of it over macroroughness can be investigated by a mathematical approach such as the stochastic approach (Guo *et al.* 2018; Mousavi *et al.* 2020). In addition, the dissolved and sediment fraction in hydraulic jump can be detected using the approach discussed by Lin *et al.* (2018). Further to tackle the multi-type uncertainty in the optimization of macroroughness in the case of hydraulic jump may use the concept of decision-making, combination, and prediction approach (Chen *et al.* 2018; Li *et al.* 2018). The past study researched this phenomenon in depth utilizing analytical, experimental, and numerical simulation approaches, which are discussed as follows:

Hughes & Flack (1984) created conceptual correlations by investigating hydraulic jumps in a horizontal rectangular flume over an artificial rough bed with smooth side walls. Hager & Bretz (1986) investigated hydraulic jumps across positive and negative steps in a rectangular prismatic channel. Ead & Rajaratnam (2002) carried out experiments on the smooth and sinusoidal corrugated beds, and investigated the characteristics of a free hydraulic jump. Elsebaie & Shabayek (2010) investigated hydraulic jump in a rectangular channel with uneven and smooth beds, with a side slope of 45 ° for two trapezoidal and triangular macroroughness and 60 ° for other trapezoidal. Afzal *et al.* (2011) performed hydraulic jumps in a rectangular channel over an uneven bed. Tokyay & Velioğlu (2012) researched hydraulic jumps across smooth and rough beds. Abbaspour *et al.* (2013) studied the hydraulic jump over just a sinusoidal corrugated bed in the rectangular channel using artificial neural networks and genetic programming. Samadi-Boroujeni *et al.* (2013) examined hydraulic jump in a rectangular channel with six smooth and six rough beds. Ahmed *et al.* (2014) experimentally explored the influence of spaced triangular strip corrugated beds on downstream submerged jump characteristics. Bayon-Barrachina & Lopez-Jimenez (2015) numerically modelled the free hydraulic jump in a rectangular channel with a smooth bed using Open FOAM. Sauida (2016) examined the submerged hydraulic jump downstream of multi-vent regulators under various gate scenarios. Hafnaoui *et al.* (2016) investigated hydraulic jump in a prismatic inclined channel with rectangular and triangular corrugation. Abbaspour *et al.* (2016) laid the plates at an angle of 50 and 90 ° at various distances from the apron in open channels with horizontal and reverse bedding to investigate the energy dissipation using a hydraulic jump. Eshkou *et al.* (2018) executed forced hydraulic jumps in a gradually diverging stilling basin Type III with baffle blocks of varying angles. Felder & Chanson (2018) performed hydraulic jumps in rectangular channels with smooth bed and micro-roughness. Al-Hashimi *et al.* (2019) used the computational fluid dynamics (CFD) approach with ANSYS Fluent to numerically model the flow over submerged weir. Kumar *et al.* (2019) tested hydraulic jumps across three roughness heights of crushed and rounded aggregates, as well as two positive bed slopes. Yildiz *et al.* (2020) numerically investigated both free and submerged hydraulic jumps using ANSYS

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

Fluent. Macián-Pérez *et al.* (2020b) attempted using Open FOAM and FLOW-3D to evaluate a traditional hydraulic jump using CFD. Mahtabi *et al.* (2020) examined free and submerged hydraulic jumps across natural and artificial roughness using a decision tree method and a multi-layer neural network. Ghaderi *et al.* (2020) performed numerical simulation for free and submerged jumps in a rectangular channel with triangular, square, and semi-oval macroroughness. Nikmehr & Aminpour (2020) used trapezoidal roughness at the bed to investigate the hydraulic jump characteristics in a rectangular flume. Maleki & Fiorotto (2021) used an analytical approach to investigate the hydraulic jump stilling basins over an uneven substrate using experimental data. Dasineh *et al.* (2021) explored the free and forced hydraulic jump in a rectangular channel on the triangular platform in different configurations and compared it to the prediction of this numerical model using artificial intelligence approaches. Ghaderi *et al.* (2021) investigated the effects of free and forced hydraulic jumps in a rectangular channel on triangular macroroughness.

Based on the aforementioned literature, it seems that the cavity region between two successive elements type energy dissipators was effective for energy dissipation, and hence, many past studies had investigated rough beds specifically in the form of strip corrugation to create a small bucket type structure. The past study considered various shapes of roughness element and found that the triangular shape yields good results. Here, the authors found that the isosceles triangular shape is more commonly used, and the right-angled shape has not been addressed in previous studies. In addition, there is also a need to reinvestigating different macroroughness arrangements such as T_1, T_2, T_3, and T_4 (Figure 3(b)) to enhance the jump characteristics, which have not been extensively explored in the past studies.

Therefore, the objective of this study is to investigate numerically the effects of different bed macroroughness configurations with their unique arrangements on submerged hydraulic jump characteristics such as tail water and submerged depth ratio, including velocity profiles. By examining these parameters, the study aims to provide further insights into the behaviour of submerged hydraulic jumps over macroroughness.

Submerged jump

The submerged hydraulic jump forms when the downstream tailwater depth is higher than the sequent depth of the free jump, and in other words, it occurs when the jump is completely submerged underwater and is not visible at the surface. A schematic diagram in Figure 1 is presents a detailed explanation of the submerged hydraulic jump phenomenon where a sluice gate is installed at the boundary of the upstream and downstream sections. The upstream section contains reservoir conditions, whereas the sluice gate acts as a hydraulic structure that releases excess water from the reservoir, which converts potential energy into kinetic energy. This excess energy must be treated near the hydraulic structure to minimize the chance of damage to the downstream bed and hydraulic structure. The hydraulic jump is a widely used phenomenon that has been adopted in the past to manage this excess energy. The length and depth of the jump are important parameters that can be managed to reduce the damage to the structure.

METHODOLOGY

The investigation was carried out using numerical simulation, and the same was validated using the experimental result of the present study as well as compared with the past studies. In the first step, experiments were performed in the laboratory of hydraulics and the same parameters were varied in the numerical simulation to get the output. A detailed sequence of methodology is presented in Figure 2.

Experimental background

An experimental setup for a hydraulic jump typically involves a horizontal rectangular channel with a specific length, width, and depth for a constant flow rate of water through the channel. The flow rate can be controlled using a flow control valve or a pump, while a flowmeter can be used to measure the flow rates. The bed of the channel can be made of different materials and configured in different ways to simulate bed roughness. The channel can also have different inlet and outlet configurations such as a sharp-crested weir to induce the hydraulic jump.

In the present study, experiments were performed in a rectangular flume of 3.0 m long, 0.5 m deep, and 0.5 m wide in the laboratory of hydraulics of Delhi Technological University, Delhi, India. In the present study, bed macroroughness in the form of solid iron plates of specification $B = 0.07$ m, $H_1 = 0.1$ m, and $H_2 = 0.05$ m (Figure 3(b)) was laid to the full width of the flume. The length of bed macroroughness remains fixed for all the experimental runs. To investigate the flow field

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

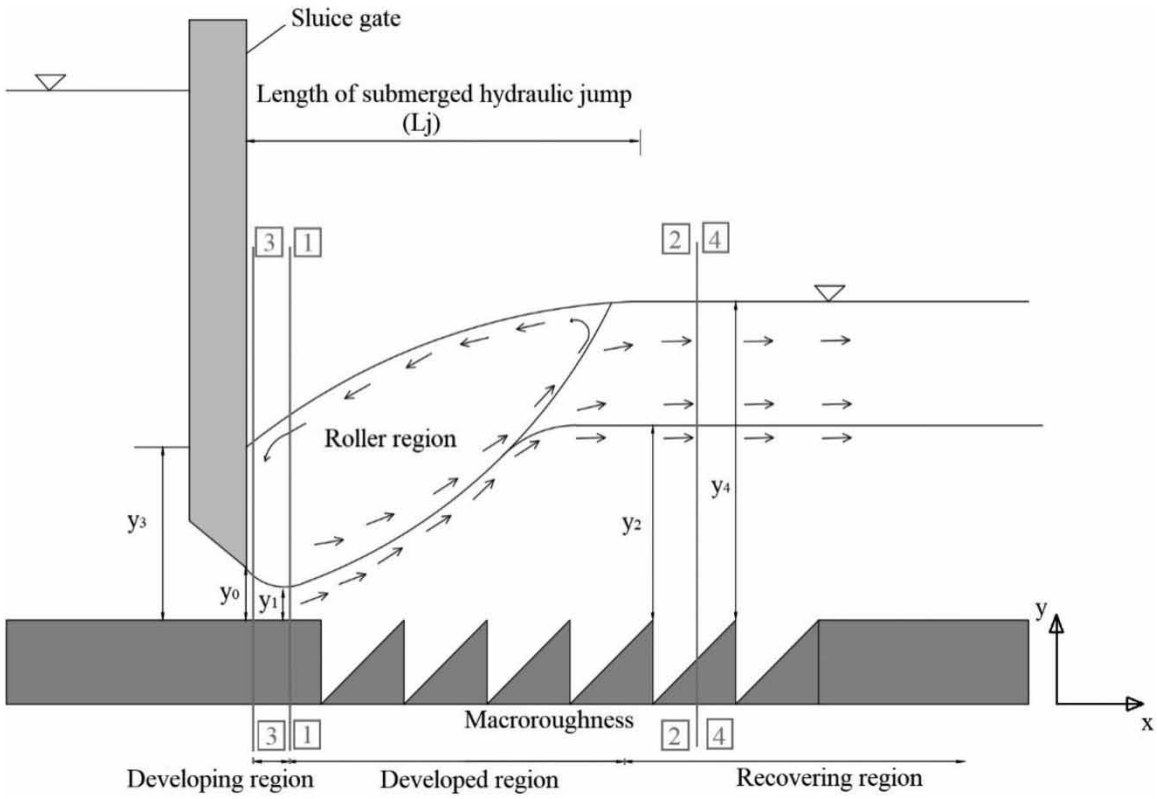


Figure 1 | Sketch of submerged hydraulic jump showing relevant parameters associated with the phenomenon.

and other parameters of the jump, different arrangements of bed roughness were considered as shown in Figure 3. A total of '25' runs ('5' for each type of bed condition) on the smooth and macrorough bed were performed on a mild bed slope with different Fr_1 ranging from 4.0 to 8.5 by varying the discharge and sluice gate opening. Flow velocity was measured with the help of a digital pitot tube at the initial section of the gate opening and where the water levels in the downstream side only. Details of flume arrangements are presented in Figure 3(a).

Numerical simulation

Numerical simulation is the process of using computer algorithms to replicate the behaviour of a real-world system or process, typically in the form of mathematical models or equations. It involves breaking down complex phenomena into smaller components that can be mathematically modelled and then using numerical methods to solve these models and generate data that can be used to understand and analyse the behaviour of the system.

Numerical simulation is used in a wide range of fields, including physics, engineering, chemistry, biology, and economics. Some common applications include simulating the behaviour of fluids, predicting weather patterns, modelling the spread of infectious diseases, and designing and testing new products.

The accuracy and reliability of numerical simulations depend on a variety of factors, including the quality of the underlying mathematical models, the accuracy of the numerical methods used to solve these models, and the quality and quantity of the input data used to initialize the simulation. As a result, numerical simulation is an iterative process that often involves refining and improving the models and methods over time.

Computational fluid dynamics

To conduct a CFD investigation of hydraulic jump characteristics, one needs to model the flow using appropriate numerical methods. Some of the most common CFD models and their simplified equations are Reynolds-averaged Navier–Stokes (RANS) equations, which is a time-averaged equation for the motion of fluids that accounts for the turbulence effects by introducing additional terms that describe the turbulent viscosity and diffusion. Bayon-Barrachina & Lopez-Jimenez (2015);

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

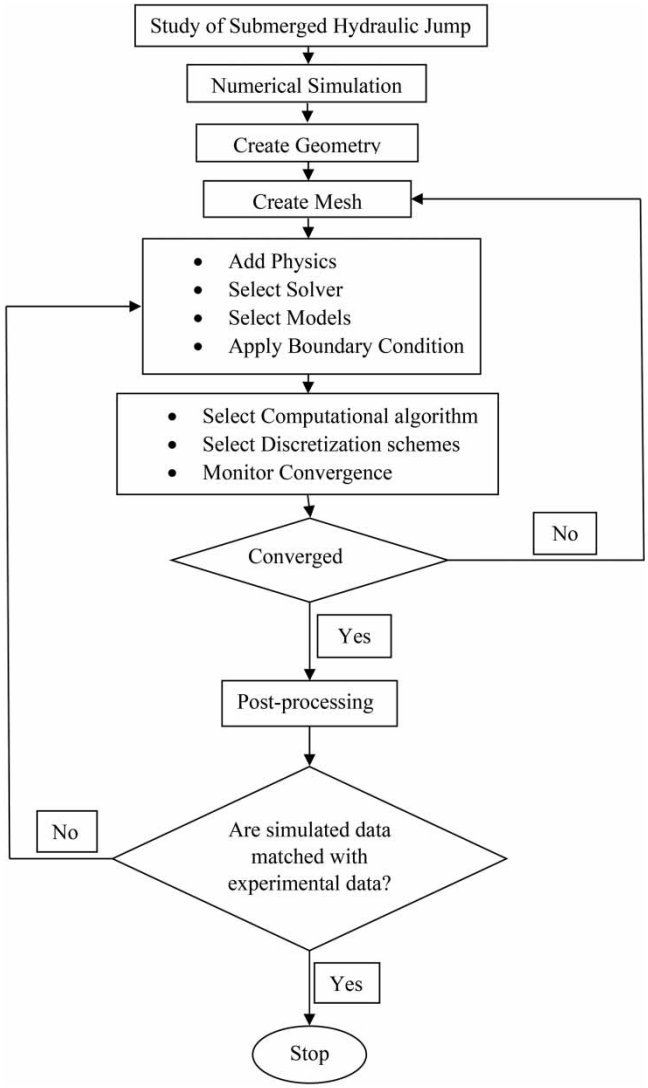


Figure 2 | Flowchart of the adopted methodology.

Dasineh *et al.* (2021); Ghaderi *et al.* (2021) used CFD for the use of the sophisticated algorithm volume of fluid (VOF) model for tracking the free surface of the flow. Here, numerical simulations were worked out using ANSYS Fluent 2022 R1, which is a well-known licensed software. Software is based on finite volume methods in the Cartesian form to solve the RANS equation. Two-phase (i.e., gas and liquid) for the present study was considered, where air is one of the forms of gas, and water is one of the forms of liquid.

Basic equations in CFD are presented in Equations (2) and (3):

(a) Continuity:

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{2}$$

(b) Navier–Stokes equation:

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial t_{ji}}{\partial x_j} \tag{3}$$

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

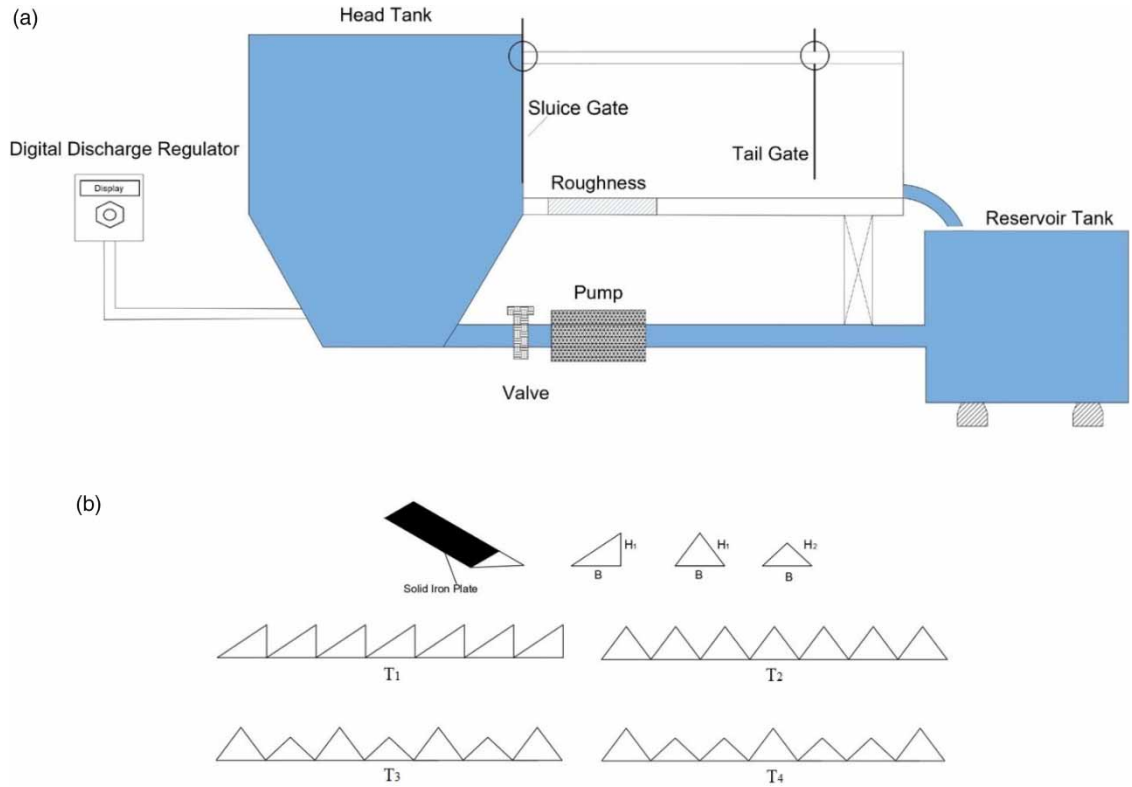


Figure 3 | (a) Sketch of experimental setup and (b) different types of roughness arrangements.

In the given context, the variables and symbols mentioned earlier are defined as follows:
u_i: Velocity vector in a three-dimensional space, where i represents the Cartesian coordinates (x, y, z).
x_i: Position vector in a three-dimensional space, where i represents the Cartesian coordinates (x, y, z).
t: Time variable.p: Pressure of the fluid. ρ: Density of the fluid.
t_{ji}: Viscous stress tensor. The subscript ij denotes the components of the tensor.
It is a mathematical construct used to describe the internal forces within a fluid caused by its viscosity.
VOF tracking equations:
The VOF equation is expressed in Equation (4):

$$\frac{\partial F}{\partial t} + \frac{1}{V_f}\left[\frac{\partial (FA_x u_x)}{\partial x} + \frac{\partial (FA_y u_y)}{\partial y} + \frac{\partial (FA_z u_z)}{\partial z}\right] = 0 \tag{4}$$

In the aforementioned equation, F is the fraction function that determines whether a cell is empty or filled. It is a function of position and represents the fraction of the cell occupied by a fluid. If a cell is empty (i.e., no fluid is present), then $F = 0$. This indicates that the fraction of the cell volume occupied by the fluid is zero. If a cell is filled (i.e., fluid is present), then $F = 1$ (Canonsburg 2013; Choufu *et al.* 2019). This indicates that the fraction of the cell volume occupied by the fluid is 1, representing a fully filled cell. The purpose of the fraction function F is to determine the distribution of fluid and the location of the free surface within a computational domain. By assigning different values to F, it is possible to determine the intermediate amounts of fluid present in a cell, allowing for the determination of the free surface position.

Typically, the free surface is determined at a position where the fraction function has an intermediate value, often $F = 0.5$. This means that at the free surface, half of the cell volume is occupied by the fluid and the other half is empty. However, it is

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

worth noting that the choice of the specific intermediate amount for determining the free surface position can vary depending on the specific application or user preferences.

Geometry and mesh

Geometry was prepared by keeping the experimental flume configuration in mind to avoid any type of mismatch (Figure 7(a)). In the next steps, mesh analysis was carried out in a meshing setup, where proper names were provided to different faces. In addition, inflation layers at the bed of the flume were provided to enhance the flow behaviour.

Turbulence model

Ansys provides various turbulence models that can be used for simulating turbulent fluid flows. Turbulence models are mathematical formulations that simulate the effects of turbulence on fluid flow, such as eddies and vortices. The most commonly used turbulence models in Ansys are the RANS models, which assume that the turbulence is statistically steady and can be averaged over time. These models include the k-ε model, the Spalart–Allmaras model, and the Reynolds Stress model, among others.

The software provides the facility of different turbulent models, but the exact and appropriate model depends on the phenomenon of the study. The k-ε model is one of the most widely used RANS models and is based on the assumption that the turbulent kinetic energy (TKE) (k) and its dissipation rate (ε) are the two main variables that determine the turbulence characteristics of the flow. Bayon-Barrachina & Lopez-Jimenez (2015) investigated the efficiency of re-normalization group (RNG) k-ε, standard k-ε, and SST k-ω turbulent models for the study of hydraulic jump over a smooth bed and stated that the most accurate model for hydraulic jump analysis is RNG k-ε followed by SST k-ω and standard k-ε. Ghaderi *et al.* (2021) said that RNG k-ε is similar to the standard k-ε model, even if it contains some refinement. Nikmehr & Aminpour (2020) proved that the RNG k-ε model efficiently predicts the hydraulic jump over a corrugated bed also. The previous studies (Salmasi & Samadi 2018; Hien & Duc 2020; Macián-Pérez *et al.* 2020a; Yildiz *et al.* 2020; Yamini *et al.* 2021) employed the turbulent k-ε model in their research. The RNG k-ε approach uses statistical techniques to obtain the averaged equations for two turbulence quantities: TKE (k) and its dissipation rate (ε). One benefit of this model is that, compared to the standard k-ε, it typically yields superior results when simulating whirling flows (Canonsburg 2013; Bayon-Barrachina & Lopez-Jimenez 2015). On the basis of the past pieces of literature, the authors attempted to employ a two-equation RNG k-ε turbulence model for the CFD technique to investigate the flow field of the submerged hydraulic jump throughout the analysis.

Turbulence equations are presented in Equations (5) and (6) as follows:

$$\frac{\partial}{\partial t}(\rho\varepsilon) + \frac{\partial}{\partial x_i}(\rho\varepsilon u_i) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_i}\right] + G_k + G_b - \rho\varepsilon - Y_M + S_k \tag{5}$$

$$\frac{\partial}{\partial t}(\rho\varepsilon) + \frac{\partial}{\partial x_i}(\rho\varepsilon u_i) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_\varepsilon}\right)\frac{\partial \varepsilon}{\partial x_i}\right] + C_{1_\varepsilon}\frac{\varepsilon}{k}(G_k + C_{3_\varepsilon}G_b) - C_{2_\varepsilon}\frac{\varepsilon^2}{k} + S_\varepsilon \tag{6}$$

$$\mu_t = C_\mu \frac{\rho k^2}{\varepsilon} \tag{7}$$

The variables and symbols in the aforementioned equation are defined as follows:

k: TKE, which represents the energy associated with the turbulent fluctuations in the fluid flow. It is a measure of the intensity of turbulence in the flow.
ε: dissipation rate of TKE. It represents the rate at which turbulent energy is converted into heat through viscous dissipation.
t: time variable.
ρ: density of the fluid.
x_i: coordinate in the i-axis of a Cartesian reference system.
μ: dynamic viscosity of the fluid.
μ_t: turbulent eddy dynamic viscosity. It is a model parameter that represents the effective viscosity associated with turbulence effects.
G_k: production term for TKE. It represents the rate at which TKE is generated within the flow due to the action of turbulent stresses.

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

G_b: buoyancy effect term. It represents the influence of buoyancy forces on the turbulent flow.

Y_M: dilatation oscillation effect term. It accounts for the effects of compression and expansion of fluid elements on the turbulence.

S_k and S_ε: source terms for TKE (k) and dissipation rate of TKE (ε), respectively. These terms represent additional contributions to the rate of change of k and ε in the turbulence model.

C_μ, C_{1_ε}, C_{2_ε}, C_{3_ε}, σ_k, and σ_ε: model parameters in the RNG k-ε turbulence model. These parameters have specific values assigned to them to govern the behaviour and accuracy of the turbulence model. In the RNG k-ε model, the specific values for these parameters were 0.084, 1.42, 1.68, 1.0, 0.7194, and 0.7194, respectively. These variables and symbols were used in the RNG k-ε turbulence model (Canonsburg 2013), which was a specific turbulence model within CFD simulations. The model equations and terms involving these variables were utilized to describe the behaviour and evolution of turbulence in the fluid flow, providing insights into turbulence-related phenomena and their impact on the overall flow characteristics.

Boundary conditions

In a hydraulic jump simulation, boundary conditions were used to define the flow conditions at the inlet and outlet of the computational domain. The boundary condition is essential in hydraulic jump analysis as the flume has different surfaces like an inlet, outlet, wall, and top. Choosing the appropriate boundary conditions depends on the specific hydraulic jump being simulated and the level of accuracy required. It is important to carefully select the boundary conditions to ensure that the simulation accurately represents the real-world hydraulic jump.

Hydraulic jump is known to be an open-channel flow, and for this, ANSYS Fluent has provided a toggle key for open-channel flow, which had been chosen for the simulation purpose in the present study. To produce the jump in the vicinity of the flume pressure inlet was decided for all the flume configurations whereas gauge pressure had opted at the top surface and outlet of the flume. The wall was treated as a no-slip wall, which assumes that the flow velocity at the wall is zero and a wall roughness constant was taken equal to 0.5. Initial conditions at the inlet and outlet were set as similar to experimental conditions.

Near-wall treatment

A critical part of CFD is near-wall treatment. It involves precisely modelling and resolving the flow behaviour near solid boundaries, as flow characteristics in this region can have a substantial impact on the overall simulation results. Wall functions are empirical correlations that estimate flow variables along the wall using information from the adjoining cell. By avoiding the resolution of the near-wall region, these functions reduce computational effort. It is important to create a mesh capable of precisely predicting the velocity gradient over the boundary layer. These criteria cannot be met for turbulent flows in complex geometry since it would necessitate an extremely fine mesh resolution near the wall, which would significantly increase the time necessary to solve the problem. The mesh should not be over-refined because the wall function resolves the flow near the wall, which may result in less accurate results (Bayon-Barrachina & Lopez-Jimenez 2015). To address this requirement, a wall function was developed that enables the usage of a 'relatively' larger mesh near the wall.

The non-dimensional distance between the wall and the first node away from the wall is given by y^+, which is an essential aspect of wall functions (Canonsburg 2013). The turbulent boundary layer is categorised into three sections based on y^+: laminar sub-layer ($y^+ < 5$), transition layer ($5 < y^+ < 30$), and turbulent ($y^+ > 30$). Near-wall treatment options in ANSYS Fluent for the k-ε turbulence model incorporate normal wall function, non-equilibrium wall function, enhanced wall treatment, and others. The present study used the standard wall function as it was the most commonly used wall function in ANSYS Fluent because it provides reasonably accurate predictions for the majority of high Reynolds number, wall-bounded flows and it lowers computational effort by eliminating the resolution of the near-wall region. The dimensionless y^+ value for all the grid systems fell between 46 and 113, which supported the need for the wall function technique. The range of inlet Reynolds number was 31,654–52,861.

Discretization scheme

ANSYS Fluent uses various discretization schemes to solve the governing equations of fluid flow numerically. These discretization schemes are used to transform the partial differential equations (PDEs) that describe the fluid flow into a set of algebraic equations that can be solved using a computer. The most commonly used discretization schemes in ANSYS Fluent are finite volume and finite element methods. Finite volume methods are used to discretize the governing equations

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

on a mesh of control volumes, while finite element methods are used to discretize the equations on a mesh of finite elements. In finite volume methods, the control volume is a small, discrete region of space that encloses a small portion of the fluid being simulated. The PDEs are then integrated over the control volume and approximated using numerical methods to obtain an algebraic equation for each control volume. These equations are then assembled into a global matrix equation that can be solved using numerical methods.

ANSYS Fluent also offers different discretization schemes for the spatial and temporal derivatives of the governing equations. For example, for the spatial derivatives, Fluent offers first-order upwind, second-order upwind, and second-order central differencing schemes. These schemes differ in the way they approximate the spatial derivatives, and they have different levels of accuracy and stability. Choosing the appropriate discretization schemes depends on the specific fluid flow being simulated and the desired level of accuracy and computational cost. It is important to carefully select the discretization schemes to ensure that the simulation accurately represents the real-world fluid flow.

Discretization is more influencing parameters in the case of numerical simulation to obtain the desired results. Bayon-Barrachina & Lopez-Jimenez (2015) stated that the RANS model, upwind schemes have less instability, and hence, the present study used the upwind schemes. The Courant number is a dimensionless parameter used in numerical simulations that represents the ratio of the time step to the distance between computational grid points. It was used to determine the stability of the simulation, as a Courant number that is too high can lead to numerical instability. In the present case of the submerged hydraulic jump simulation, the maximum Courant number used was 0.45, which was within the stable range for this type of simulation. The SIMPLE (Semi-Implicit Method for Pressure-Linked Equations) pressure–velocity coupling scheme was used, which was a widely used algorithm for solving fluid flow problems. The density and momentum equations were discretized using second-order upwind and first-order upwind schemes, respectively.

Stability and convergence criteria

Stability and convergence are critical factors in CFD modelling as they ensure the accuracy and reliability of the simulation results. Stability criteria in CFD modelling refer to the ability of the numerical solution to remain bounded as the simulation progresses. Convergence criteria refer to the ability of the numerical solution to converge to a solution that accurately represents the real-world fluid flow. In CFD modelling, the solution is considered converged when the numerical solution changes negligibly with each iteration. Convergence can be monitored by examining the residuals of the governing equations, which represent the error between the numerical solution and the exact solution. The most common convergence criterion is based on the magnitude of the residuals. When the residuals drop below a specified tolerance, the solution is considered converged.

For the present study to examine the convergence criteria, residuals were set to 10^{-6}. The stability of the numerical model for the hydraulic jump was quite difficult as the hydraulic jump was abrupt in nature. Here, the stability of the jump was achieved by the hit and trial method – first, by measuring the discharge equilibrium curve for the inlet and outlet, and second, by knowing the inlet and outlet pressure distribution. The stable conditions of the jump were noticed after 10 s of time steps.

Mesh sensitivity

Mesh sensitivity analysis is a statistical technique used to evaluate the sensitivity of a model output to variations in input parameters. It involves varying one or more input parameters over a predefined range of values, running the model for each set of input parameters, and analysing the resulting output to determine which input parameter has the most significant impact on the output. Mesh sensitivity analysis typically involves creating a grid of input parameter values and running the model for each combination of values in the grid. The output from each model run is then analysed using statistical methods to determine the sensitivity of the output to each input parameter.

Mesh sensitivity analysis was conducted on two basic parameters of submerged hydraulic jump – first on longitudinal velocity profile considering the basic procedure discussed by Celik *et al.* (2008) and second on the tailwater depth ratio as adopted by Bayon-Barrachina & Lopez-Jimenez (2015). Five mesh sizes of coarse to fine variety were considered. In the initial stage of the numerical simulation, a coarse mesh size of 50 mm was selected to plot the longitudinal velocity profile of submerged hydraulic jump at $x = 1.33$ m in the direction of flow. Numerically plotted data were compared with the experimental measured data, shown in in Figure 4. From Figure 4, it can be inferred that the 50 mm grid size is not sufficient to obtain the desired result, and hence, finer mesh sizes such as 30, 10, 7, and 5 mm were tested to obtain the desired result.

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

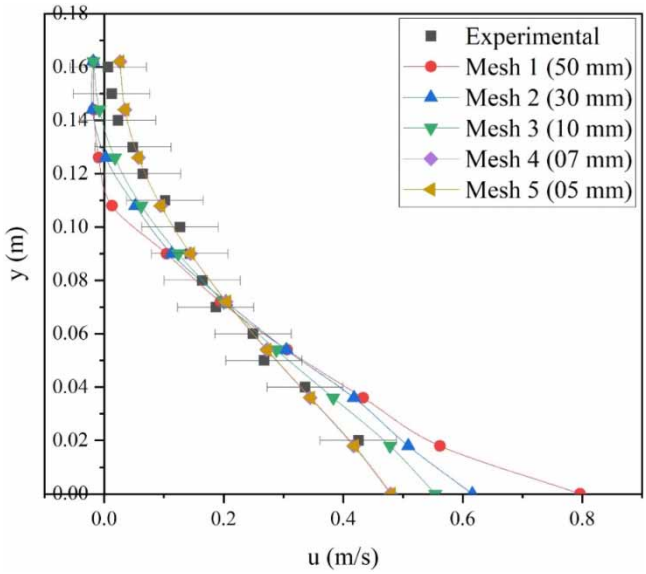


Figure 4 | Mesh size sensitivity analysis for longitudinal velocity profile in case of submerged jump over smooth bed at $Fr_1 = 4.6$.

Mesh sensitivity analysis result reflected that 7 and 5 mm mesh size has almost same data points, and hence, further reduction in the mesh size was not executed. The maximum percentage error of 5.35% was reported by comparing the numerically obtained longitudinal velocity profile with experimental findings. Similarly, the numerically obtained the tailwater depth ratio (y_4/y_1) for $Fr_1 = 4.6$ is compared with the experimental findings and achieved the least percentage error of 1.41% at 7 mm mesh size, which is tabulated in Table 1. Therefore, to avoid the computational cost, 7 mm mesh size was used throughout the simulation work.

RESULTS AND DISCUSSION

Validation of numerical simulation

Validation of numerical simulation refers to the process of determining whether the results of a computer simulation accurately represent the real-world phenomena being studied. It involves comparing the results of the simulation to experimental or observational data to assess the accuracy and reliability of the simulation. Accuracy and reliability of the simulation can involve quantifying the differences between the simulation results and the experimental or observational data and identifying areas where the simulation may need to be improved.

The comparison between numerical and experimental models on basic parameters of jump such as submerged depth ratio (y_3/y_1) and tail water depth ratio (y_4/y_1) are presented in Figure 5, which indicates a good agreement between the two models, with a mean error less than 6%. This level of agreement confirms the capability of CFD modelling to accurately simulate complex fluid dynamics problems such as submerged hydraulic jump. The derived equations for smooth and macrorough beds are

Table 1 | Mesh size sensitivity analysis for tailwater depth ratio in case of submerged jump over smooth bed

Mesh	Mesh size (mm)	$(y_4/y_1)_{Experimental}$	$(y_4/y_1)_{CFD}$	Percentage error
1	50	6.40	5.97	6.72
2	30	6.40	6.06	5.32
3	10	6.40	6.14	4.07
4	07	6.40	6.31	1.41
5	05	6.40	6.31	1.41

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

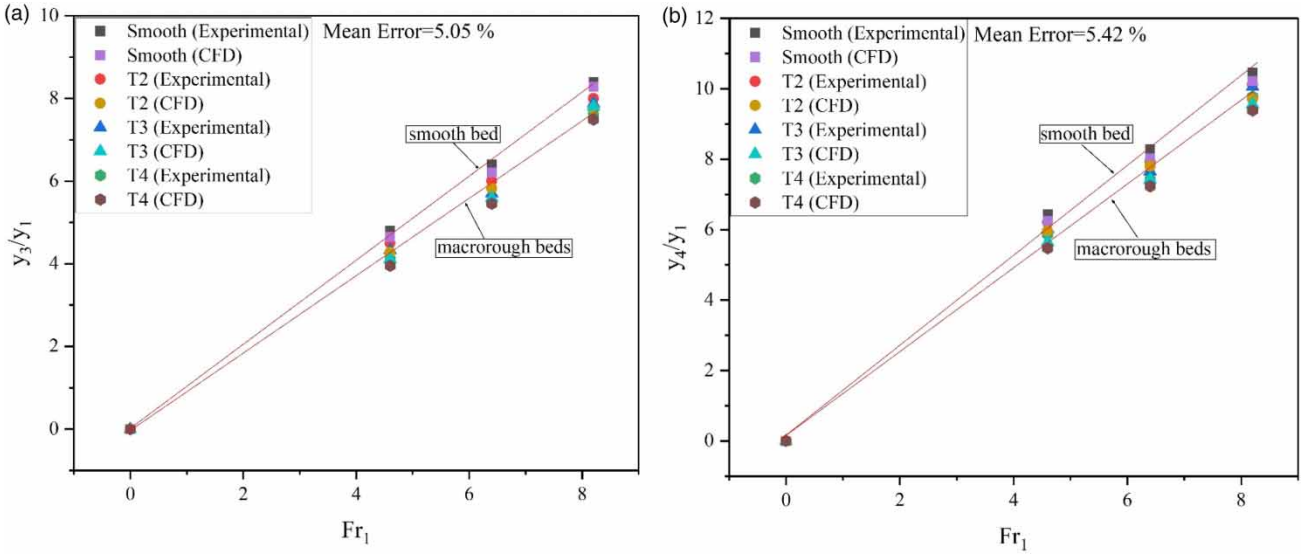


Figure 5 | Comparison of a numerical model with an experimental model for basic parameters of submerged hydraulic jump: (a) y_3/y_1 and (b) y_4/y_1.

presented in Equations (8)–(11):

$$y_3/y_1 = 1.0172\mathrm{Fr}_1 + 0.0174, \quad R^2 = 0.998 \quad \text{(smooth bed)} \tag{8}$$

$$y_3/y_1 = 0.9965\mathrm{Fr}_1 + 0.5886, \quad R^2 = 0.979 \quad \text{(macrorough beds)} \tag{9}$$

$$y_4/y_1 = 1.1113\mathrm{Fr}_1 + 0.1408, \quad R^2 = 0.995 \quad \text{(smooth bed)} \tag{10}$$

$$y_4/y_1 = 0.9313\mathrm{Fr}_1 + 0.5967, \quad R^2 = 0.979 \quad \text{(macrorough beds)} \tag{11}$$

Velocity distribution

In a hydraulic jump, the velocity distribution changes abruptly from a supercritical flow with high velocity to a subcritical flow with lower velocity. The velocity distribution in a hydraulic jump is complex and varies depending on the specific conditions of the flow. In the supercritical flow region upstream of the jump, the velocity is relatively constant and high. At the jump location, the velocity decreases rapidly to a subcritical flow region, where the velocity varies significantly across the depth of the flow. In the subcritical flow region downstream of the jump, the velocity is lower and varies with the depth, with the lowest velocities occurring near the bottom of the flow. This phenomenon is commonly referred to as the 'recovering zone' (Figure 1), where the fluid gradually recovers its downstream velocity profile after the hydraulic jump. The recovering

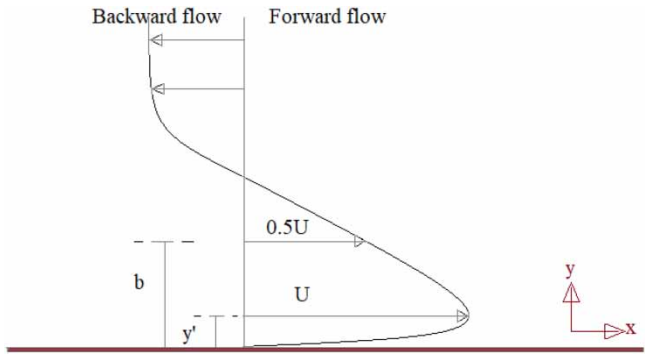


Figure 6 | Typical longitudinal velocity profile in a submerged hydraulic jump.

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

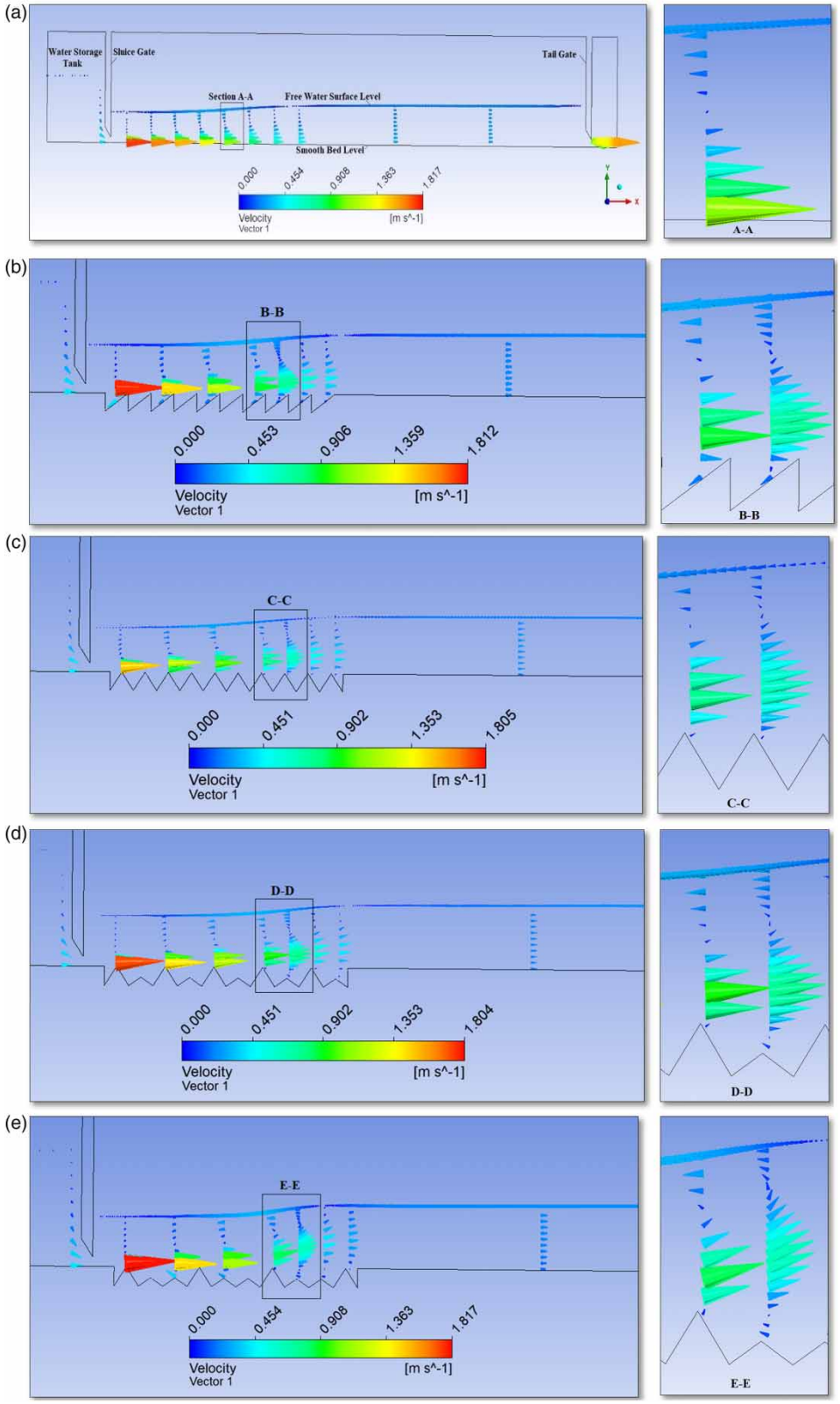


Figure 7 | Pictorial representation of the longitudinal velocity profile along the submerged hydraulic jump for $Fr_1 = 4.6$: (a) smooth bed, (b) T_1, (c) T_2, (d) T_3, and (e) T_4.

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

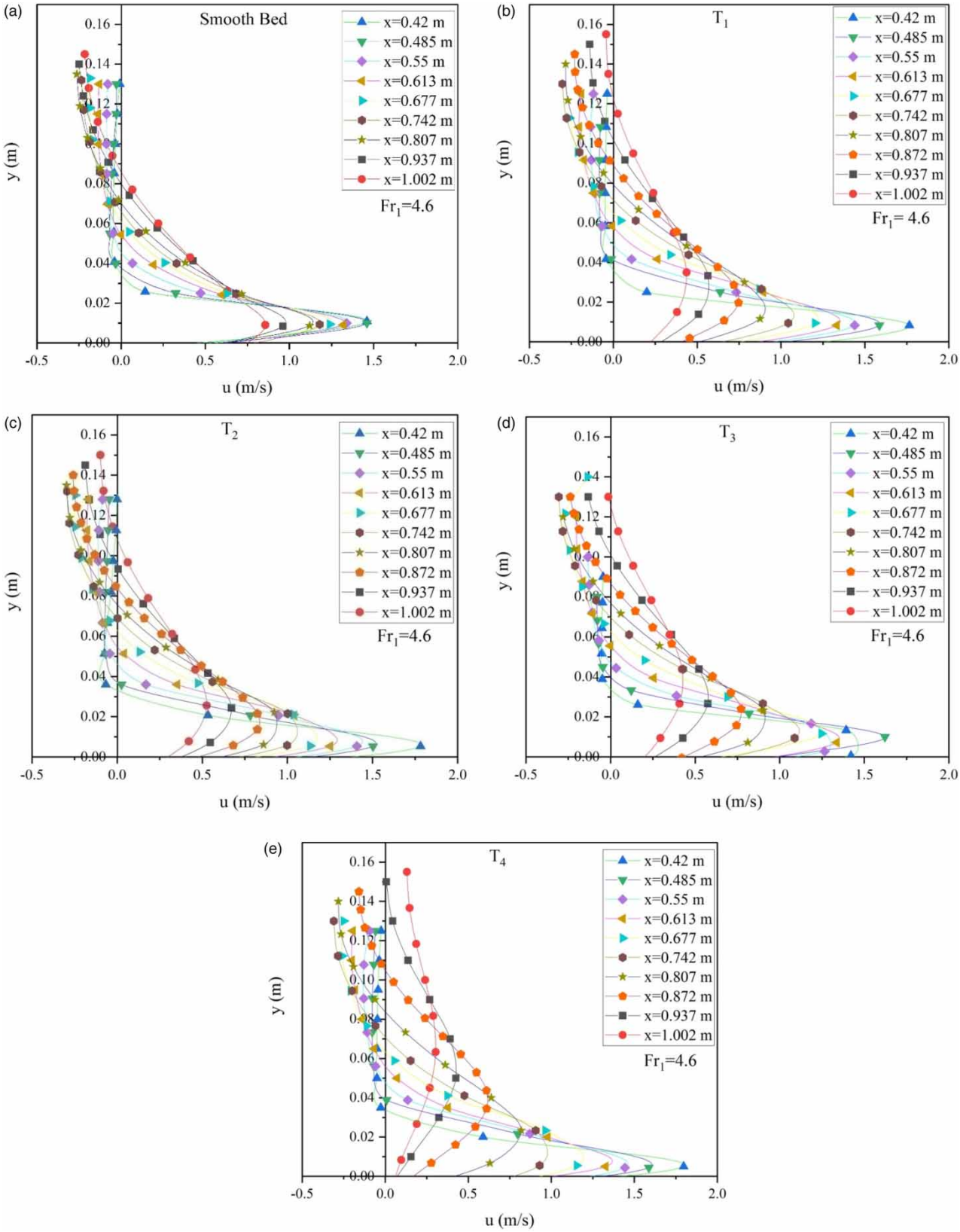


Figure 8 | Longitudinal velocity distribution at different sections of the flume in a submerged hydraulic jump over (a) smooth bed, (b) T_1, (c) T_2, (d) T_3, and (e) T_4.

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

zone is characterized by a gradually decreasing flow depth and a gradual increase in velocity as the fluid moves downstream. The extent of the recovering zone depends on several factors such as the initial Froude number (Fr_1), bed roughness, and channel geometry. The recovering zone is an important aspect to consider in hydraulic engineering design as it affects the downstream flow conditions and energy dissipation.

Wang & Chanson (2016) suggested that the upstream Froude number and velocity magnitude are inversely connected. Figure 6 depicts the general velocity profile in the hydraulic Jump (Rajaratnam 1967), where y' is the vertical distance from the bed level where maximum velocity can be found as U and at a vertical distance of b, and velocity reduces to almost half of the maximum velocity. This is a typical velocity variation in a fully formed roller zone, just above point b velocity stagnation takes place and reverse flow occurs.

Typical pictorial representations of numerically obtained longitudinal velocity profile of submerged hydraulic jump on smooth and macrorough beds are presented as follows.

On observing the results of Figure 7, it can be inferred that the velocity decreases at the top most section of the gate opening as the mean flow velocity is maximum in the vicinity of the gate. After the development of the jump, velocity in the roller region reverses towards the backward direction with a stagnation point. Such a description is valid for smooth and macrorough beds. Velocity distribution confirms the past velocity distribution of Rajaratnam (1967); Ead & Rajaratnam (2002); Ghaderi *et al.* (2021). In Figure 8, the graphical velocity distribution is shown for different sections within the flume, which also confirms the earlier description.

In Figure 9, the dimensionless velocity distribution of submerged hydraulic jump on smooth and macroroughness arrangements are plotted together for $Fr_1 = 4.6$. Here, the vertical axis represents the ratio of y/b, where y is the distance from the bed and b is the length scale at which velocity if half of the maximum velocity. The horizontal axis represents the ratio of u/U, where u is the time-averaged longitudinal velocity at any section and U is the maximum longitudinal velocity at any section.

The dimensionless velocity distribution graph is typically plotted for the point where the velocity is only half the maximum velocity, which is at a distance of b from the jump toe. This is because the velocity distribution is relatively flat in the downstream region and does not show significant variation. To present the dimensionless velocity distribution, additional calculations were performed to normalize the velocity values based on the mean velocity and the height of the jump. For the validation purpose, past studies (Rajaratnam 1967; Jesudhas *et al.* 2018; Ghaderi *et al.* 2021)were taken as references and presented in contrast with present data in Figure 9, and it was found that the trend of dimensionless velocity profiles is more or less similar to the past research and found that the maximum horizontal forward velocity occurred at lesser flow depth in comparison to studies by Rajaratnam (1967) and Jesudhas *et al.* (2018).

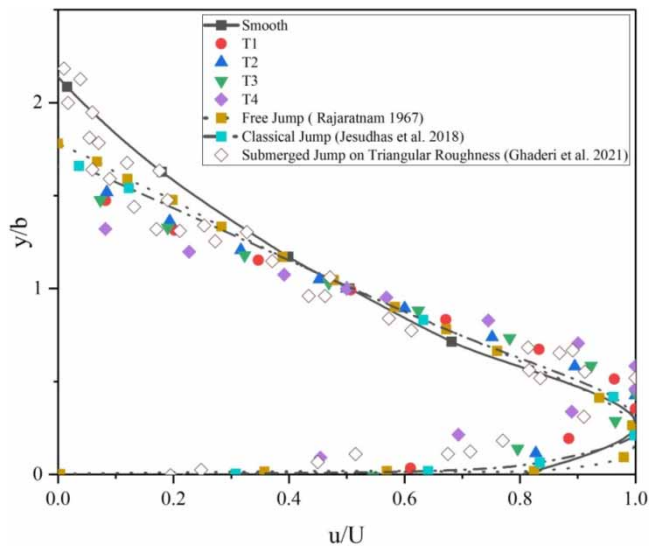


Figure 9 | Combined dimensionless flow velocity distribution of submerged hydraulic jump for smooth and macroroughness arrangements at $Fr_1 = 4.6$.

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

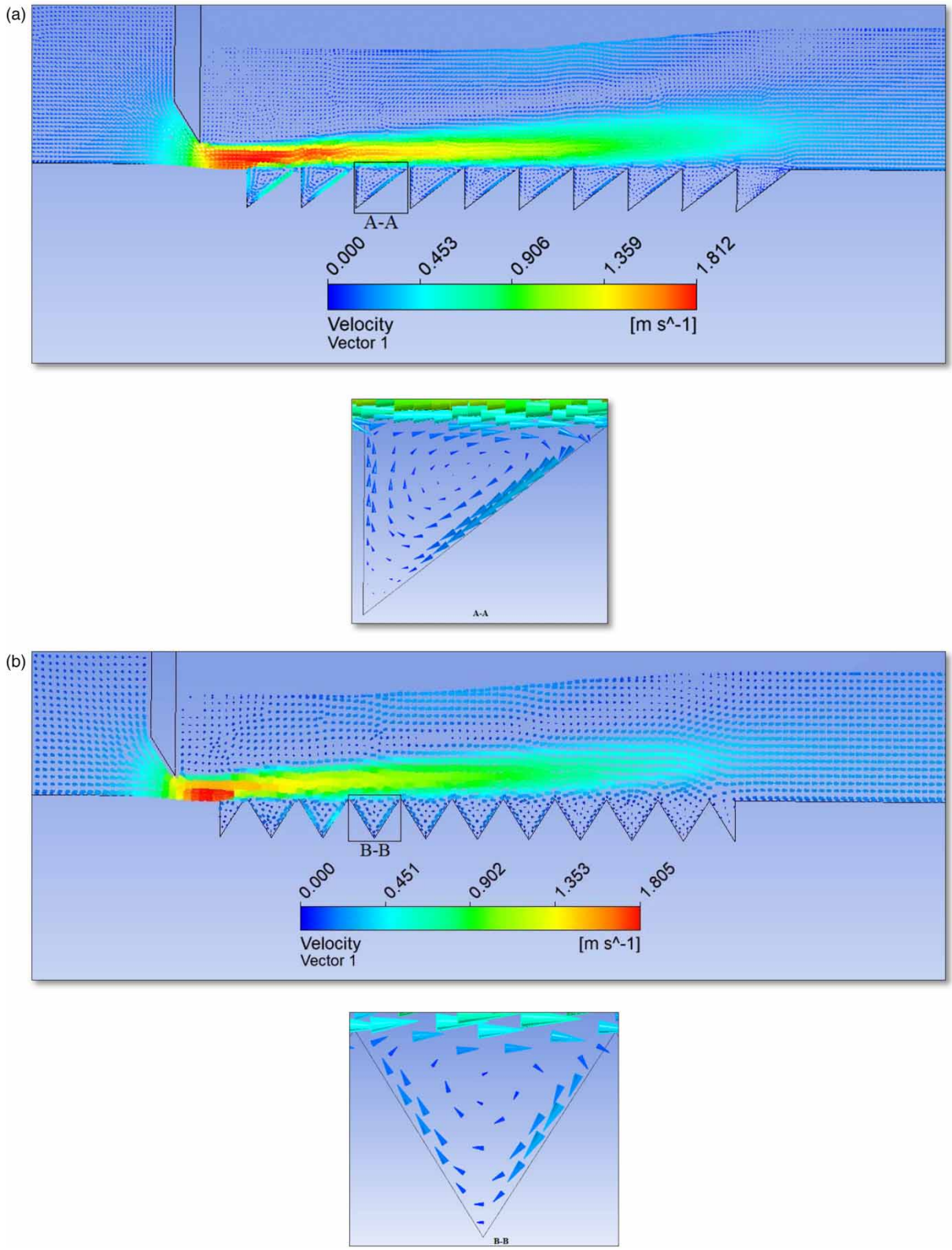


Figure 10 | Velocity vector field over the rough bed in case of submerged hydraulic jump: (a) T_1, (b) T_2, (c) T_3, and (d) T_4. (*continued.*).

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

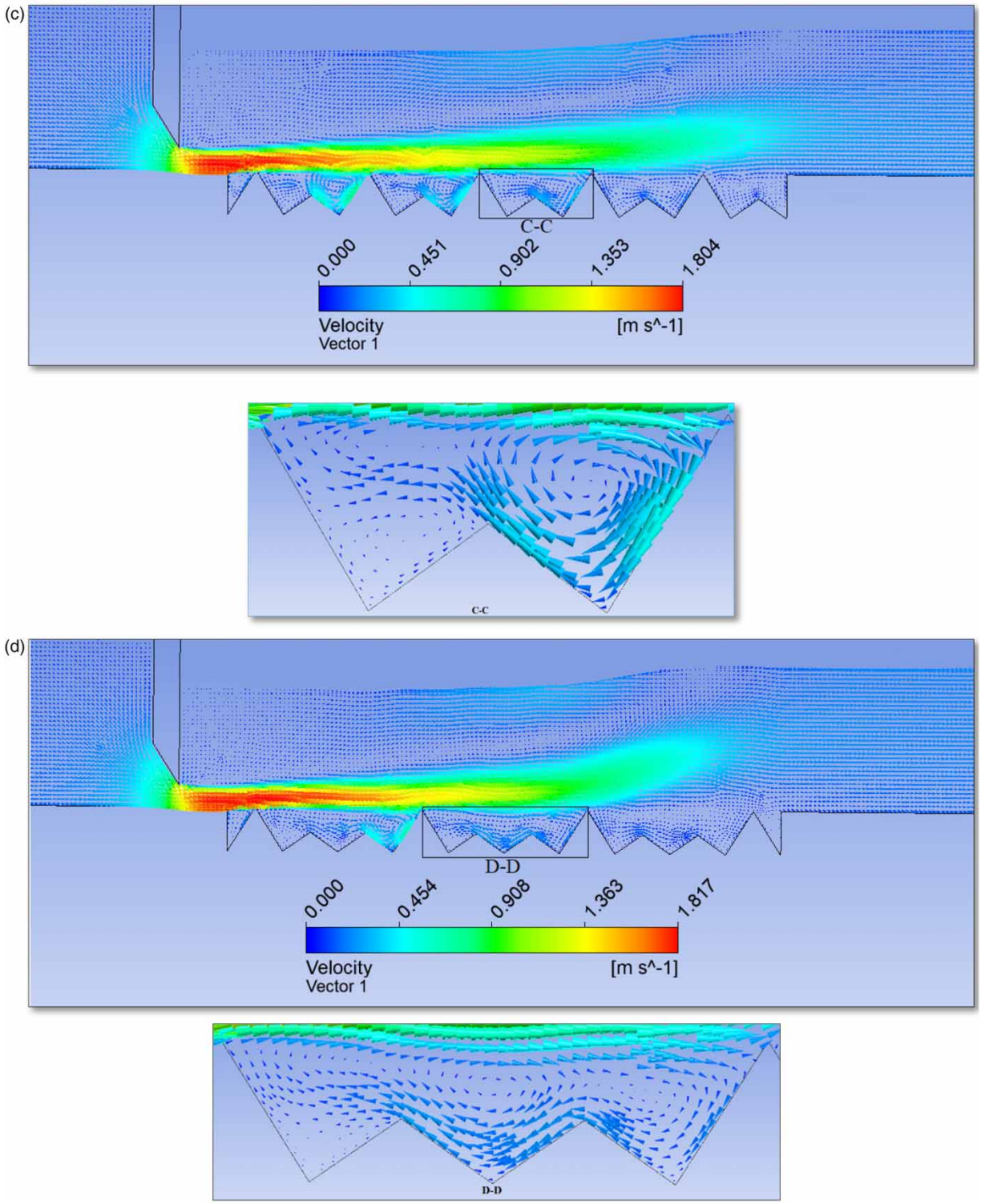


Figure 10 | Continued.

Longitudinal velocity vector variation over rough beds

Figure 10 shows the velocity vector variation on macrorough beds of the submerged hydraulic jump. From the results of Figure 10, it can be seen that the flow in the cavity region is significant in reducing excess kinetic energy. The direction of

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

circulation of velocity vectors depends on the arrangement of the macrorough bed. It was observed that the hydraulic jump over the macrorough bed has a clockwise circulation for T_1 and T_2 arrangements, whereas a combination of clockwise and anti-clockwise circulation was observed for T_3 and T_4 arrangements. The formation of a large circulation in the vicinity of the cavity region under the fully developed portion may be the reason of reduction in excess kinetic energy.

In addition, it is interesting to note that the velocity in the cavity region between two successive macroroughness elements was quite lesser than the mean velocity, which suggests that the flow in this region behaves differently from the rest of the flow. Finally, it seems that the presence of the cavity region between two successive elements may affect the behaviour of the flow, especially for higher initial Froude numbers.

Turbulent kinetic energy

The TKE can be obtained by averaging the velocities in all three directions, i.e., u_x, u_y, and u_z. Figure 11 shows the TKE over different sections of the flume for different macroroughness arrangements in comparison to the smooth bed. It can be

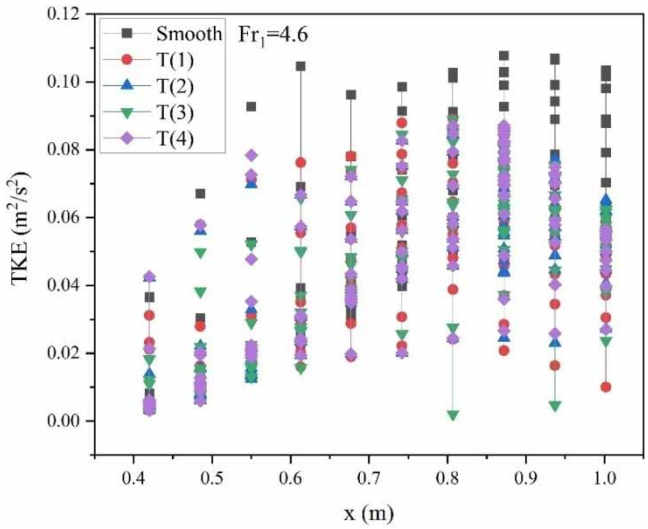


Figure 11 | Turbulent kinetic energy distribution at different sections of flume over different roughness configurations within the submerged hydraulic jump.

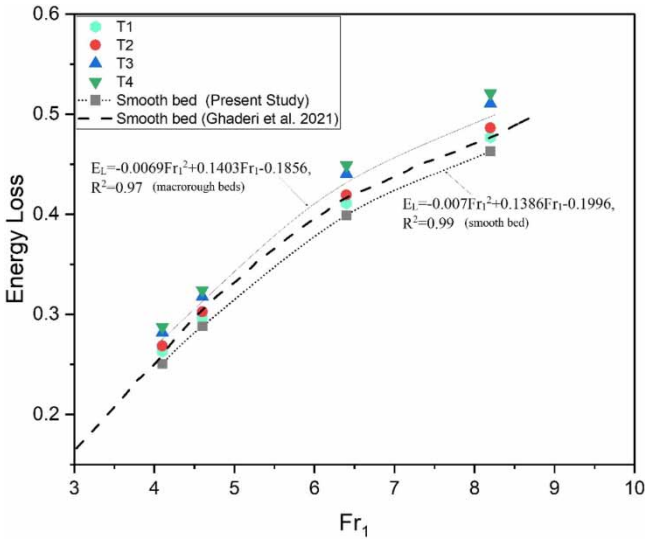


Figure 12 | Energy loss of submerged hydraulic jump over smooth and macroroughness bed versus Fr_1.

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

observed from Figure 11 that TKE decreases over the section, and it is evident that the TKE is maximum in the case of smooth bed and decreases with the provision of macroroughness to the bed.

Energy loss

Ghaderi *et al.* (2020, 2021) found that the energy loss over macroroughness was much higher than the smooth bed. Ghaderi *et al.* (2020) stated that a free jump dissipates more energy compared to the submerged hydraulic jump over macroroughness. In the present study, kinetic energy at a particular section is calculated using Equation (12), and the relative energy loss through the submerged hydraulic jump is calculated using Equation (13).

$$E_i = y_i + \frac{u_i^2}{2g} \tag{12}$$

$$E_L = \frac{E_3 - E_4}{E_3} \tag{13}$$

Here, the longitudinal sections within the jump were indicated by the subscript 'i' (i.e., $i = 0, 1, 2, 3, 4$). E_L represents relative energy loss through submerged hydraulic jump, whereas E_3 and E_4 are the specific energies calculated at the upstream and downstream ends of the submerged hydraulic jump respectively. Figure 12 depicts the variation of energy loss over different channel bed conditions.

$$E_L = -0.007\ \mathrm{Fr}_1^2 + 0.1386\ \mathrm{Fr}_1 - 0.1996, \quad R^2 = 0.99 \quad \text{(smooth bed)} \tag{14}$$

$$E_L = -0.0069\ \mathrm{Fr}_1^2 + 0.1403\ \mathrm{Fr}_1 - 0.1856, \quad R^2 = 0.97 \quad \text{(macrorough beds)} \tag{15}$$

Figure 12 shows that macrorough beds are dissipating more energy compared to a smooth bed. Equations for energy loss over smooth and macrorough beds are derived and written in Equations (14) and (15).

CONCLUSION

Overall, the research aimed to enhance the understanding of submerged hydraulic jumps by exploring various flow characteristics such as tailwater depth ratio, submerged depth ratio, including longitudinal velocity profile, and flow pattern in the cavity region over different shapes of macroroughness with their unique arrangements. Based on the numerical investigation, the key findings of the present study are presented as follows:

1. ANSYS Fluent, popular CFD software, is capable of predicting complex phenomena such as submerged hydraulic jump over smooth and macrorough beds.
2. The trend of velocity profiles obtained for submerged hydraulic jumps over macrorough beds is in good agreement with past studies.
3. It is reported that the cavity region between two successive macrorough elements leads to a decrease in both the tailwater and submerged depth ratios, as well as an increase in the energy dissipation of the submerged jump.
4. Clockwise and anti-clockwise circulation of the velocity vector is observed in between the cavity region of macroroughness elements in the case of submerged hydraulic jumps, which may be the reason for the reduction in tailwater depth ratio, submerged depth ratio, and longitudinal velocity.
5. In the cavity region between two successive macroroughness elements, it is found that the velocity of the fluid is quite less than the mean velocity.
6. The right-angle and isosceles triangular macroroughness have almost similar jump characteristics in the case of a submerged hydraulic jump.
7. By varying the arrangements of macroroughness elements, better energy dissipation can be achieved in the case of a submerged hydraulic jump.
8. For the basic parameters of submerged hydraulic jump, such as tailwater depth ratio and submerged depth ratio, the mean error between numerically derived results and experimental findings is determined to be less than 6%.

As a final remark for future research, optimization of possible configurations with their unique arrangements may be investigated. Furthermore, this numerical study may help hydraulic engineers in designing and modelling of energy dissipators.

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

ACKNOWLEDGEMENTS

The authors want to express their sincere gratitude to the anonymous reviewers for contributing their time to this article.

FUNDING

Delhi Technological University has provided the laboratory for the experiments and the ANSYS 2022 R1 license version for the numerical simulations.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abbaspour, A., Farsadizadeh, D. & Ghorbani, M. A. 2013 Estimation of hydraulic jump on corrugated bed using artificial neural networks and genetic programming. *Water Sci. Eng.* **6**, 189–198. https://doi.org/10.3882/j.issn.1674-2370.2013.02.007.

Abbaspour, A., Parvini, S. & Hosseinzadeh Dalir, A. 2016 Effect of buried plates on scour profiles downstream of hydraulic jump in open channels with horizontal and reverse bed slopes. *Water Sci. Eng.* **9**, 329–335. https://doi.org/10.1016/j.wse.2017.01.003.

Afzal, N., Bushra, A. & Seena, A. 2011 Analysis of turbulent hydraulic jump over a transitional rough bed of a rectangular channel: Universal relations. *J. Eng. Mech.* **137**, 835–845. https://doi.org/10.1061/(asce)em.1943-7889.0000294.

Ahmed, H. M. A., El Gendy, M. & Mirdan, A. M. H. *et al.* 2014 Effect of corrugated beds on characteristics of submerged hydraulic jump. *Ain Shams Eng. J.* **5**, 1033–1042. https://doi.org/10.1016/j.asej.2014.06.006.

Al-Hashimi, S. A. M., Saeed, K. A. & Nahi, T. N. 2019 Experimental and CFD modeling of hydraulic jumps forming at submerged weir. *J. Inst. Eng. Ser. A* **100**, 487–493. https://doi.org/10.1007/s40030-019-00381-z.

Bayon-Barrachina, A. & Lopez-Jimenez, P. A. 2015 Numerical analysis of hydraulic jumps using OpenFOAM. *J. Hydroinf.* **17**, 662–678. https://doi.org/10.2166/hydro.2015.041.

Canonsburg, T. D. 2013 *ANSYS Fluent Theory Guide*. ANSYS Inc, USA 15317, p. 814.

Celik, I. B., Ghia, U. & Roache, P. J. 2008 Procedure for estimation and reporting of uncertainty due to discretization in CFD applications. *J. Fluids Eng.* **130** (7), 0780011–0780013.

Chen, J. S., Chen, W. G., Li, J. & Sun, P. 2018 A generalized model for wind turbine faulty condition detection using combination prediction approach and information entropy. *J. Enviro. Inf.* **32**, 14–24. https://doi.org/10.3808/jei.201800393.

Choufu, L., Abbasi, S. & Pourshahbaz, H. *et al.* 2019 Investigation of flow, erosion, and sedimentation pattern around varied groynes under different hydraulic and geometric conditions: A numerical study. *Water (Switzerland)* **11**. https://doi.org/10.3390/w11020235.

Dasineh, M., Ghaderi, A. & Bagherzadeh, M. *et al.* 2021 Prediction of hydraulic jumps on a triangular bed roughness using numerical modeling and soft computing methods. *Mathematics* **9**. https://doi.org/10.3390/math9233135.

Ead, S. A. & Rajaratnam, N. 2002 Hydraulic jumps on corrugated beds. *J. Hydraul. Eng.* **128**, 656–663. https://doi.org/10.1061/(asce)0733-9429(2002)128:7(656).

Elsebaie, I. H. & Shabayek, S. 2010 Formation of hydraulic jumps on corrugated beds. 40–51.

Eshkou, Z., Dehghani, A. A. & Ahmadi, A. 2018 Forced hydraulic jump in a diverging stilling basin using angled baffle blocks. *J. Irrig. Drain. Eng.* **144**, 06018004. https://doi.org/10.1061/(asce)ir.1943-4774.0001328.

Felder, S. & Chanson, H. 2018 Air–water flow patterns of hydraulic jumps on uniform beds macroroughness. *J. Hydraul. Eng.* **144**, 04017068. https://doi.org/10.1061/(asce)hy.1943-7900.0001402.

Ghaderi, A., Dasineh, M., Aristodemo, F. & Ghahramanzadeh, A. 2020 Characteristics of free and submerged hydraulic jumps over different macroroughnesses. *J. Hydroinf.* **22**, 1554–1572. https://doi.org/10.2166/HYDRO.2020.298.

Ghaderi, A., Dasineh, M., Aristodemo, F. & Aricò, C. 2021 Numerical simulations of the flow field of a submerged hydraulic jump over triangular macroroughnesses. *Water (Switzerland)* **13**. https://doi.org/10.3390/w13050674.

Guo, Y., Wang, Q. & Zhang, D. *et al.* 2018 A stochastic-process-based method for assessing frequency regulation ability of power systems with wind power fluctuations. *J. Enviro. Inf.* **32**, 45–54. https://doi.org/10.3808/jei.201800394.

Hafnaoui, M. A., Carvalho, R. F. & Debabeche, M. 2016 Prediction of hydraulic jump location in some types of prismatic channels using numerical prediction of hydraulic jump location in some types of prismatic channels using numerical modelling. https://doi.org/10.15142/T3D01F.

Hager, W. H. & Bremen, R. 1989 Sequent depths: Le ressaut hydraulique classique: Étude des hauteurs conjuguées. *J. Hydraul. Res.* **27**, 565–585. https://doi.org/10.1080/00221688909499111.

Hager, W. H. & Bretz, N. V. 1986 Ressauts hydrauliques par-dessus des marches positives et négatives. *J. Hydraul. Res.* **24**, 237–253. https://doi.org/10.1080/00221688609499303.

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

Hager, W. H., Bremen, R. & Kawagoshi, N. 1990 Length of roller Ressaut hydraulique classique: Longueur du rouleau. *J. Hydraul. Res.* **28**, 591–608. https://doi.org/10.1080/00221689009499048.

Hien, L. T. T. & Duc, D. H. 2020 Numerical simulation of free surface flow on spillways and channel chutes with wall and step abutments by coupling turbulence and air entrainment models. *Water (Switzerland)* **12**. https://doi.org/10.3390/w12113036.

Hughes, W. C. & Flack, J. E. 1984 Hydraulic jump properties over a rough bed. *J. Hydraul. Eng.* **110**, 1755–1771. https://doi.org/10.1061/(asce)0733-9429(1984)110:12(1755).

Jesudhas, V., Balachandar, R., Roussinova, V. & Barron, R. 2018 Turbulence characteristics of classical hydraulic jump using DES. *J. Hydraul. Eng.* **144**, 04018022. https://doi.org/10.1061/(asce)hy.1943-7900.0001427.

Kumar, M., Kumar, S. & Bidhu, S. 2019 Determination of sequent depth of hydraulic jump over sloping floor with rounded and crushed aggregates using experimental and ANN model. *Water Sci. Technol. Water Supply* **19**, 2240–2247. https://doi.org/10.2166/ws.2019.104.

Li, W., Bao, Z., Huang, G. H. & Xie, Y. L. 2018 An inexact credibility chance-constrained integer programming for greenhouse gas mitigation management in regional electric power system under uncertainty. *J. Environ. Inf.* **31**, 111–122. https://doi.org/10.3808/jei.201500326.

Lin, C., Ma, R. H. & Wu, Z. P. *et al.* 2018 Detection of the sensitive inflowing river indicators related to non-point source organic pollution: A case study of Taihu lake. *J. Environ. Inf.* **32**, 98–111. https://doi.org/10.3808/jei.201600343.

Macián-Pérez, J. F., Bayón, A. & García-Bartual, R. *et al.* 2020a Characterization of structural properties in high reynolds hydraulic jump based on CFD and physical modeling approaches. *J. Hydraul. Eng.* **146**, 04020079. https://doi.org/10.1061/(asce)hy.1943-7900.0001820.

Macián-Pérez, J. F., Vallés-Morán, F. J. & Sánchez-Gómez, S. *et al.* 2020b Experimental characterization of the hydraulic jump profile and velocity distribution in a stilling basin physical model. *Water (Switzerland)* **12**. https://doi.org/10.3390/w12061758.

Mahtabi, G., Chaplot, B., Azamathulla, H. M. & Pal, M. 2020 Classification of hydraulic jump in rough beds. *Water (Switzerland)* **12**, 1–11. https://doi.org/10.3390/w12082249.

Maleki, S. & Fiorotto, V. 2021 Hydraulic jump stilling basin design over rough beds. *J. Hydraul. Eng.* **147**, 04020087. https://doi.org/10.1061/(asce)hy.1943-7900.0001824.

Mousavi, S. N., Steinke, R. & Teixeira, E. D. *et al.* 2020 Predictive modeling the free hydraulic jumps pressure through advanced statistical methods. *Mathematics* **8**. https://doi.org/10.3390/math8030323.

Nikmehr, S. & Aminpour, Y. 2020 Numerical simulation of hydraulic jump over rough beds. *Period Polytech. Civ. Eng.* **64**, 396–407. https://doi.org/10.3311/PPci.15292.

Rajaratnam, N. 1967 *Hydraulic Jumps*. Academic Press Inc.

Salmasi, F. & Samadi, A. 2018 Experimental and numerical simulation of flow over stepped spillways. *Appl. Water Sci.* **8**, 1–11. https://doi.org/10.1007/s13201-018-0877-5.

Samadi-Boroujeni, H., Ghazali, M., Gorbani, B. & Fattahi Nafchi, R. 2013 Effect of triangular corrugated beds on the hydraulic jump characteristics. *Can. J. Civ. Eng.* **40**, 841–847. https://doi.org/10.1139/cjce-2012-0019.

Sauida, M. F. 2016 Prediction of hydraulic jump length downstream of multi-vent regulators using artificial neural networks. *Ain Shams Eng. J.* **7**, 819–826. https://doi.org/10.1016/j.asej.2015.12.005.

Tokyay, N. D. & Velioğlu, D. 2012 Effects of bed roughness on the characteristics of hydraulic jumps. 17–19.

Wang, H. & Chanson, H. 2016 Self-similarity and scale effects in physical modelling of hydraulic jump roller dynamics, air entrainment and turbulent scales. *Environ. Fluid Mech.* **16**, 1087–1110. https://doi.org/10.1007/s10652-016-9466-z.

Yamini, O. A., Mousavi, S. H., Kavianpour, M. R. & Ghaleh, R. S. 2021 Hydrodynamic performance and cavitation analysis in bottom outlets of dam using CFD modelling. *Adv. Civ. Eng.* **2021**. https://doi.org/10.1155/2021/5529792.

Yildiz, A., Marti, A. I., Yarar, A. & Yilmaz, V. 2020 Determination of position of hydraulic jump in a flume by using CFD and comparison with experiential results. *Rom. J. Ecol. Environ. Chem.* **2**, 78–85. https://doi.org/10.21698/rjeec.2020.211.

First received 17 April 2023; accepted in revised form 20 November 2023. Available online 28 November 2023

Downloaded from http://iwaponline.com/jh/article-pdf/doi/10.2166/hydro.2023.302/1330025/jh2023302.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 September 2023, accepted 29 October 2023, date of publication 6 November 2023, date of current version 9 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3330429

RESEARCH ARTICLE

On a Novel Model for ECG Signals and Its Statistical Properties

AYAN KUMAR DUTTA[1], SHIV DUTT JOSHI[1], AND BREJESH LALL[1,2], (Member, IEEE)
[1]Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi 110016, India
[2]Bharti School of Telecommunication Technology and Management, Indian Institute of Technology Delhi, New Delhi 110016, India

Corresponding author: Ayan Kumar Dutta (Ayan.Kumar.Dutta@ee.iitd.ac.in)

This work was supported by the Ministry of Education (MoE), formerly the Ministry of Human Resource Development (1985–2020), Government of India.

ABSTRACT An accurate mathematical ECG model helps comprehend the heart's workings, which in turn helps identify various heart-related abnormalities. A typical ECG waveform consists of recurrent QRS complexes at almost regular intervals. In the most general scenario, the amplitude, the period after which it repeats, and the shape of the QRS complex may change from pulse to pulse. This paper proposes a mathematical model for the ECG signals capable of capturing these features. Since statistical analysis of the the general model is more complex, we make certain assumptions to make the problem analytically tractable. With these assumptions at our disposal, we provide closed-form expressions for time-varying covariance/spectra of the process generated by the proposed model. In order to demonstrate the approximation capability of the model, that is, the ability to model ECG signals, we synthesize ECG signals for different health states by first extracting the QRS complex of that health state and using the mean time period also extracted from the realization of the same health state. We find that the degree of non-stationarity of the real-world ECG and the one generated by the model match closely. In reality, the signal for each state appears to be a close match. Based on our experimental findings, we observe that the measure of the degree of non-stationarity interestingly falls in different ranges, depending upon the type of heart-related abnormality. This observation makes our conjecture that this measure can serve as a biomarker for various related abnormalities.

INDEX TERMS Electrocardiogram, almost-cyclostationary, time-varying covariance, time-varying spectrum.

I. INTRODUCTION

An electrocardiogram (ECG or EKG) is a non-invasive diagnostic procedure used to identify issues connected to the heart. This examination is entirely risk-free and will not adversely affect the patient's health. The ECG is the biological signal most frequently used for diagnosing cardiac conditions. ECG is obtained by placing electrodes in precise areas on the patient's chest, arms, and legs. Lead wires are used to make the connection between the electrodes and the electrocardiogram equipment. After that, the electrical activity of the heart is monitored and then evaluated in order to determine any abnormalities that may be present in the patient's heart. The temporal differences between individual waves and their duration, amplitude levels, polarity, and shape carry essential clinical information that can be extracted from them. Not only are these qualities relevant, but their development over time, their fluctuations from beat to beat or the long term, their responses to alterations in heart rate, and their interactions may also have substantial diagnostic as well as therapeutic implications. Important clinical information is obtained from the changes in wave duration, amplitude levels, polarity, and shape across different time intervals [10]. If we have an accurate mathematical model, we will be able to understand better and identify various heart disorders. This work is an

The associate editor coordinating the review of this manuscript and approving it for publication was Ganesh Naik.

© 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/

effort in that direction. A mathematical model that faithfully reproduces the actual signal of an electrocardiogram (ECG) can provide a more in-depth understanding of how the heart works and can also be used to evaluate the heart's overall health.

In the literature, investigations have been carried out to identify particular rhythm abnormalities that typically come before significant arrhythmias by mathematical modeling. The authors of [1] propose a Markov model of cardiac R-wave rhythm. They used Wald's sequential test to categorize heart rhythms without assigning a priori probabilities. The authors in [2], address the inverse problem in a novel electric multipole expansion (EME) cardiac model. The EME model estimates the potentials many cardiac models produce without solving the forward problem for each cardiac model. The authors of [3] analyzed typical ECG signals using the well-known homomorphic analysis and modeling approaches [12], [13]. The authors in [4] propose an approach for screening obstructive sleep apnea (OSA) that is based on the recording of an electrocardiogram (ECG) while the patient is asleep. Another study [5] offers a 2D deep convolutional neural network ECG arrhythmia classification algorithm (CNN). ECG time domain data from five heartbeat types-NOR, LBB, RBB, PVC, and APC-were first transformed into time-frequency spectrograms by short-time Fourier transform. The 2D-CNN recognizes and categorizes ECG arrhythmia types using the five arrhythmia kinds' spectrograms. The authors in [6] construct a depression detection model using task-state heart rate variability (HRV) parameters and examined how therapy affects them. Using several electric dipoles and a whole model of the human body, a new and straightforward ECG generation model is proposed in [7]. The authors in [8] provide a novel hidden Markov model (HMM) method for real-time beat segmentation and ECG categorization. The HMM framework is found to be extremely useful for electrocardiogram (ECG) beat detection, segmentation, and classification. The authors in [9] propose a signal-dependent orthogonal transform and a parametric representation of the transform coefficients vector to model ECGs. In [14], the authors utilize deep learning techniques to predict the four significant cardiac abnormalities: abnormal heartbeat, myocardial infarction, history of myocardial infarction, and normal person classes.

Despite the substantial quantity of research, the limitations of the research mentioned above must be addressed. For example, Wald's standard sequential probability ratio test used in [1] depends on the alternative hypothesis's relative risk, which may delay or miss a vital signal. A composite alternative hypothesis-based maximum sequential probability ratio test (MaxSPRT) is a superior choice [20]. Again, observe that, due to incorrect modeling of small patches of electrical activity during the early and late periods of QRS and differences between Durrer et al. surface potential maps and typical empirically obtained maps, [2] cannot construct a realistic discrete dipole cardiac model. The study described in [3] had certain limitations, including the absence of models of higher-order ECG signals with abnormalities to find an exact match between the signal and its model. Additionally, no classification has been provided for the abnormalities observed. In addition, modeling was not performed on large datasets. The study described in [4] was limited to sleep apnea detection, and the classification was not based on a beat-by-beat basis, which may not be appropriate for a real-time apnea detection system. In [8], there is no quantitative analysis of the effect of various wavelets on system performance. In [6] and [7] there is no classification of heart-related abnormalities. In [21], the author proposed that the ECG signal model is oscillatory, almost cyclostationary, and works with much greater observation intervals than the traditional cyclostationary model. However we should note that due to the relatively large number of normal beats used in the average, a transient phenomenon such as a brief string of anomalous beats may go undetected if the observation interval is too long. Due to the numerous limitations identified in the research conducted to date, it would be more beneficial to develop an accurate ECG model for a deeper comprehension of the heart and as a byproduct of improved classification. This manuscript models the ECG as the sum of quasi-periodically shifted waveforms with infinitesimally small random fluctuations at each pulse interval.

The authors of [22] and [23] corroborate that the cyclostationary model can be used to model ECG signals. In order to carry out any processing, samples of continuous-time signals are taken, indeed meeting the minimum sampling rate requirement. It is well known that the sampled version of a continuous-time signal need not be periodic in discrete-time cases. Such a signal, though, may appear almost periodic. Motivated by this, for a random process, if the sampling interval is not a multiple of the mean period of a continuous-time cyclostationary process, then a sampled discrete-time process is said to be almost cyclostationary, abbreviated as ACS [24], although the ACS process may have origin different from this too. The authors in [25] and [26] demonstrates that the discrete-time signal, comprised of samples from a generalized ACS signal, can be classified as a discrete-time ACS signal. Consequently, in this context, we also present the sampled counterpart of the continuous-time cyclic covariance and cyclic spectrum of our proposed electrocardiogram model.

The main contributions of the paper are as follows:

1) We model the ECG as the sum of quasi-periodically shifted waveforms with small random fluctuations at each pulse interval.
2) We provide expressions for the time-varying covariance and time-varying spectrum for the signal generated from the proposed model. We then use these expressions to obtain an expression to measure the degree of non-stationarity ($\mathcal{MDNST}$) proposed earlier in [15].

3) We empirically obtain the optimum value of Δ, a parameter used in the model to characterize delay fluctuations around our model's mean value.
4) We obtain $\mathcal{MDNST}$ ranges for diverse heart abnormalities and observe that different heart conditions result in distinct $\mathcal{MDNST}$ ranges. Therefore, we conjecture that these could serve as biomarkers for various associated abnormalities.

The paper is organized as follows: Section II gives prerequisites for the subsequent tasks. In Section III, we propose a model for the electrocardiogram signal. Here we derive closed-form expressions for time-varying covariance, time-varying spectra, and their coefficients, namely cyclic covariance and cyclic spectrum. Sampled version of the proposed electrocardiogram model is given in Section IV. In Section V and Section VI, respectively, we describe about the relationship between continuous and discrete mean components and the relationship between continuous and discrete covariance components in more detail. In Section VII, the cyclic spectrum of the proposed model is presented as a theorem. Section VIII provides the measure for the degree of non-stationarity for our proposed model. We then present our results, simulations and discussions in Section IX. We conclude in Section X.

II. NOTATIONS AND PRELIMINARIES

Let $\mathcal{Y}(t)$ be a random process. The first and second-order statistics of such a random process are given by its mean function:

$$\mu_{\mathcal{Y}}(t) \triangleq \mathcal{E}(\mathcal{Y}(t)), -\infty < t < \infty, \tag{1}$$

and its covariance function

$$\mathcal{C}_{\mathcal{Y}\mathcal{Y}}(t,\mathcal{T}) \triangleq \mathcal{E}\left(\widetilde{\mathcal{Y}}(t)\,\widetilde{\mathcal{Y}}^{*}(t-\mathcal{T})\right), -\infty < t,\mathcal{T} < \infty, \tag{2}$$

where, $\widetilde{\mathcal{Y}}(t) \triangleq \mathcal{Y}(t) - \mu_{\mathcal{Y}}(t)$, and $*$ represents the complex conjugate.

Definition 1: A process is called wide sense almost-cyclostationary (WS-ACS) [19] if its covariance function $\mathcal{C}_{\mathcal{Y}\mathcal{Y}}(t,\mathcal{T})$ admits the trigonometric representation given below:

$$\mathcal{C}_{\mathcal{Y}\mathcal{Y}}(t,\mathcal{T}) = \sum_{\alpha\in\mathbb{S}_{\mathcal{T}}} C_{\mathcal{Y}\mathcal{Y}}(\alpha,\mathcal{T}) \exp(j\alpha t), \tag{3}$$

where,

$$\begin{aligned} &C_{\mathcal{Y}\mathcal{Y}}(\alpha,\mathcal{T}) \\ &\quad\triangleq \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2}\mathcal{C}_{\mathcal{Y}\mathcal{Y}}(t,\mathcal{T})\exp(-j\alpha t)\,dt \\ &\quad\triangleq \left\langle \mathcal{C}_{\mathcal{Y}\mathcal{Y}}(t,\mathcal{T})\exp(-j\alpha t)\right\rangle_t, \end{aligned} \tag{4}$$

is the Fourier series coefficient at cycle frequency α and $\mathbb{S}_{\mathcal{T}}$ is the countable set of frequencies α, which in general could be incommensurate, i.e.,

$$\mathbb{S}_{\mathcal{T}} = \left\{\alpha\in\mathbb{R} | C_{\mathcal{Y}\mathcal{Y}}(\alpha,\mathcal{T}) \neq 0\right\}, \tag{5}$$

Note that, $\langle\cdot\rangle_t \triangleq \lim_{T_0\to\infty}\frac{1}{T_0}\int_{-T_0/2}^{T_0/2}(\cdot)\,dt$, represents the time averaging operation. Notably, if the frequencies α's are harmonically related, the process is wide sense cyclostationary (WSCS). The coefficients $C_{\mathcal{Y}\mathcal{Y}}(\alpha,\mathcal{T})$ are called cyclic covariances.

Definition 2: A sampled continuous-time random process $\mathcal{Y}_{\text{sam}}$ is wide sense almost-cyclostationary if its covariance function $\mathcal{C}_{\mathcal{Y}_{\text{sam}}\mathcal{Y}_{\text{sam}}}(n,r)$ admits the trigonometric representation given below:

$$\begin{aligned} \mathcal{C}_{\mathcal{Y}_{\text{sam}}\mathcal{Y}_{\text{sam}}}(n,r) &= \mathcal{E}\left(\mathcal{Y}_{\text{sam}}(n)\,\mathcal{Y}^{*}_{\text{sam}}(n-r)\right) \\ &= \sum_{\alpha_{\text{sam}}\in\mathbb{S}_r} C_{\mathcal{Y}_{\text{sam}}\mathcal{Y}_{\text{sam}}}(\alpha_{\text{sam}},r)\,e^{j\alpha_{\text{sam}}n}, \end{aligned} \tag{6}$$

where, $\mathbb{S}_r = \left\{0 \le \alpha_{\text{sam}} < 2\pi | C_{\mathcal{Y}_{\text{sam}}\mathcal{Y}_{\text{sam}}}(\alpha_{\text{sam}},r) \neq 0\right\}$, and,

$$\begin{aligned} C_{\mathcal{Y}_{\text{sam}}\mathcal{Y}_{\text{sam}}}(\alpha_{\text{sam}},r) \triangleq \lim_{N\to\infty}\frac{1}{2N+1}\sum_{n=-N}^{N} \\ \mathcal{C}_{\mathcal{Y}_{\text{sam}}\mathcal{Y}_{\text{sam}}}(n,r)\times\exp(-j\alpha_{\text{sam}}n). \end{aligned} \tag{7}$$

Observe that when α_{sam}'s are harmonically related, then (6) represents a WSCS process.

III. MODEL FOR THE ELECTROCARDIOGRAM SIGNAL

In this section, we present our proposed model for ECG signals. Let us look into the typical ECG signal (Fig. 1). We take note of the following: There is a recurring ripple (pulse) $p^{C}(t)$, known as QRS complex. The distance between two ripples is almost same. If we denote the mean value of the time interval between successive ripples as T, it is clear that the time interval between kth ripple and $(k+1)$ th ripple, where $k \in \mathbb{Z}$ would be T plus some zero mean fluctuation. We denote zero mean fluctuation for the kth QRS complex as β_k. So the kth ripple would appear at the instant $kT + \beta_k$.

The idea is as follows: Consider the continuous-time random process $\mathcal{X}^{C}(t) = \sum_{k\in\mathbb{Z}}\mathcal{A}_k p_k^{C}(t)$, where $p_k^{C}(t)$ is the QRS complex appearing at the kth instance (with respect to the zeroth QRS complex $p_0^{C}(t)$, that appears at the origin). In general, the QRS complex shape can change from instance to instance, we here, for the sake of analytical tractability, assume that kth instance QRS complex $p_k^{C}(t)$ is just a shifted version of the first QRS complex $p_0^{C}(t)$ (taken as reference), appearing at the origin. Here the superscript **C** denotes the continuous-time signal. Using the aforementioned assumptions, the following representation is proposed for a realization of the electrocardiogram signal:

$$\mathcal{X}^{C}(t) = \sum_{k\in\mathbb{Z}}\mathcal{A}_k p_k^{C}(t) = \sum_{k\in\mathbb{Z}}\mathcal{A}_k p^{C}(t-[kT+\beta_k]). \tag{8}$$

The schematic diagram of (8) is shown in Fig. 1. The average heart rate is denoted as $\alpha \triangleq 2\pi\times\frac{1}{T}$, and sampling period is denoted as $T_s \triangleq \frac{1}{f_s} = \frac{T_0}{N}$, where, T_0 is the observation interval and N denotes the number of samples taken.

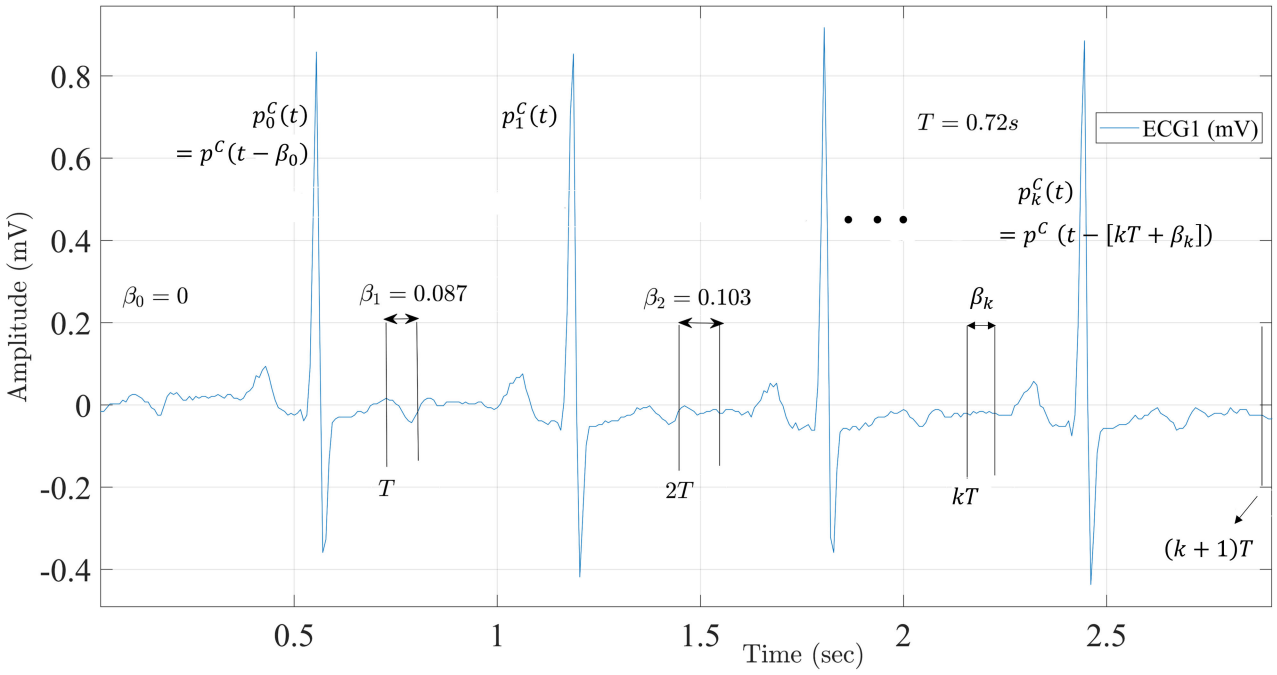


FIGURE 1. A schematic representation of our proposed model experimented upon a Normal sinus rhythm.

A. TIME-VARYING MEAN OF THE PROPOSED MODEL

With the proposed model (eqn. (8)) at our disposal, we now study the statistical features of the signal generated by the proposed model. Later, we will determine these statistical features for actual ECG signals and comment on the similarity between the two.

The mathematical analysis of the general model, as can be seen, would be highly complex. Hence, in order to make the problem mathematically tractable, we make certain assumptions. The analysis has been carried out under the following assumptions:

1) The coefficients $\mathcal{A}_k'$s, which in the general case could be time-varying and also dependent on each other, we assume all the coefficients to be constant value $'1'$ (and also the same).
2) The fluctuations about the mean period, β_k's, which could generally be dependent, with time-varying statistics, are assumed to be independent and identically distributed for analytical tractability. We assume them to be zero mean and uniformly distributed.
3) The QRS complexes also, in general, could vary in shape and size. We assume, again, for mathematical simplicity, the QRS complexes to be translated versions of one prototype signal.

Let β_k's have the following (uniform) distribution given as,

$$f_\beta(\xi) = \begin{cases} \frac{1}{\Delta} & \frac{-\Delta}{2} \le \xi \le \frac{\Delta}{2} \\ 0 & \text{elsewhere} \end{cases}.$$

The time-varying mean of the proposed model (represented by (8)) is given as:

$$\mu_{\mathcal{X}^C}(t) \triangleq \mathcal{E}\left(\mathcal{X}^C(t)\right), -\infty < t < \infty, \tag{9}$$

$$\Rightarrow \mu_{\mathcal{X}^C}(t) = \mathcal{E}\left[\sum_{\widetilde{k}\in\mathbb{Z}} p^C\left(t - \left[\widetilde{k}T + \beta\right]\right)\right], \tag{10}$$

$$= \sum_{\widetilde{k}\in\mathbb{Z}} \int_{-\infty}^{\infty} p^C\left(\left[t - \widetilde{k}T\right] - \widetilde{\xi}\right) f_\beta\left(\widetilde{\xi}\right) d\widetilde{\xi}, \tag{11}$$

$$= \sum_{\widetilde{k}\in\mathbb{Z}} q\left(t - \widetilde{k}T\right), \tag{12}$$

which is clearly a periodic signal with period T, where, $q(t) = \left(p^C \star f_\beta\right)(t)$, and $\star$ represents the convolution operation. The probability density function of β, for simplicity, is taken to be uniformly distributed with support as $\frac{-\Delta}{2}$ to $\frac{\Delta}{2}$. An interesting observation that we made while dealing with real-world signals and the one generated by our model is that the optimum value of Δ remains invariant under different health conditions of the patient.

B. TIME-VARYING COVARIANCE OF THE PROPOSED MODEL

In this subsection, we derive expressions for the time-varying covariance function of the process generated from the proposed model with the simplifications mentioned above. Let us investigate the form of cyclic covariance, the coefficients occurring in (3), for the time-varying covariance

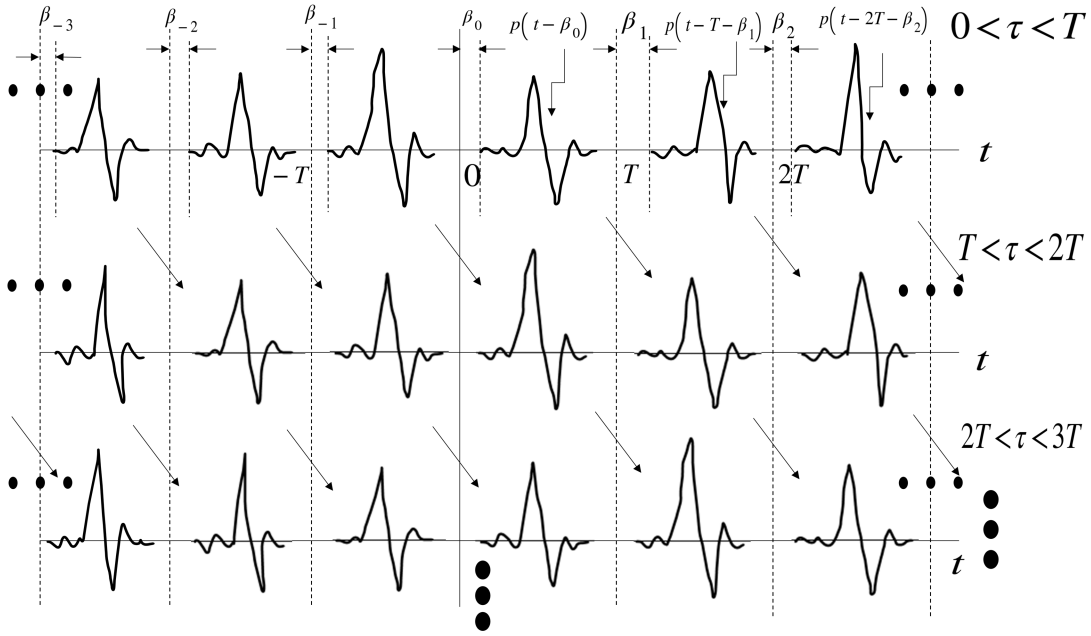


FIGURE 2. Schematic diagram for various ranges of $\mathcal{T}$.

for the process of the proposed model. Clearly, we have:

$$\mathcal{C}_{\mathcal{X}^C\mathcal{X}^C}(t,\mathcal{T}) \triangleq \mathcal{E}\left(\left[\mathcal{X}^C(t)-\mu_{\mathcal{X}^C}(t)\right]\left[\mathcal{X}^{C*}(t-\mathcal{T})-\mu^*_{\mathcal{X}^C}(t-\mathcal{T})\right]\right), \tag{13}$$

$\forall t$ and $\forall \mathcal{T}$. Observe that (13) reduces to:

$$\mathcal{C}_{\mathcal{X}^C\mathcal{X}^C}(t,\mathcal{T}) = \mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(t,\mathcal{T}) - \mathcal{M}_{\mathcal{X}^C}(t,\mathcal{T}), \tag{14}$$

where, we have, $\mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(t,\mathcal{T}) \triangleq \mathcal{E}\left[\mathcal{X}^C(t)\,\mathcal{X}^{C*}(t-\mathcal{T})\right]$ and $\mathcal{M}_{\mathcal{X}^C}(t,\mathcal{T}) \triangleq \mu_{\mathcal{X}^C}(t)\,\mu^*_{\mathcal{X}^C}(t-\mathcal{T})$. The covariance function, $\mathcal{C}_{\mathcal{X}^C\mathcal{X}^C}(t,\mathcal{T})$ of (14), has two components one due to time-varying correlation $\mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(t,\mathcal{T})$ and the other related to the time-varying mean $\mathcal{M}_{\mathcal{X}^C}(t,\mathcal{T})$. We will call the first component as correlation component and second as mean component for convenience. We assume the trigonometric representation of $\mathcal{C}_{\mathcal{X}^C\mathcal{X}^C}(t,\mathcal{T})$ for every $\mathcal{T}$ as: $\mathcal{C}_{\mathcal{X}^C\mathcal{X}^C}(t,\mathcal{T}) \triangleq \sum_{\alpha\in\mathbb{S}_\mathcal{T}} C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T})\,e^{j\alpha t} - \sum_{\alpha\in\mathbb{S}_\mathcal{T}} M_{\mathcal{X}^C}(\alpha,\mathcal{T})\,e^{j\alpha t}$, where,

$$\mathbb{S}_\mathcal{T} \triangleq \left\{\alpha\in\mathbb{R} | C_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T})\neq 0\right\}. \tag{15}$$

The coefficients of the cyclic covariance are given as:

$$C_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T}) = \left\langle \mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(t,\mathcal{T})\,e^{-j\alpha t}\right\rangle_t - \left\langle \mathcal{M}_{\mathcal{X}^C}(t,\mathcal{T})\,e^{-j\alpha t}\right\rangle_t, \tag{16}$$

$$= C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T}) - M_{\mathcal{X}^C}(\alpha,\mathcal{T}). \tag{17}$$

Even though $\alpha\in\mathbb{S}_\mathcal{T}$, in the correlation component may differ from $\alpha\in\mathbb{S}_\mathcal{T}$ in the mean component. Nevertheless, given that both are countable, their union is also countable.

1) EVALUATION OF VARIOUS RANGES OF $\mathcal{T}$

Different ranges of delay $\mathcal{T}$ (Fig. 2) would result in different overlaps; therefore, each range must be evaluated separately.

2) FOR $0<\mathcal{T}<T$

The time-varying autocorrelation has the following form:

$$\begin{aligned}\mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(1)}(t,\mathcal{T}) &= \sum_{\substack{k_1,k_2\in\mathbb{Z}\\ k_1\neq k_2}} \mathcal{E}\left[p^C\left(t-\left[k_1T+\beta_{k_1}\right]\right)\right] \\ &\quad\times \mathcal{E}\left[p^{C*}\left(t-\mathcal{T}-\left[k_2T+\beta_{k_2}\right]\right)\right] \\ &\quad+\sum_{k_1\in\mathbb{Z}} \mathcal{E}\{p^C\left(t-\left[k_1T+\beta_{k_1}\right]\right) \\ &\quad\times p^{C*}\left(t-\mathcal{T}-\left[k_1T+\beta_{k_1}\right]\right)\}, \qquad (18)\\ &= \sum_{k\in\mathbb{Z}} \mathcal{E}\{p^C\left(t-[kT+\beta_k]\right) \\ &\quad\times p^{C*}\left(t-\mathcal{T}-[kT+\beta_k]\right)\} \\ &\quad+\text{other terms zero.} \qquad (19)\end{aligned}$$

Now assuming that the model would, in general, generate an almost cyclostationary process, we consider the following representation for its time-varying autocorrelation function:

$$\mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(1)}(t,\mathcal{T}) = \sum_{\alpha\in\mathbb{S}_\mathcal{T}} C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(1)}(\alpha,\mathcal{T})\,e^{j\alpha t}, \tag{20}$$

where, the cyclic autocorrelation is given by:

$$C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(1)}(\alpha,\mathcal{T}) = \left\langle \mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(1)}(t,\mathcal{T})\,e^{-j\alpha t}\right\rangle_t. \tag{21}$$

Substituting (19) in (21), we obtain,

$$C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(1)}(\alpha,\mathcal{T}) = \sum_{k\in\mathbb{Z}} \lim_{T_0\to\infty} \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} \int_{-\infty}^{\infty} f_\beta(\xi)\, d\xi \times p^C([t-kT]-\xi) \times p^{C*}([t-kT]-\mathcal{T}-\xi)\, e^{-j\alpha t} dt. \tag{22}$$

We define $s \triangleq t - kT - \xi$, and noting that, $\int_{-\infty}^{\infty} p^C(s)\, p^{C*}(s-\mathcal{T})\, e^{-j\alpha s} ds = \frac{[P(\alpha)\star e^{-j\alpha\mathcal{T}} P^*(-\alpha)]}{2\pi}$, we have, then, (22) reduces to:

$$C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(1)}(\alpha,\mathcal{T}) = F_\beta(\alpha) \times \frac{1}{2\pi}\left(P(\alpha)\star e^{-j\alpha\mathcal{T}} P^*(-\alpha)\right) \times \lim_{T_0\to\infty} \frac{1}{T_0} \sum_{k\in\mathbb{Z}} e^{-j\alpha kT}, \tag{23}$$

where, $F_\beta(\alpha) = \int_{-\infty}^{\infty} f_\beta(\xi)\, e^{-j\alpha\xi} d\xi$. Let $T_0 = 2NT$. Observe that, $\lim_{T_0\to\infty} \frac{1}{T_0} \sum_{k\in\mathbb{Z}} e^{-j\alpha kT} = \lim_{N\to\infty} \frac{1}{(2N+1)T} \times \frac{\sin\left(\frac{(\alpha T)(2N+1)}{2}\right)}{\sin\left(\frac{(\alpha T)}{2}\right)} = \delta_{\alpha,\frac{2m\pi}{T}}$, where, $\delta_{\alpha,\frac{2m\pi}{T}} = \begin{cases} 1 & \alpha = \frac{2m\pi}{T} \\ 0 & \text{else} \end{cases}$. Hence, (23) reduces to:

$$C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(1)}(\alpha,\mathcal{T}) = \begin{cases} \frac{F_\beta(\alpha)}{2\pi T}\left(P(\alpha)\star e^{-j\alpha\mathcal{T}} P^*(-\alpha)\right) \\ \frac{(\alpha T)}{2} = m\pi,\ m\in\mathbb{Z} \\ 0 \quad \text{else} \end{cases} \tag{24}$$

$$= \sum_{m\in\mathbb{Z}} \frac{F_\beta(\alpha)}{2\pi T}\left(P(\alpha)\star e^{-j\alpha\mathcal{T}} P^*(-\alpha)\right) \times \delta_{\alpha,\frac{2m\pi}{T}}. \tag{25}$$

3) FOR $T < \mathcal{T} < 2T$

The time-varying autocorrelation has the form:

$$\mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(2)}(t,\mathcal{T}) = \sum_{k\in\mathbb{Z}} \mathcal{E}\{p^C([t-kT]-\beta_k) \times p^{C*}([t-\mathcal{T}-(k-1)T]-\beta_{k-1})\} + \text{other terms zero}. \tag{26}$$

Again using same reasoning as given before equation (20) we can write:

$$\mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(2)}(t,\mathcal{T}) = \sum_{\alpha\in\mathbb{S}_\mathcal{T}} C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(2)}(\alpha,\mathcal{T})\, e^{j\alpha t}, \tag{27}$$

where, the cyclic autocorrelation is:

$$C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(2)}(\alpha,\mathcal{T}) = \lim_{T_0\to\infty} \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} \mathcal{C}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(2)}(t,\mathcal{T})\, e^{-j\alpha t} dt. \tag{28}$$

Substituting (26) in (28), and defining $s \triangleq t - kT$, we have:

$$C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(2)}(\alpha,\mathcal{T}) = \lim_{T_0\to\infty} \frac{1}{T_0} \sum_{k\in\mathbb{Z}} \int_{-T_0/2}^{T_0/2} \left(p^C\star f_\beta\right)(s) \times \left(p^{C*}\star f_\beta\right) \times (s-[\mathcal{T}-T])\, e^{-j\alpha(s+kT)} ds. \tag{29}$$

Let $q(\cdot) \triangleq \left(p^C\star f_\beta\right)(\cdot)$. Then (29) reduces to:

$$C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(2)}(\alpha,\mathcal{T}) = \sum_{m\in\mathbb{Z}} \frac{1}{2\pi T}\left[Q(\alpha)\star e^{-j\alpha(\mathcal{T}-T)} Q^*(-\alpha)\right] \times \delta_{\alpha,\frac{2m\pi}{T}}, \tag{30}$$

where,

$$\int_{-\infty}^{\infty} q(s)\, q^*(s-[\mathcal{T}-T])\, e^{-j\alpha s} ds = \frac{1}{2\pi}\left(Q(\alpha)\star e^{-j\alpha(\mathcal{T}-T)} Q^*(-\alpha)\right). \tag{31}$$

4) FOR GENERAL INTERVAL $GT < \mathcal{T} < (G+1)T; G \in \mathbb{Z}^+$

The cyclic autocorrelation is:

$$C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C(g)}(\alpha,\mathcal{T}) = \sum_{m\in\mathbb{Z}} \frac{1}{2\pi T}\left[Q(\alpha)\star e^{-j\alpha(\mathcal{T}-gT)} Q^*(-\alpha)\right] \times \delta_{\alpha,\frac{2m\pi}{T}}. \tag{32}$$

Now, let us define the following function:

$$\Re(\mathcal{T}) = \begin{cases} 1 & 0 \le \mathcal{T} \le T \\ 0 & \text{else}, \end{cases} \tag{33}$$

then the expression for $C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T})$ is given by:

$$\begin{aligned} C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T}) &= \{[\sum_{m,g\in\mathbb{Z}} \frac{1}{2\pi} \times \frac{1}{T} \times \delta_{\alpha,\frac{2m\pi}{T}} \times \left[Q(\alpha)\star e^{-j\alpha(\mathcal{T}-gT)} Q^*(-\alpha)\right]] \times \Re(\mathcal{T}-gT)\} \\ &+ \{[\sum_{m\in\mathbb{Z}} F_\beta(\alpha) \times \frac{1}{T} \times \delta_{\alpha,\frac{2m\pi}{T}} \times \left[\frac{1}{2\pi}\left(P(\alpha)\star e^{-j\alpha\mathcal{T}} P^*(-\alpha)\right)\right]] \times \Re(\mathcal{T})\} \\ &- \{[\sum_{m\in\mathbb{Z}} \frac{1}{2\pi} \times \frac{1}{T} \times \delta_{\alpha,\frac{2m\pi}{T}} \times \left[Q(\alpha)\star e^{-j\alpha\mathcal{T}} Q^*(-\alpha)\right]] \times \Re(\mathcal{T})\}. \end{aligned} \tag{34}$$

Again,

$$\mathcal{M}_{\mathcal{X}^C}(t,\mathcal{T}) = \sum_{\alpha\in\mathbb{S}_{\mathcal{T}}} M_{\mathcal{X}^C}(\alpha,\mathcal{T})\, e^{j\alpha t}, \tag{35}$$

$$\Rightarrow M_{\mathcal{X}^C}(\alpha,\mathcal{T}) = \lim_{T_0\to\infty}\frac{1}{T_0}\int_{-T_0/2}^{T_0/2}\mathcal{M}_{\mathcal{X}^C}(t,\mathcal{T})\, e^{-j\alpha t}dt, \tag{36}$$

$$= \left\langle \sum_{\widetilde{k},\widetilde{k_1}\in\mathbb{Z}} q\left(t-\widetilde{k}T\right) q^*\left(t-\mathcal{T}-\widetilde{k_1}T\right) e^{-j\alpha t}\right\rangle_t, \tag{37}$$

where, we have obtained (37), by substituting (12), (13) in (36). The above discussion can be summarized in terms of the following theorem:

Theorem 1: The cyclic covariances of the proposed model represented by (8) is given by:

$$C_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T}) = C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T}) - M_{\mathcal{X}^C}(\alpha,\mathcal{T}), \tag{38}$$

where, for each cyclic frequency α, the coefficients $C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T})$ are given by equation (34) and the coefficients $M_{\mathcal{X}^C}(\alpha,\mathcal{T})$ are given by (37).

With the expression for time-varying covariance of the process generated by the proposed model, some remarks are in order:

1) It is interesting to observe that the cyclic frequencies for any $\mathcal{T}$ are the same. Since they are all harmonically related, the process is cyclostationary.
2) We made certain assumptions, in the model, such as independent β_k's, the exact shape of the QRS complex at any translation, constant amplitudes of the QRS complexes, etc. The general model can have β_k's dependent, QRS pulse shape can vary, and the amplitudes also, in general, could be dependent random variables. So the assumptions used for mathematical tractability lead to α's being harmonically related, which need not be the case in general. Hence it is safe to assume that the model, in its proper generality, would lead to almost-cyclostationarity.

Since the synthesis/processing of the process will invariably involve sampling, it would be pertinent to derive appropriate results for the sampled version of the signal. In the following section, we take up this issue and derive the results for the sampled data case.

IV. SAMPLING THE ECG AND THE RECONSTRUCTED SIGNAL

As stated earlier, to synthesize or process a signal, one must first appropriately sample the given signal. In this section, we provide results for the discrete-time (sampled version) counterpart of the continuous case discussed in Section III. The sampled version of equation (8) is given by:

$$\mathcal{X}^C(nT_s) = \sum_{k\in\mathbb{Z}} \mathcal{A}_k p^C\left((nT_S) - [kT + \beta_k]\right). \tag{39}$$

Here, we assume that the mean period T is an integral multiple of the sampling period T_s, i.e., $\widehat{T} = MT_s$, where M is an integer, and $\widehat{T}$ is the approximated mean period. Note that T_s should be reasonably small. Also, for analytical simplicity, we assume β_k's to be integral multiple of T_s, i.e., $\widehat{\beta_k} = \ell_k T_s$, where, $\widehat{\beta_k}$ is the approximated zero mean fluctuation for the kth QRS complex. So, with the above assumptions, we can write the sampled and approximated version of the electrocardiogram signal as:

$$\widehat{\mathcal{X}^C}(nT_s) = \sum_{k\in\mathbb{Z}} \mathcal{A}_k p^C\left((nT_S) - [kMT_s + \ell_k T_s]\right), \tag{40}$$

$$\Rightarrow \widehat{\mathcal{X}^C}(nT_s) = \sum_{k\in\mathbb{Z}} \mathcal{A}_k p^C\left(T_s\left(n - [kM + \ell_k]\right)\right). \tag{41}$$

We denote the discrete-time version, of the QRS complex as $p^D(n) \triangleq p^C((nT_S))$. Then, discretized version of (41) is:

$$\widehat{\mathcal{X}_{\text{sam}}}(n) = \sum_{k\in\mathbb{Z}} \mathcal{A}_k p^D\left(n - [kM + \ell_k]\right), \tag{42}$$

$$= \sum_{k\in\mathbb{Z}} \mathcal{A}_k\left[p^D(n)\otimes\delta\left(n - [kM + \ell_k]\right)\right], \tag{43}$$

$$= p^D(n)\otimes\sum_{k\in\mathbb{Z}} \mathcal{A}_k\delta\left(n - [kM + \ell_k]\right). \tag{44}$$

where, $\otimes$ represents discrete-time convolution. Observe that here, we assume the sampling process picks up the same samples of all the delayed versions of the QRS complex. If the sampling is reasonably high, then it is easy to see that this assumption is justified. Let $\widehat{\alpha}$ be an approximation of the average heart rate $\alpha \triangleq 2\pi \times \frac{1}{T}$, obtained by counting the number of R peaks in the observation interval T_0, that is $\widehat{\alpha} = 2\pi \times \frac{\text{Number of R peaks in } T_0 \text{ interval}}{T_0}$.

V. RELATION BETWEEN CONTINUOUS AND DISCRETE MEAN COMPONENTS

Approximating the representation of mean component (refer eqn. (14)) as trigonometric expansion of (3), for every $\mathcal{T}$, we can write:

$$\mathcal{M}_{\mathcal{X}^C}(t,\mathcal{T})\,|_{t=nT_s} \triangleq \sum_{\alpha\in\mathbb{S}_{\mathcal{T}}} M_{\mathcal{X}^C}(\alpha,\mathcal{T})\exp(j\alpha t)\,|_{t=nT_s}, \tag{45}$$

$$= \sum_{\alpha\in\mathbb{S}_{\mathcal{T}}} M_{\mathcal{X}^C}(\alpha,\mathcal{T})\exp(j\alpha nT_s). \tag{46}$$

Again, assuming similar representation, in the sampled case, we write the discretized part of mean component as:

$$\mathcal{M}_{\mathcal{X}_{\text{sam}}}(n,\mathcal{T}) = \sum_{\alpha_{\text{sam}}\in\mathbb{S}_{\mathcal{T}}} M^{\text{sam}}_{\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}},\mathcal{T})\exp(j\alpha_{\text{sam}}n). \tag{47}$$

A. EXPRESSION FOR $M^{SAM}_{\mathcal{X}_{SAM}}(\alpha_{SAM}, \mathcal{T})$

This can be found out as follows by using (7), (46), (47):

$$M^{\text{sam}}_{\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}}, \mathcal{T}) \triangleq \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \mathcal{M}_{\mathcal{X}_{\text{sam}}}(n, \mathcal{T})\, e^{-j\alpha_{\text{sam}} n}, \tag{48}$$

$$= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \sum_{\alpha\in\mathbb{S}_{\mathcal{T}}} M_{\mathcal{X}^C}(\alpha, \mathcal{T}) \exp(j\alpha n T_s) \times \exp(-j\alpha_{\text{sam}} n), \tag{49}$$

$$= \sum_{\alpha\in\mathbb{S}_{\mathcal{T}}} M_{\mathcal{X}^C}(\alpha, \mathcal{T}) \times \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \exp(j(\alpha T_s - \alpha_{\text{sam}}) n), \tag{50}$$

$$\Rightarrow M^{\text{sam}}_{\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}}, \mathcal{T}) = M_{\mathcal{X}^C}(\alpha, \mathcal{T})\,|_{\alpha_{\text{sam}}=\alpha T_s \mod 2\pi}. \tag{51}$$

Note that (51) establishes the relation between components of the continuous domain and discrete domain mean components.

VI. RELATION BETWEEN CONTINUOUS AND DISCRETE COVARIANCE COMPONENTS

Let $\mathcal{X}_{\text{sam}}(n) \triangleq \mathcal{X}^C(t)\,|_{t=nT_s}$ be the discrete-time process obtained by uniformly sampling the continuous-time process $\mathcal{X}^C(t)$ with period T_s and sampling frequency $f_s = \frac{1}{T_s}$. Let $\mathbb{S}_{m_1T_s} \triangleq \mathbb{S}_{\mathcal{T}}|_{\mathcal{T}=m_1T_s}$, where $\mathbb{S}_{\mathcal{T}}$ is already defined in equation (5). For analytical convenience, we denote the sampled covariance function $\mathcal{C}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(n, m_1)$ as follows:

$$\mathcal{C}_{\mathcal{X}^C\mathcal{X}^C}(t, \mathcal{T})\,|_{t=nT_s, \mathcal{T}=m_1T_s} \triangleq \mathcal{E}\left(\mathcal{X}^C(t)\, \mathcal{X}^{C*}(t-\mathcal{T})\right)|_{t=nT_s, \mathcal{T}=m_1T_s}, \tag{52}$$

$$\triangleq \mathcal{E}\left(\mathcal{X}_{\text{sam}}(n)\, \mathcal{X}^*_{\text{sam}}(n-m_1)\right) = \mathcal{C}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(n, m_1). \tag{53}$$

Approximating the representation of the correlation component (refer eqn. (14)) as trigonometric expansion of (3), for every $\mathcal{T}$, we can write:

$$\mathcal{E}\left(\mathcal{X}^C(t)\, \mathcal{X}^{C*}(t-\mathcal{T})\right)|_{t=nT_s, \mathcal{T}=m_1T_s} = \sum_{\alpha\in\mathbb{S}_{\mathcal{T}}} C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \mathcal{T}) \exp(j\alpha t)\,|_{t=nT_s, \mathcal{T}=m_1T_s} = \sum_{\alpha\in\mathbb{S}_{m_1T_s}} C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, m_1T_s) \exp(j\alpha n T_s). \tag{54}$$

Now, assuming similar representation, for every m_1, in the sampled case, we write the discretized part of correlation component as:

$$\mathcal{E}\left(\mathcal{X}_{\text{sam}}(n)\, \mathcal{X}^*_{\text{sam}}(n-m_1)\right) = \sum_{\alpha_{\text{sam}}\in\widetilde{\mathbb{S}}} C^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}}, m_1) \exp(j\alpha_{\text{sam}} n), \tag{55}$$

where, $\widetilde{\mathbb{S}} = \left\{\alpha_{\text{sam}} \in [0, 2\pi) | C^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}}, m_1) \neq 0\right\}$.

A. EXPRESSION FOR $C^{SAM;CORR}_{\mathcal{X}_{SAM}\mathcal{X}_{SAM}}(\alpha_{SAM}, M_1)$

This can be found out as shown below:

$$C^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}}, m_1) \triangleq \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \mathcal{E}\left(\mathcal{X}_{\text{sam}}(n)\, \mathcal{X}^*_{\text{sam}}(n-m_1)\right) \times \exp(-j\alpha_{\text{sam}} n), \tag{56}$$

$$= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \exp(-j\alpha_{\text{sam}} n) \times \mathcal{E}\left(\mathcal{X}^C(t)\, \mathcal{X}^{C*}(t-\mathcal{T})\right)|_{t=nT_s, \mathcal{T}=m_1T_s}, \tag{57}$$

$$= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \sum_{\alpha\in\mathbb{S}_{m_1T_s}} C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, m_1T_s) \times \exp(j\alpha n T_s) \times \exp(-j\alpha_{\text{sam}} n), \tag{58}$$

$$= \sum_{\alpha\in\mathbb{S}_{m_1T_s}} C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, m_1T_s) \times \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \exp(j(\alpha T_s - \alpha_{\text{sam}}) n), \tag{59}$$

$$\Rightarrow C^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}}, m_1) = C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \mathcal{T})\,|_{\alpha_{\text{sam}}=\alpha T_s \mod 2\pi; \mathcal{T}=m_1T_s}. \tag{60}$$

Note that (60) establishes the relation between components of the continuous domain and discrete domain covariance components.

VII. TIME-VARYING SPECTRUM OF THE PROPOSED MODEL

This section contains expressions for the time-varying spectrum and the cyclic spectrum, and the following theorem:

Theorem 2: The cyclic spectrum of the proposed model represented by (8) is given by:

$$S_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \omega) = S^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \omega) - M_{\mathcal{X}^C}(\alpha, \omega), \tag{61}$$

where, $S^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \omega)$ and $M_{\mathcal{X}^C}(\alpha, \omega)$ are represented by equations (65), (69) respectively. Here, the cyclic covariance $C_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \mathcal{T})$ and cyclic spectrum $S_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \omega)$ are Fourier transform pairs, that is $C_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \mathcal{T}) \longleftrightarrow S_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \omega)$, and also, $C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \mathcal{T}) \longleftrightarrow S^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \omega)$; $M_{\mathcal{X}^C}(\alpha, \mathcal{T}) \longleftrightarrow M_{\mathcal{X}^C}(\alpha, \omega)$.

Proof: For the non-stationary process, generated from our model, equation (14) provides how its covariance is changing with time. Taking its Fourier transform $\mathcal{F}$, for each t, with respect to $\mathcal{T}$, we get the time-varying spectra of the process, as given below:

$$\mathcal{S}_{\mathcal{X}^C\mathcal{X}^C}(t, \omega) = \mathcal{S}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(t, \omega) - \mathcal{F}\left(\mathcal{M}_{\mathcal{X}^C}(t, \mathcal{T})\right). \tag{62}$$

Now, observe that, $\mathcal{S}^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(t, \omega) = \sum_\alpha S^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \omega)\, e^{j\alpha t}$, where the coefficients $S^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \omega)$ and $C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \mathcal{T})$ are Fourier transform pairs as shown:

$$S^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \omega) = \int_{-\infty}^{\infty} C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha, \mathcal{T})\, e^{-j\omega\mathcal{T}}\, d\mathcal{T}. \tag{63}$$

Substituting (34) in (63), we obtain:

$$\begin{aligned}
&S^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\omega)\\
&=\Bigg[\left\{\int_{-\infty}^{\infty}\sum_{g}q^*(s-[\mathcal{T}-gT])\,\Re(\mathcal{T}-gT)\,e^{-j\omega\mathcal{T}}d\mathcal{T}\right\}\\
&\quad\times\left\{\int_{-\infty}^{\infty}q(s)\,e^{-j\alpha s}ds\right\}\left\{\lim_{T_0\to\infty}\frac{1}{T_0}\sum_{k\in\mathbb{Z}}e^{-j\alpha kT}\right\}\Bigg]\\
&\quad+\Bigg[\int_{-\infty}^{\infty}p^*(s-\mathcal{T})\,\Re(\mathcal{T})\,e^{-j\omega\mathcal{T}}d\mathcal{T}\times\int_{-\infty}^{\infty}p(s)\,e^{-j\alpha s}ds\\
&\quad\left\{\int_{-\infty}^{\infty}f_\beta(\xi)\,e^{-j\alpha\xi}d\xi\right\}\left\{\lim_{T_0\to\infty}\frac{1}{T_0}\sum_{k\in\mathbb{Z}}e^{-j\alpha kT}\right\}\Bigg]\\
&\quad-\Bigg[\left\{\int_{-\infty}^{\infty}\sum_{g\in\mathbb{Z}}q^*(s-[\mathcal{T}-gT])\,\Re(\mathcal{T})\,e^{-j\omega\mathcal{T}}d\mathcal{T}\right\}\\
&\quad\times\left\{\int_{-\infty}^{\infty}q(s)\,e^{-j\alpha s}ds\right\}\left\{\lim_{T_0\to\infty}\frac{1}{T_0}\sum_{k\in\mathbb{Z}}e^{-j\alpha kT}\right\}\Bigg]. \qquad (64)
\end{aligned}$$

Let, $\mathcal{T}-gT=\widetilde{\mathcal{T}}$, and after performing algebraic manipulations, (64) reduces to:

$$\begin{aligned}
S^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\omega)&=\left\{\left[\left\{\Re(s)\,e^{-j\omega s}\right\}\star q(s)\right]\right.\\
&\quad\left.\times Q(\alpha)\times\left\{\sum_{g\in\mathbb{Z}}e^{-j\omega gT}\frac{1}{T}\sum_{m\in\mathbb{Z}}\delta_{\alpha,\frac{2m\pi}{T}}\right\}\right\}\\
&\quad+\left\{\left[\left\{\Re(s)\,e^{-j\omega s}\right\}\star p(s)\right]\times P(\alpha)\right.\\
&\quad\left.\times F_\beta(\alpha)\times\left\{\frac{1}{T}\sum_{m\in\mathbb{Z}}\delta_{\alpha,\frac{2m\pi}{T}}\right\}\right\}\\
&\quad-\left\{\sum_{g\in\mathbb{Z}}\left[\left\{\Re(s)\,e^{-j\omega s}\right\}\star q(s+gT)\right]\right.\\
&\quad\left.\times Q(\alpha)\times\left\{\frac{1}{T}\sum_{m\in\mathbb{Z}}\delta_{\alpha,\frac{2m\pi}{T}}\right\}\right\}. \qquad (65)
\end{aligned}$$

Again,

$$M_{\mathcal{X}^C}(\alpha,\omega)=\mathcal{F}\left(M_{\mathcal{X}^C}(\alpha,\mathcal{T})\right). \qquad (66)$$

Substituting (36) in (66), we obtain

$$\begin{aligned}
&M_{\mathcal{X}^C}(\alpha,\omega)\\
&=\lim_{T_0\to\infty}\frac{1}{T_0}\int_{-T_0/2}^{T_0/2}\sum_{\widetilde{k},\widetilde{k}_1\in\mathbb{Z}}q\left(t-\widetilde{k}T\right)\\
&\quad\times\left[\int q^*\left(t-\mathcal{T}-\widetilde{k}_1T\right)e^{-j\omega\mathcal{T}}d\mathcal{T}\right]\times e^{-j\alpha t}dt. \qquad (67)
\end{aligned}$$

Let $t-\mathcal{T}-\widetilde{k}_1T=\widetilde{\mathcal{T}}$. Then,

$$\begin{aligned}
&M_{\mathcal{X}^C}(\alpha,\omega)\\
&=\lim_{T_0\to\infty}\frac{1}{T_0}\int_{-T_0/2}^{T_0/2}\left[\int q\left(\widetilde{\mathcal{T}}\right)e^{-j\omega\widetilde{\mathcal{T}}}d\widetilde{\mathcal{T}}\right]^*\\
&\quad\times\sum_{\widetilde{k},\widetilde{k}_1}q\left(t-\widetilde{k}T\right)e^{-j\alpha t}e^{-j\omega\left(t-\widetilde{k}_1T\right)}dt, \qquad (68)\\
&=Q^*(\omega)\sum_{\widetilde{k}_1\in\mathbb{Z}}e^{j\omega\widetilde{k}_1T}\left\langle\sum_{\widetilde{k}\in\mathbb{Z}}q\left(t-\widetilde{k}T\right)e^{-j(\alpha+\omega)t}\right\rangle_t. \qquad (69)
\end{aligned}$$

Combining (65) & (69), we get the cyclic spectrum of the process generated using the proposed model. Clearly, $\alpha=0$ provides the stationary part of the process, and the information about non-stationarity is contained in the alphas other than zeros. Since we have this decoupling of stationarity and non-stationary parts, we can use the results of [15] to get the form to measure the degree of non-stationarity of this process. ■

A. GENERAL OBSERVATIONS

We make the following observations, which make the relationship between time-varying spectra of continuous-time and its sampled version explicit:

1) Consider the cyclic mean corresponding to the sampled data case as below:

$$\begin{aligned}
&M^{\text{sam}}_{\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}},\omega)\\
&=\int M^{\text{sam}}_{\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}},\mathcal{T})\,e^{-j\omega\mathcal{T}}d\mathcal{T}, \qquad (70)\\
&=\int M_{\mathcal{X}^C}(\alpha,\mathcal{T})\,|_{\alpha_{\text{sam}}=\alpha T_s\ \bmod 2\pi}\,e^{-j\omega\mathcal{T}}d\mathcal{T}, \qquad (71)\\
&=M_{\mathcal{X}^C}(\alpha,\omega)\,|_{\alpha_{\text{sam}}=\alpha T_s\ \bmod 2\pi}. \qquad (72)
\end{aligned}$$

2) Consider the cyclic covariance corresponding to the sampled data case as below:

$$\begin{aligned}
&S^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}},\Omega)\\
&=\sum_{m_1\in\mathbb{Z}}C^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}},m_1)\exp(-j\Omega m_1), \qquad (73)\\
&=\sum_{m_1\in\mathbb{Z}}C^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\mathcal{T})\,|_{|\alpha_{\text{sam}}=\alpha T_s\ \bmod 2\pi;\mathcal{T}=m_1T_s}\\
&\quad\times\exp(-j\Omega m_1), \qquad (74)\\
&=S^{\text{corr}}_{\mathcal{X}^C\mathcal{X}^C}(\alpha,\Omega)\,|_{|\alpha_{\text{sam}}=\alpha T_s\ \bmod 2\pi;\mathcal{T}=m_1T_s}. \qquad (75)
\end{aligned}$$

VIII. MEASURE FOR THE DEGREE OF NON STATIONARITY

At this point, we can define two kinds of measures for the degree of non-stationarities. One is based on the covariance function, and the other is based on the correlation function. Conceptually, the $\mathcal{MDNST}$ is the departure from being stationary in both cases. Here, we consider the $\mathcal{MDNST}$ based on a time-varying correlation function. This is due to the fact that the correlation function captures complete variation,

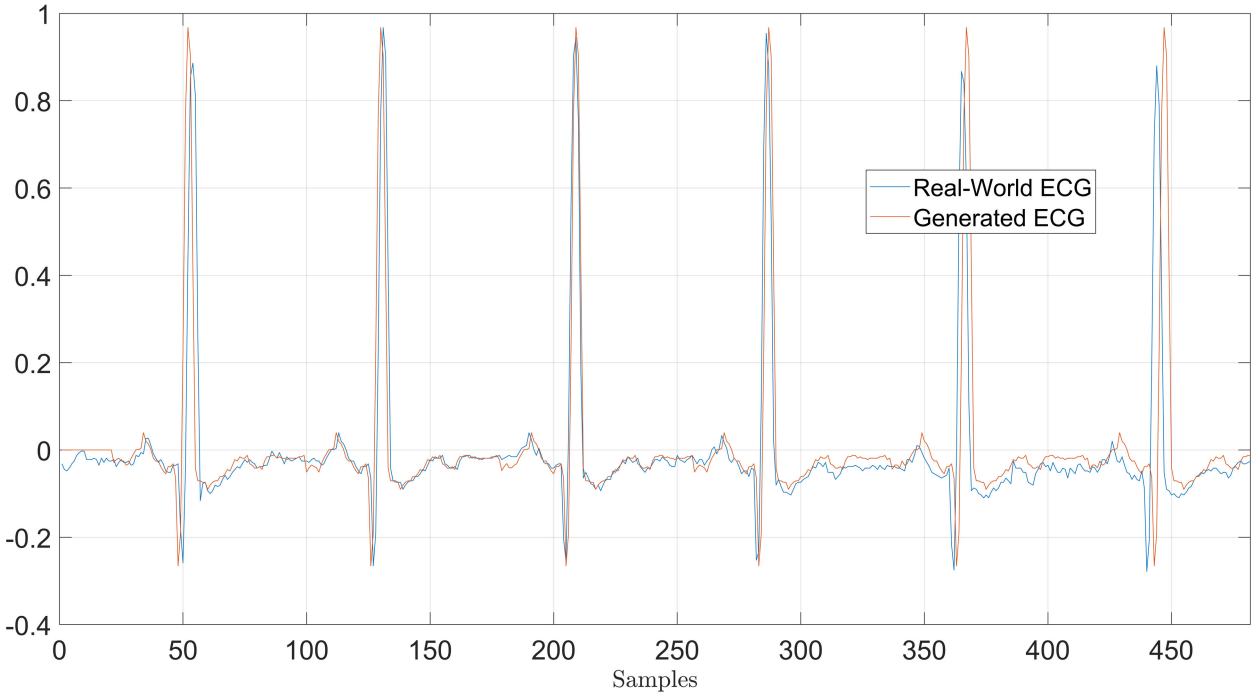


FIGURE 3. Reconstructed electrocardiogram of a normal sinus rhythm person.

unlike the covariance function, which captures time variation, which is only over and above the time variation of the mean function. Using (60), (75), we define $\mathcal{MDNST}$ as shown:

$$\mathcal{MDNST}\left(\mathcal{X}_{\text{sam}}\right) \triangleq 1 - \frac{\int \left| C^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(0, m_1) \right|^2 dm_1}{\sum_{\alpha_{\text{sam}}} \int \left| C^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}}, m_1) \right|^2 dm_1}, \tag{76}$$

$$= 1 - \frac{\int \left| S^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(0, \Omega) \right|^2 d\Omega}{\sum_{\alpha_{\text{sam}}} \int \left| S^{\text{sam;corr}}_{\mathcal{X}_{\text{sam}}\mathcal{X}_{\text{sam}}}(\alpha_{\text{sam}}, \Omega) \right|^2 d\Omega}. \tag{77}$$

IX. NUMERICAL RESULTS AND DISCUSSIONS

In this section, experiments are conducted to validate the efficacy of the mathematical model proposed in Section II for the ECG signal. Multiple ECG data from the PhysioNet database were analyzed. We perform experiments in this section to demonstrate the usefulness of the proposed model. In this study, ECG datasets from the MIT-BIH Normal Sinus Rhythm Database [16], the Apnea-ECG Database [17] and the MIT-BIH Supraventricular Arrhythmia Database [18] are used.We synthesize the ECG signal using our proposed model (eqn. 44), which extracts the QRS complex from a specific health condition and uses the average heart rate estimated from the ECG of the given health state.

A. ELECTROCARDIOGRAM RECONSTRUCTION

In this subsection we reconstruct the electrocardiogram using eqn. (44) for each subject (person) chosen from normal sinus rhythm database [16], sleep apnea database [17] and supraventricular arrhythmia database [18]. To verify the efficacy of the proposed model for the electrocardiogram signal, we rerun the analysis from Section IV on all of the subjects for whom electrocardiogram measurements are available in [16], [17], and [18]. This paper is condensed by including only the results from three subjects. The other test subjects achieved similar outcomes. We provide the reconstructed signal for each of these cases having observation interval $T_0 = 1$ min:

1) Normal sinus rhythm electrocardiogram having $f_s = 128$ Hz, and estimated mean period $\widehat{T} \approx \frac{81}{128} \approx 0.65$ seconds, as shown in Fig. 3.
2) Sleep Apnea electrocardiogram having $f_s = 100$ Hz, and estimated mean period $\widehat{T} \approx \frac{91}{100} = 0.91$ seconds, as shown in Fig. 4.
3) Supraventricular arrhythmia electrocardiogram having $f_s = 128$ Hz, and estimated mean period $\widehat{T} \approx \frac{125}{128} \approx 0.98$ seconds, is shown in Fig. 5.

B. ELECTROCARDIOGRAM CLASSIFICATION

In order to assess the effectiveness of our proposed definition for quantifying the degree of non-stationarity in electrocardiogram signals, we conducted a simulation study of the definitions provided in Section VIII. This study

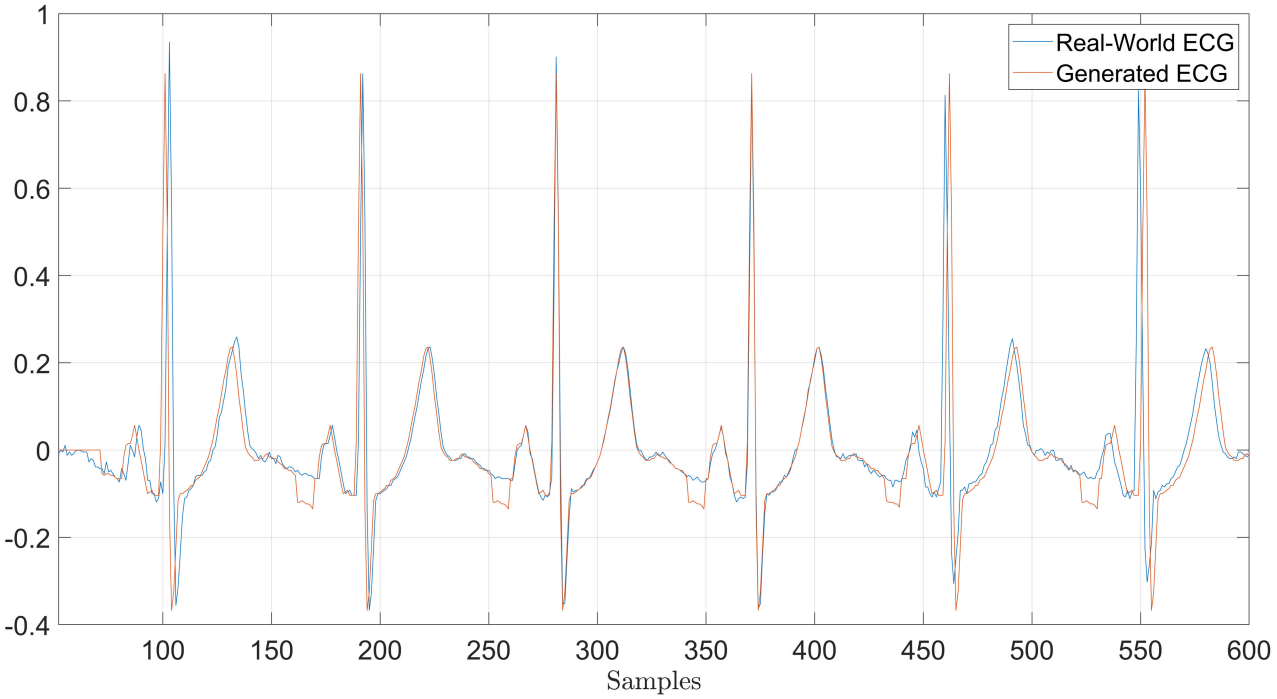


FIGURE 4. Reconstructed electrocardiogram of a sleep apnea subject.

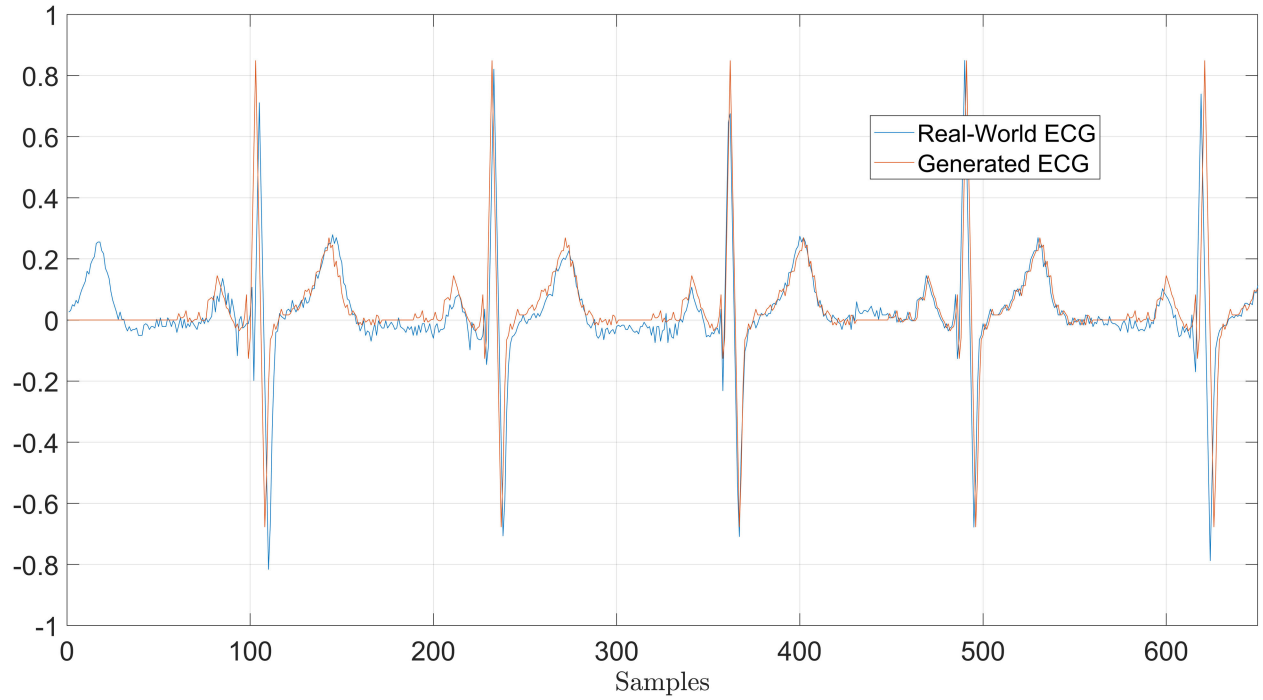


FIGURE 5. Reconstructed electrocardiogram of a supraventricular arrhythmia subject.

is carried out for all subjects whose ECG readings were accessible in [16], [17], and [18]. This experiment was conducted to demonstrate our proposed measure's efficacy in achieving accurate classification. Tables 1 to 18 have been condensed to exclusively present the findings of six individuals in each category of normal sinus rhythm,

TABLE 1. $\mathcal{MDNST}$ **for normal sinus rhythm (subject 1).**

$f_s = 128$Hz	Normal Sinus Rhythm (Subject 1)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.2295	0.2295
$\triangle = 0.02$	0.2744	0.2295
$\triangle = 0.03$	0.3099	0.2295
$\triangle = 0.08$	0.5774	0.2295
$\triangle = 0.25$	0.7837	0.2295

TABLE 2. $\mathcal{MDNST}$ **for normal sinus rhythm (subject 2).**

$f_s = 128$Hz	Normal Sinus Rhythm (Subject 2)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.1881	0.1881
$\triangle = 0.02$	0.2220	0.1881
$\triangle = 0.03$	0.2479	0.1881
$\triangle = 0.08$	0.4442	0.1881
$\triangle = 0.25$	0.6511	0.1881

TABLE 3. $\mathcal{MDNST}$ **for normal sinus rhythm (subject 3).**

$f_s = 128$Hz	Normal Sinus Rhythm (Subject 3)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.1724	0.1724
$\triangle = 0.02$	0.2586	0.1724
$\triangle = 0.03$	0.3370	0.1724
$\triangle = 0.08$	0.7143	0.1724
$\triangle = 0.25$	0.6884	0.1724

TABLE 4. $\mathcal{MDNST}$ **for normal sinus rhythm (subject 4).**

$f_s = 128$Hz	Normal Sinus Rhythm (Subject 4)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.2053	0.2053
$\triangle = 0.02$	0.2437	0.2053
$\triangle = 0.03$	0.2702	0.2053
$\triangle = 0.08$	0.4537	0.2053
$\triangle = 0.25$	0.7626	0.2053

TABLE 5. $\mathcal{MDNST}$ **for normal sinus rhythm (subject 5).**

$f_s = 128$Hz	Normal Sinus Rhythm (Subject 5)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.1525	0.1525
$\triangle = 0.02$	0.1861	0.1525
$\triangle = 0.03$	0.2112	0.1525
$\triangle = 0.08$	0.3813	0.1525
$\triangle = 0.25$	0.7293	0.1525

TABLE 6. $\mathcal{MDNST}$ **for normal sinus rhythm (subject 6).**

$f_s = 128$Hz	Normal Sinus Rhythm (Subject 6)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.0982	0.0982
$\triangle = 0.02$	0.1285	0.0982
$\triangle = 0.03$	0.1558	0.0982
$\triangle = 0.08$	0.3814	0.0982
$\triangle = 0.25$	0.6797	0.0982

TABLE 7. $\mathcal{MDNST}$ **for sleep apnea (subject 1).**

$f_s = 100$Hz	Sleep Apnea (Subject 1)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.3593	0.3593
$\triangle = 0.02$	0.3884	0.3593
$\triangle = 0.03$	0.4464	0.3593
$\triangle = 0.08$	0.7437	0.3593
$\triangle = 0.25$	0.7368	0.3593

TABLE 8. $\mathcal{MDNST}$ **for sleep apnea (subject 2).**

$f_s = 100$Hz	Sleep Apnea (Subject 2)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.3716	0.3716
$\triangle = 0.02$	0.3855	0.3716
$\triangle = 0.03$	0.4055	0.3716
$\triangle = 0.08$	0.5387	0.3716
$\triangle = 0.25$	0.7676	0.3716

TABLE 9. $\mathcal{MDNST}$ **for sleep apnea (subject 3).**

$f_s = 100$Hz	Sleep Apnea (Subject 3)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.3764	0.3764
$\triangle = 0.02$	0.3951	0.3764
$\triangle = 0.03$	0.4385	0.3764
$\triangle = 0.08$	0.5278	0.3764
$\triangle = 0.25$	0.5607	0.3764

TABLE 10. $\mathcal{MDNST}$ **for sleep apnea (subject 4).**

$f_s = 100$Hz	Sleep Apnea (Subject 4)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.3585	0.3585
$\triangle = 0.02$	0.3734	0.3585
$\triangle = 0.03$	0.4127	0.3585
$\triangle = 0.08$	0.6953	0.3585
$\triangle = 0.25$	0.8001	0.3585

sleep apnea, and supraventricular arrhythmia. The remaining subjects attained comparable results.

1) DISCUSSION

- We observe that the observation interval has no influence on $\mathcal{MDNST}$.
- We compute $\mathcal{MDNST}$ for various patients. For normal sinus rhythm, sleep apnea, and supraventricular arrhythmia cases. We observe that the $\mathcal{MDNST}$ for the proposed model and the real-world ECG are identical for $\triangle = 0.01$. Tables 1 to 18 contain the results.
- Tables 1 to 18 provide $\mathcal{MDNST}$ for electrocardiogram with the proposed model, with QRS complex extracted from electrocardiogram of the corresponding health condition, for various values of parameter Δ. Moreover, in the same tables, the $\mathcal{MDNST}$ of the real-world

TABLE 11. $\mathcal{MDNST}$ for sleep apnea (subject 5).

$f_s = 100$Hz	Sleep Apnea (Subject 5)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.3503	0.3503
$\triangle = 0.02$	0.3812	0.3503
$\triangle = 0.03$	0.4347	0.3503
$\triangle = 0.08$	0.7582	0.3503
$\triangle = 0.25$	0.7386	0.3503

TABLE 12. $\mathcal{MDNST}$ for sleep apnea (subject 6).

$f_s = 100$Hz	Sleep Apnea (Subject 6)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.3598	0.3598
$\triangle = 0.02$	0.4890	0.3598
$\triangle = 0.03$	0.5847	0.3598
$\triangle = 0.08$	0.7059	0.3598
$\triangle = 0.25$	0.7065	0.3598

TABLE 13. $\mathcal{MDNST}$ for supraventricular arrhythmia (subject 1).

$f_s = 128$Hz	Supraventricular Arrhythmia (Subject 1)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.5143	0.5143
$\triangle = 0.02$	0.5520	0.5143
$\triangle = 0.03$	0.5834	0.5143
$\triangle = 0.08$	0.7505	0.5143
$\triangle = 0.25$	0.7672	0.5143

TABLE 14. $\mathcal{MDNST}$ for supraventricular arrhythmia (subject 2).

$f_s = 128$Hz	Supraventricular Arrhythmia (Subject 2)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.6265	0.6265
$\triangle = 0.02$	0.6308	0.6265
$\triangle = 0.03$	0.6329	0.6265
$\triangle = 0.08$	0.6479	0.6265
$\triangle = 0.25$	0.7479	0.6265

TABLE 15. $\mathcal{MDNST}$ for supraventricular arrhythmia (subject 3).

$f_s = 128$Hz	Supraventricular Arrhythmia (Subject 3)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.5959	0.5959
$\triangle = 0.02$	0.6013	0.5959
$\triangle = 0.03$	0.6062	0.5959
$\triangle = 0.08$	0.6604	0.5959
$\triangle = 0.25$	0.7495	0.5959

ECG of the patient is also provided. For various classes, we observe that the range of $\mathcal{MDNST}$ is distinct.

- We conjecture the following ranges for the cases mentioned above:
  1) Normal Sinus Rhythm: $0.07 \leq \mathcal{MDNST} \leq 0.25$.
  2) Sleep Apnea: $0.30 \leq \mathcal{MDNST} \leq 0.40$.

TABLE 16. $\mathcal{MDNST}$ for supraventricular arrhythmia (subject 4).

$f_s = 128$Hz	Supraventricular Arrhythmia (Subject 4)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.5696	0.5696
$\triangle = 0.02$	0.6119	0.5696
$\triangle = 0.03$	0.6455	0.5696
$\triangle = 0.08$	0.7596	0.5696
$\triangle = 0.25$	0.8051	0.5696

TABLE 17. $\mathcal{MDNST}$ for supraventricular arrhythmia (subject 5).

$f_s = 128$Hz	Supraventricular Arrhythmia (Subject 5)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.6185	0.6185
$\triangle = 0.02$	0.6303	0.6185
$\triangle = 0.03$	0.6397	0.6185
$\triangle = 0.08$	0.6702	0.6185
$\triangle = 0.25$	0.7118	0.6185

TABLE 18. $\mathcal{MDNST}$ for supraventricular arrhythmia (subject 6).

$f_s = 128$Hz	Supraventricular Arrhythmia (Subject 6)	
$\triangle$ values	$\mathcal{MDNST}$ (Proposed Model)	$\mathcal{MDNST}$ (Without Model)
$\triangle = 0.01$	0.6961	0.6961
$\triangle = 0.02$	0.7031	0.6961
$\triangle = 0.03$	0.7082	0.6961
$\triangle = 0.08$	0.7128	0.6961
$\triangle = 0.25$	0.6513	0.6961

TABLE 19. $\mathcal{MDNST}$ for or different values of observation intervals for normal sinus rhythm (subject-1).

$f_s = 360$Hz	Normal Sinus Rhythm (Subject 1)		
$\triangle$ values	$\mathcal{MDNST}$ (Proposed model)		
	$T_0 = 10$ secs	$T_0 = 30$ secs	$T_0 = 1$ min
$\triangle = 0.01$	0.2295	0.2295	0.2295
$\triangle = 0.02$	0.2744	0.2744	0.2744
$\triangle = 0.03$	0.3099	0.3099	0.3099
$\triangle = 0.08$	0.5774	0.5774	0.5774
$\triangle = 0.25$	0.7837	0.7837	0.7837

3) Supraventricular Arrhythmia: $0.50 \leq \mathcal{MDNST} \leq 0.75$.

C. EFFECT OF OBSERVATION INTERVAL ON $\mathcal{MDNST}$

To assess the impact of observation intervals on the estimation of $\mathcal{MDNST}$, we calculate the $\mathcal{MDNST}$ for different values of observation intervals and provide the findings in Table 19. Due to the consistent findings across all classes, the outcomes are presented exclusively for a single class and subject in the interest of conciseness.

D. MODEL COMPARISON

In this subsection, we compare our proposed model with previous models as shown:

1) The authors in [1] apply Wald's sequential test to a Markovian heart rhythm interval model and graph

the predicted number of observations under pairwise testing of three rhythm classes as a function of error rate. Such a test has a significant drawback: the sample size needed to complete it cannot be established before data collecting begins, reducing test subject resources. Our model, in contrast, applies to datasets available in [16], [17], and [18] for all subjects.

2) Inaccurate modeling of small electrical activity during early and late QRS periods and discrepancies between Durrer et al. surface potential maps and empirically acquired maps prevent a realistic discrete dipole cardiac model [2]. In contrast, the mathematical model proposed in our paper (eqn. 8), though we have presented the analysis with simplified assumptions for analytical convenience, can represent the temporal differences between individual waves and their duration, amplitude levels, polarity, and shape of the QRS complex. This renders our proposed mathematical model more plausible and tractable from a mathematical standpoint.
3) The research presented in [3] had a few limitations, one of which was that it needed to include models of higher-order ECG signals that had abnormalities, making it difficult to discover an exact match between the signal and its model. In addition, no classification has been offered for the abnormalities that have been detected. Furthermore, the utilization of modeling techniques was not conducted on large datasets. In our paper, we have addressed these limitations. They are listed below:
   - The electrocardiogram is modeled as a second-order almost-cyclostationary (ACS) process in our paper. In other words, the autocorrelation function is an almost periodic function of time whose coefficients, known as cyclic autocorrelation functions, depend on the lag parameter but whose frequencies (second-order) cycle frequencies do not.
   - With the aid of our proposed model (eqn.44), we were able to reconstruct the electrocardiogram as depicted in Fig. 3, Fig. 4 and Fig. 5. The reconstructed electrocardiogram closely matched the actual electrocardiogram.
4) It is clear from eqn. (3.6) and eqn. 2.24 (b) and subsequent discussion of [21] that $\dot{\epsilon}(t)$ would be just a constant; hence, the model may not lead to a time-varying heart rate. Also, if one observes eqn. (2.2) of [21], it first considers a deterministic periodic component, representing the mean of ECG, added with a cyclostationary term, modeling the variation about the mean. Then, this signal is time-warped to model the almost periodicity. Since the mean and cyclostationary part will get time warped, even the mean component now will not be periodic, which one would expect. However, in our proposed model, the gap between two QRS complexes is modeled as an average component of the period, plus its variation around the mean, modeled as zero mean, uniformly distributed random variables, leading to the mean of the process as periodic as one would expect.

X. CONCLUSION AND SCOPE FOR FUTURE WORK

In this paper, we propose a mathematical model for the electrocardiogram (ECG) signal and also provide expressions for the $\mathcal{MDNST}$ of the signal generated by the proposed model. By comparing the degrees of non-stationarity for our proposed model with the degrees of non-stationarity for a real-world ECG, we conclude:

1) For a particular $\triangle$ value, where $\triangle$ is the parameter associated with the probability density function of departure from the average period, the degrees of the non-stationarity match.
2) As anticipated, the non-stationarity associated with our proposed model increases as the $\triangle$ value increases.
3) Our model is suitable for classifying various heart-related abnormalities based on the $\mathcal{MDNST}$ range.
4) As mentioned, we generated (reconstructed) the electrocardiogram (ECG) signal while comparing it with real-world ECG, with QRS complexes extracted from the corresponding disease and also using the average heart rate estimated from the ECG of the given health state. This observation also leads us to infer that the QRS complex shape is not invariant. The variation in the QRS complex may be of immense diagnostic importance.

Some of the exciting directions for future explorations could be as follows:

- Analysis of a more general model where $\mathcal{A}_k'$s, β_k's, and QRS complex variations and dependence can be considered.
- Adaptive estimation of $\mathcal{MDNST}$, hence modeling its time variation for various heart-related abnormalities.
- Modeling of QRS and integrating with the proposed model.

REFERENCES

[1] E. Tsui and E. Wong, "A sequential approach to heart-beat interval classification (corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 5, pp. 596–599, Sep. 1975, doi: 10.1109/TIT.1975.1055425.

[2] B. N. Cuffin and D. N. Geselowitz, "Studies of the electrocardiogram using realistic cardiac and torso models," *IEEE Trans. Biomed. Eng.*, vol. BME-24, no. 3, pp. 242–252, May 1977, doi: 10.1109/TBME.1977.326209.

[3] I. S. N. Murthy, M. R. Rangaraj, K. J. Udupa, and A. K. Goyal, "Homomorphic analysis and modeling of ECG signals," *IEEE Trans. Biomed. Eng.*, vol. BME–26, no. 6, pp. 330–344, Jun. 1979, doi: 10.1109/TBME.1979.326562.

[4] M. O. Mendez, A. M. Bianchi, M. Matteucci, S. Cerutti, and T. Penzel, "Sleep apnea screening by autoregressive models from a single ECG lead," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 12, pp. 2838–2850, Dec. 2009, doi: 10.1109/TBME.2009.2029563.

[5] J. Huang, B. Chen, B. Yao, and W. He, "ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network," *IEEE Access*, vol. 7, pp. 92871–92880, 2019, doi: 10.1109/ACCESS.2019.2928017.

[6] Y. Xing, N. Rao, M. Miao, Q. Li, Q. Li, X. Chen, Q. Zhang, and J. Wu, "Task-state heart rate variability parameter-based depression detection model and effect of therapy on the parameters," *IEEE Access*, vol. 7, pp. 105701–105709, 2019, doi: 10.1109/ACCESS.2019.2932393.
[7] T. Nakane, T. Ito, N. Matsuura, H. Togo, and A. Hirata, "Forward electrocardiogram modeling by small dipoles based on whole-body electric field analysis," *IEEE Access*, vol. 7, pp. 123463–123472, 2019, doi: 10.1109/ACCESS.2019.2938409.
[8] R. V. Andreao, B. Dorizzi, and J. Boudy, "ECG signal analysis through hidden Markov models," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 8, pp. 1541–1549, Aug. 2006, doi: 10.1109/TBME.2006.877103.
[9] H. Baali, R. Akmeliawati, M. J. E. Salami, A. Khorshidtalab, and E. Lim, "ECG parametric modeling based on signal dependent orthogonal transform," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1293–1297, Oct. 2014, doi: 10.1109/LSP.2014.2332425.
[10] P. Laguna, J. P. M. Cortés, and E. Pueyo, "Techniques for ventricular repolarization instability assessment from the ECG," *Proc. IEEE*, vol. 104, no. 2, pp. 392–415, Feb. 2016, doi: 10.1109/JPROC.2015.2500501.
[11] A. V. Dandawate and G. B. Giannakis, "Statistical tests for presence of cyclostationarity," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2355–2369, Sep. 1994, doi: 10.1109/78.317857.
[12] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Amer.*, vol. 45, no. 2, pp. 458–465, Feb. 1969, doi: 10.1121/1.1911395.
[13] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio Electroacoustics*, vol. 16, no. 2, pp. 221–226, Jun. 1968, doi: 10.1109/TAU.1968.1161965.
[14] M. B. Abubaker and B. Babayigit, "Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods," *IEEE Trans. Artif. Intell.*, vol. 4, no. 2, pp. 373–382, Apr. 2023, doi: 10.1109/TAI.2022.3159505.
[15] A. K. Dutta, S. D. Joshi, and B. Lall, "Measure for degree of time variance and measure for degree of non-stationarity: Application to discrete LPTV systems," *Signal Process.*, vol. 183, Jun. 2021, Art. no. 107995, doi: 10.1016/j.sigpro.2021.107995.
[16] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000. [Online]. Available: http://circ.ahajournals.org/content/101/23/e215.full
[17] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ECG database," in *Proc. Comput. Cardiol.*, vol. 27, 2000, pp. 255–258.
[18] S. D. Greenwald, "Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information," Ph.D. thesis, Harvard-MIT Division Health Sci. Technol., Cambridge, MA, USA, 1990.
[19] L. Izzo and A. Napolitano, "The higher order theory of generalized almost-cyclostationary time series," *IEEE Trans. Signal Process.*, vol. 46, no. 11, pp. 2975–2989, Nov. 1998, doi: 10.1109/78.726811.
[20] M. Kulldorff, R. L. Davis, M. Kolczak, E. Lewis, T. Lieu, and R. Platt, "A maximized sequential probability ratio test for drug and vaccine safety surveillance," *Sequential Anal.*, vol. 30, no. 1, pp. 58–78, Jan. 2011, doi: 10.1080/07474946.2011.539924.
[21] A. Napolitano, "Modeling the electrocardiogram as oscillatory almost-cyclostationary process," *IEEE Access*, vol. 10, pp. 13193–13209, 2022, doi: 10.1109/ACCESS.2022.3147500.
[22] S. Mihandoost and M. C. Amirani, "Cyclic spectral analysis of electrocardiogram signals based on GARCH model," *Biomed. Signal Process. Control*, vol. 31, pp. 79–88, Jan. 2017, doi: 10.1016/j.bspc.2016.07.012.
[23] M. Haritopoulos, J. Roussel, C. Capdessus, and A. K. Nandi, "Cyclostationarity-based estimation of the foetus subspace dimension from ECG recordings," in *Proc. XIII Medit. Conf. Med. Biol. Eng. Comput.* Cham, Switzerland: Springer, 2014, pp. 726–729.
[24] S. Horstmann, D. Ramírez, and P. J. Schreier, "Joint detection of almost-cyclostationary signals and estimation of their cycle period," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1695–1699, Nov. 2018, doi: 10.1109/LSP.2018.2871961.
[25] L. Izzo and A. Napolitano, "Sampling of generalized almost-cyclostationary signals," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1546–1556, Jun. 2003, doi: 10.1109/TSP.2003.811236.
[26] A. Napolitano, *Cyclostationary Processes and Time Series*. Amsterdam, The Netherlands: Elsevier 2019, doi: 10.1016/C2017-0-04240-4.

AYAN KUMAR DUTTA received the M.Tech. degree in communication engineering from the Vellore Institute of Technology, Vellore, in 2016. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Indian Institute of Technology Delhi. His research interests include multirate signal processing, statistical signal processing, biomedical signal processing, and mathematical modeling.

SHIV DUTT JOSHI received the Ph.D. degree from the Indian Institute of Technology (IIT) Delhi, New Delhi, India, in 1988. He has been a Professor with the Department of Electrical Engineering, IIT Delhi, since March 2000. His research interests include the development of fast algorithms for statistical signal processing, signal representation, multi-scale signal modeling and processing, category theoretical approach to signal representation and processing, and image processing.

BREJESH LALL (Member, IEEE) received the B.E. and M.E. degrees in electronics and communications engineering from the Delhi College of Engineering, Delhi University, in 1991 and 1992, respectively, and the Ph.D. degree in multirate signal processing from Indian Institute of Technology (IIT) Delhi, New Delhi, in 1999. From 1997 to 2005, he was with the Digital Signal Processing Group, Hughes Software Systems. Since 2005, he has been a Professor with the Department of Electrical Engineering, IIT Delhi. He has contributed to research and teaching in the general area of signal processing.

On Transfer of Adversarial Robustness from Pretraining to Downstream Tasks

Laura F. Nern
Yahoo Research
laurafee.nern@yahooinc.com

Harsh Raj
Delhi Technological University
harsh777111raj@gmail.com

Maurice Georgi
Hyundai Mobis
georgimaurice@gmail.com

Yash Sharma
University of Tübingen
yash.sharma@bethgelab.org

Abstract

As large-scale training regimes have gained popularity, the use of pretrained models for downstream tasks has become common practice in machine learning. While pretraining has been shown to enhance the performance of models in practice, the transfer of robustness properties from pretraining to downstream tasks remains poorly understood. In this study, we demonstrate that the robustness of a linear predictor on downstream tasks can be constrained by the robustness of its underlying representation, regardless of the protocol used for pretraining. We prove (i) a bound on the loss that holds independent of any downstream task, as well as (ii) a criterion for robust classification in particular. We validate our theoretical results in practical applications, show how our results can be used for calibrating expectations of downstream robustness, and when our results are useful for optimal transfer learning. Taken together, our results offer an initial step towards characterizing the requirements of the representation function for reliable post-adaptation performance.[1]

1 Introduction

Recently, there has been a rise of models (BERT [31], GPT-3 [7], and CLIP [43]) trained on large-scale datasets which have found application to a variety of downstream tasks [6]. Such pretrained representations enable training a predictor, e.g., a linear "head", for substantially increased performance in comparison to training from scratch [10, 11], particularly when training data is scarce.

Generalizing to circumstances unseen in the training distribution is significant for real-world applications [2, 60, 29], and is crucial for the larger goal of robustness [27], where the aim is to build systems that can handle extreme, unusual, or adversarial situations. In this context, adversarial attacks [51, 5], i.e., inducing model failure via minimal perturbations to the input, provide a stress test for predictors, and have remained an evaluation framework where models trained on large-scale data have consistently struggled [53, 21, 22], particularly when, in training, adversarials are not directly accounted for [23, 40].

With that said, theoretical evidence [8] has been put forth that posits overparameterization is required for adversarial robustness, which is notable given its ubiquity in training on large-scale datasets. Thus, given the costs associated with large-scale deep learning, we are motivated to study how adversarial robustness transfers from pretraining to downstream tasks.

[1] https://github.com/lf-tcho/robustness_transfer

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

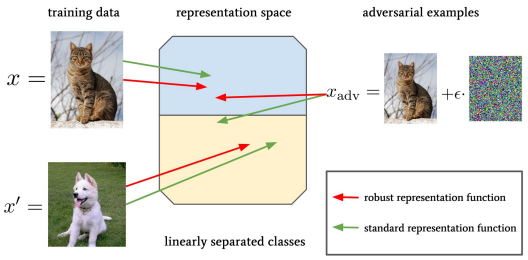


Figure 1: **Overview**. We distinguish between a robust (red) and non-robust (green) representation by its sensitivity to adversarial attacks, and use both for linear separation of the representation space, e.g., binary classification. This illustrates the dependence of the classifier on the representation's robustness, i.e., the predictor can only be robust if x_{adv} and x can be linearly separated from x' in representation space.

For this, considering the representation function to be the penultimate layer of the downstream predictor, we contribute theoretical results which bound the robustness of the downstream predictor by the robustness of the representation function. Fig. 1 illustrates the intuition behind our theory, and shows that a linear classifier's robustness is dependent on the robustness of the pretrained representation. Our theoretical results presented in Sec. 3 and Sec. 4 are empirically studied in Sec. 5, where we validate the results in practice, and in Sec. 6 to show how the theory can be used for calibrating expectations of the downstream robustness in a self-supervised manner [4]. In all, our theoretical results show how the quality of the underlying representation function affects robustness downstream, providing a lens on what can be expected upon deployment. We present the following contributions:

- we theoretically study the relationship between the robustness of the representation function, referred to as *adversarial sensitivity (AS) score*, and the prediction head. In Thm. 1, we prove a bound on the difference between clean and adversarial loss in terms of the AS score. In Thm. 2, we prove a sufficient condition for robust classification again conditioned on the robustness of the representation function.
- we evaluate the applicability and consistency of our theory in applications in Sec. 5.1 and Sec. 5.2. Furthermore, in Sec. 6, we illustrate the utility of the AS score to predict expected robustness transfer via linear probing.
- in Sec. 6.2, we study how linear probing compares to alternatives, like full finetuning, in robustness transfer for robust performance downstream.

2 Related Work

Transfer Learning: While empirical analyses of pretraining abound [33, 12, 62, 42, 3, 48], a number of theoretical analyses have also been contributed [56, 52, 19, 14, 49, 35, 38]. The majority of the theoretical studies aim to understand the benefit of transfer learning with respect to improved performance and data requirements compared to training from scratch. Therefore, the referenced studies include results on sample complexity [52, 19, 14, 49] and generalization bounds [38] for fine-tuning and linear probing in various scenarios. In this work, we primarily focus on the linear probing approach and perform a theoretical analysis to investigate its efficiency in transferring adversarial robustness.

Robustness: There is a direct connection between the mathematical formulation of adversarial robustness and the definition of a local Lipschitz constant of the network [25], but computing this quantity is intractable for general neural networks. However, there exists a line of work providing more practical approaches in certifiable robustness [54, 44, 15, 37, 36, 46], which can provide guarantees on the detection of adversarial examples. Relative to empirical approaches [23, 40, 18, 48, 55], certifiable methods have comparatively struggled to scale to datasets such as ImageNet [17]. Furthermore, likely due to computational expense, adversarial training has not yet been adopted by models trained on large-scale datasets [31, 7, 43]. Still, a number of works have incorporated adversarial training into

pretraining, in particular, self-supervised contrastive learning [32, 30, 9, 20]. With that said, our work is focused on the analysis of robustness irrespective of the pretraining approach.

Regarding analysis, [59] studied the trade-off between accuracy and robustness, and showed that, under a set of separation conditions on the data, a robust classifier is guaranteed to exist. Their work focuses on proving the robustness of a network itself and therefore requires Lipschitzness of the model. Further, [8] showed that overparameterization is needed for learning such smooth (i.e. Lipschitz) functions, which is critical for robustness. On the other hand, we specifically study the relationship between the robustness of the representation function and the prediction head, a scenario consistent with the practice of pretraining for transfer learning.

Transfer Learning & Robustness: The question of how robustness is transferred from pretraining to downstream task has been studied empirically in the prior literature [50, 58, 28, 63]. Notably, while [50] observe that robust representations underly robust predictors, [63] also observe that non-robust representations underly non-robust predictors. An empirical analysis of adversarial robustness transfer for the special case of linear probing is included in ,e.g., [50]. Our work not only derives theoretical results and concrete proof for the empirical phenomena reported in these studies but also provides further empirical investigation.

3 Robustness Bound

Our first theoretical result concerns the discrepancy between the standard and adversarial loss of a neural network f. The considered loss function ℓ is evaluated on a dataset of input samples $\mathbb{X} = \{x_1, \ldots, x_n\} \subset \mathcal{X} \subset \mathbb{R}^d$ and labels $\mathbb{Y} = \{y_1, \ldots, y_n\}$, where x_i is sampled from $\mathcal{X}$ with probability distribution $\mathcal{P}(\mathcal{X})$.
For the network, we consider a model $f : \mathbb{R}^d \to \mathbb{R}^c$ with a final linear layer, i.e. $f(x) = f_{W,\theta}(x) = Wg_\theta(x)$. The function $g_\theta : \mathbb{R}^d \to \mathbb{R}^r$ can be considered as a representation function for the input parameterized by θ, and $W \in \mathbb{R}^{c \times r}$ is the weight matrix of the linear layer. Note that this composition is both standard practice [10, 11], and has been extensively studied [56, 52, 19] in the literature.

For a given data pair (x_i, y_i), we are interested in the stability of the model's prediction of y_i with respect to (w.r.t.) perturbations of the input x_i as our robustness measure. Specifically, for $\delta \in \mathbb{R}^d$, we find the maximum loss $\ell(f(x_i + \delta), y)$ under a specified constraint $\|\delta\| < \epsilon$ based on some norm $\|\cdot\|$. Note that the defined measure is commonplace in the literature on adversarial attacks [51, 23].

We introduce the following notation for the standard loss $\mathcal{L}(f_{W,\theta})$ and the adversarial loss $\mathcal{L}_{\text{adv}}(f_{W,\theta})$,

$$\mathcal{L}(f_{W,\theta}) := \mathbb{E}_{\mathcal{P}(\mathcal{X})}[\ell(f_{W,\theta}(x), y)],$$
$$\mathcal{L}_{\text{adv}}(f_{W,\theta}) := \mathbb{E}_{\mathcal{P}(\mathcal{X})}\left[\max_{\|\delta\|<\epsilon} \ell(f_{W,\theta}(x+\delta), y)\right].$$

We aim to bound the effect of adversarial attacks in terms of their impact on the representation function, i.e. $\|g_\theta(x_i) - g_\theta(x_i + \delta)\|_2$. To obtain such a result for a general loss function ℓ, the variation in the loss must be bounded in relation to the proportional change in $f(x_i)$. This condition is formalized by the concept of Lipschitzness,

Definition 1. *A function $\ell : \mathbb{R}^d \to \mathbb{R}$ is C-Lipschitz in the norm $\|\cdot\|_\alpha$ on $\mathcal{R} \subset \mathbb{R}^d$, if*

$$|\ell(r) - \ell(r')| < C \|r - r'\|_\alpha,$$

for any $r, r' \in \mathcal{R}$.

The following theorem is valid for any loss function ℓ that satisfies Defn. 1 w.r.t. L_1, L_2 or L_∞ norm.

Theorem 1. *We assume that the loss $\ell(r, y)$ is C_1-Lipschitz in $\|\cdot\|_\alpha$, for $\alpha \in \{1, 2, \infty\}$, for $r \in f(\mathcal{X})$ with $C_1 > 0$ and bounded by $C_2 > 0$, i.e., $0 \leq \ell(r, y) \leq C_2$ $\forall r \in f(\mathcal{X})$. Then, for a subset $\mathbb{X}$ independently drawn from $\mathcal{P}(\mathcal{X})$, the following holds with probability of at least $1 - \rho$,*

$$\mathcal{L}_{\text{adv}}(f_{W,\theta}) - \mathcal{L}(f_{W,\theta}) \;\leq\; L_\alpha(W)\, C_1 \frac{1}{n} \sum_{i=1}^{n} \max_{\|\delta\|<\epsilon} \|g_\theta(x_i + \delta) - g_\theta(x_i)\|_2 \;+\; C_2 \sqrt{\frac{\log(\rho/2)}{-2n}},$$

where

$$L_\alpha(W) := \begin{cases} \|W\|_2 & , \textit{ if } \|\cdot\|_\alpha = \|\cdot\|_2, \\ \sum_i \|W_i\|_2 & , \textit{ if } \|\cdot\|_\alpha = \|\cdot\|_1, \\ \max_i \|W_i\|_2 & , \textit{ if } \|\cdot\|_\alpha = \|\cdot\|_\infty. \end{cases}$$

The proof of the theorem can be found in Appx. A.1. Note that the norm used to define the perturbation δ is arbitrary and only needs to be consistent on both sides of the bound.

Proof sketch: First, we apply Hoeffding's inequality, which requires a fixed function $f_{W,\theta}$, independently drawn samples $\{x_1, \ldots, x_n\}$, and an upper bound C_2 on the loss function. Next, we bound the average losses utilizing the C_1-Lipschitzness in $\|\cdot\|_\alpha$ of the loss function. Further, the Cauchy-Schwarz inequality is used to seperate the norms of W and the representation function g_θ.

Intuition: Thm. 1 shows that the effect an ϵ-perturbation has on the representation function g_θ indeed upper bounds the difference between the standard and adversarial loss. Thus, we characterize the relationship between how vulnerable the linear predictor and the underlying representation is to adversarial examples. Further, Thm. 1 formalizes the impact the weight matrix W has on robustness transfer and thereby provides a theoretical argument for regularization in this context.

The statement of Thm. 1 is formulated in terms of the expectation of the losses over $\mathcal{P}(\mathcal{X})$. In particular, the second summand on the right hand-side (RHS) of the bound illustrates the finite-sample approximation error between the expectation and the average over n samples in $\mathbb{X}$. The following lemma provides a version of the theorem for the average difference of the losses over a fixed finite sample $\mathbb{X}$, which can directly be concluded from the proof of Thm. 1 in Appx. A.1. We use this lemma to study our theoretic result in the experiments in Sec. 5.1.

Lemma 1 (Finite-sample version of Thm. 1)**.** *We assume that the loss $\ell(r, y)$ is C-Lipschitz in $\|\cdot\|_\alpha$, for $\alpha \in \{1, 2, \infty\}$, for $r \in f(\mathcal{X})$ with $C > 0$. Then, the following holds for any dataset $(\mathbb{X}, \mathbb{Y})$,*

$$\frac{\frac{1}{n}\sum_{i=1}^{n} \max_{\|\delta\|<\epsilon} \ell(f_{W,\theta}(x_i+\delta), y_i) - \ell(f_{W,\theta}(x_i), y_i)}{L_\alpha(W)\, C} \;\leq\; \frac{1}{n}\sum_{i=1}^{n} \max_{\|\delta\|<\epsilon} \|g_\theta(x_i+\delta) - g_\theta(x_i)\|_2.$$

Lemma 1 shows that the normalized difference between average clean and adversarial loss can be bound by the average robustness of the representation function on the RHS, which we denote as the representation function's *adversarial sensitivity (AS) score* on the dataset $\mathbb{X}$.

3.1 Valid Loss Functions

To elaborate on the applicability of the result, we discuss exemplary scenarios in which the assumptions of our theory are fulfilled in practice. As a first example, we consider a classification task where we use the softmax cross-entropy loss. While the cross-entropy loss itself is not Lipschitz continuous, since the logarithm is not, the softmax cross-entropy loss is indeed Lipschitz. We define the softmax cross-entropy loss as follows,

$$\ell_{CE}(f(x), y) = -\log\left(p_y(x)\right) = -\log\left(\frac{\exp\{f(x)_y\}}{\sum_{i=1}^{c}\exp\{f(x)_i\}}\right),$$

where $p_y(x)$ denotes the y-th entry of the softmax function applied to the final layer, $f(x)_y$ is the y-th entry of the model output with $f(x)_y = W_y g(x)$ and W_y being the y-th row of the matrix W. In Lemma 3 (in Appx. A), we prove that ℓ_{CE} is 2-Lipschitz in $\|\cdot\|_\infty$.

Further, the theory is also valid for regression tasks, where the mean square error (MSE)

$$\ell_{MSE}(f(x), y) = \|f(x) - y\|_2$$

is used to measure the loss. This loss function is 1-Lipschitz in $\|\cdot\|_2$, since, by the reverse triangle property, it holds that

$$|\|f(x_1) - y\|_2 - \|f(x_2) - y\|_2| \leq \|f(x_1) - f(x_2)\|_2.$$

Additional losses which satisfy Defn. 1 include the logistic, hinge, Huber and quantile loss [13].

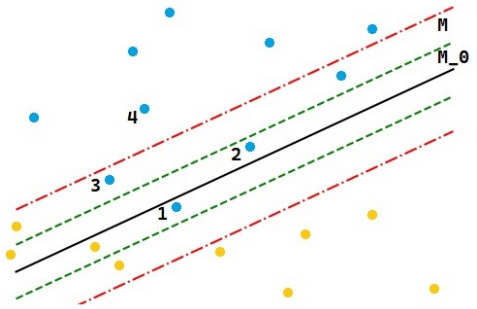


Figure 2: The graphic shows two classes of datapoints and a linear classifier's decision boundary (black). The line M_0 (green) is the optimal margin defining the boundary between a successful or failed adversarial attack. Line M (red) represents the margin derived from Thm. 2 which is an upper bound on M_0. Four datapoints from the blue class are highlighted: 1 is wrongly classified, 2 is a non-robust example, 3 is robust but does not fulfill equation (1), and 4 is a robust classification fulfilling the bound in (1).

4 Robustness Classification Criterion

In Sec. 3, we presented a theoretical result independent of any particular downstream task and valid for various loss functions. Here, we focus on classification tasks, and, in this context, we provide an additional statement that studies the robustness of a linear classifier w.r.t. adversarial attacks. We refer to a classifier as robust if there does not exist a bounded input perturbation that changes the classifier's prediction.

In particular, we consider the classifier

$$c(x) = \operatorname{argmax}_i f(x)_i,$$

where $f(x) = W g_\theta(x)$ is the model as defined in Sec. 3. The following theorem provides insights on how the robustness of the underlying representation function g_θ impacts the robustness of a linear classifier. Specifically, it formally describes how the robustness of the representation, i.e. the AS score on a datapoint x, determines the set of weight matrices which yield a robust classifier.

Theorem 2. *The classifier $c(x)$ is robust in x, i.e., $c(x+\delta) = c(x)$ for $\delta \in \mathbb{R}^d$ with $\|\delta\| < \epsilon$, if for $x \in \mathcal{X}$ with $c(x) = y$ it holds that*

$$\|g_\theta(x+\delta) - g_\theta(x)\|_2 \;\leq\; \min_{j \neq y} \frac{|f(x)_y - f(x)_j|}{\|W_y - W_j\|_2}. \tag{1}$$

The proof of the theorem can be found in Appx. A.2.

Proof sketch: To prove robustness, we need to show that $f(x+\delta)_y > f(x+\delta)_j$ if $f(x)_y > f(x)_j$. The proof consists of rewriting this condition and applying the Cauchy-Schwarz inequality to separately bound in terms of the difference in W and g_θ.

Intuition: Thm. 2 shows that the effect an ϵ-perturbation has on the representation function g_θ determines the necessary class separation by the linear predictor for robust classification to be ensured. Specifically, the RHS of Thm. 2 can be understood as a normalized margin between class logits for x, where the classification margin refers to the difference between the classification score for the true class and maximal classification score from all false classes. Thus, the theorem states that if the margin is larger than the effect of an ϵ-perturbation on the representation function, then the classifier is robust. Note that if a datapoint does not satisfy the bound, i.e. LHS is greater than the RHS, the robustness of the classifier is ambiguous. In this case, the RHS margin is not the exact separation, or the optimal margin, between robust and non-robust classification. Fig. 2 provides intuition for this interpretation of the result in a binary classification scenario.

5 Empirical Study of Theory

The following experiments serve to validate the theoretical results in practical applications. For the finite sample version of Thm. 1, Lemma 1, we consider a classification task in Sec. 5.1, and a regression task in Appx. B.2. In Sec. 5.2, we study the robustness condition established in Thm. 2.

In the following sections, we train a linear predictor on top of a pretrained representation function for a specific downstream task, i.e., we perform transfer learning via linear probing. To study robustness transfer, we use robustly pretrained representation functions and examine if the transferred robustness is consistent with the theoretical bounds. As our robust representation function, we use models robustly pretrained on CIFAR-100 [1, 34], and ImageNet [47, 17]. Both models are available on RobustBench [16] and are adversarially trained using the softmax cross-entropy loss and L_∞-attacks. For our downstream tasks, we consider CIFAR-10 [34], Fashion-MNIST [57] and Intel Image[2]. Further experimental details can be found in Appx. B.1.

5.1 Classification - Cross-Entropy Loss

Here, we empirically evaluate Lemma 1, the finite sample version of Thm. 1, on classification tasks. As per Sec. 3.1, for softmax cross-entropy, we can apply Lemma 1 with Lipschitz constant $C = 2$ and mapping function $L_\infty(W)$.

In Tab. 1, we present the left-hand side (LHS), the scaled difference between clean and robust cross-entropy (CE) loss, and the right-hand side (RHS), the representation function's AS score, of Lemma 1. Note that, for LHS, the attack on f maximizes the CE loss, while for RHS, the attack on g_θ maximizes the impact on the representation function, i.e. $\|g_\theta(x_i) - g_\theta(x_i + \delta)\|_2$. See Appx. B.1 for further details.

The results in Tab. 1 validate our theoretical predictions in practice, as the LHS is indeed always upper bounded by the RHS. Furthermore, we see signs of correlation between the LHS and RHS, i.e. significant changes on one side of the bound are accompanied by similar changes on the other. Tab. 2 further highlights the connection between our result and weight norm regularization, as it confirms that a smaller weight norm, i.e., smaller $L_\alpha(W)$, corresponds to a smaller decrease in robustness, i.e., a smaller difference between the losses.

Table 1: For the CIFAR-100 and ImageNet pretraining datasets, the LHS and RHS, i.e., the AS score, of Lemma 1 are computed for each of the CIFAR-10, Fashion-MNIST and Intel Image downstream tasks. Additionally, $L_\infty(W)$ and the absolute difference between clean and adversarial CE loss are provided.

	CIFAR-100			ImageNet		
	CIFAR-10	F-MNIST	Intel Image	CIFAR-10	F-MNIST	Intel Image
Diff. CE	0.74	0.8	0.47	4.14	9.4	0.14
$L_\infty(W)$	4.4	4.6	3.1	1.9	2.0	1.4
LHS	0.08	0.09	0.08	1.11	2.33	0.05
AS score	0.98	2.11	1.05	15.7	53.2	2.75

Table 2: The same datasets and metrics are considered as in Tab. 1. Here, however, L2 weight matrix regularization ($\lambda = 0.01$) is employed for all downstream tasks.

	CIFAR-100			ImageNet		
	CIFAR-10	F-MNIST	Intel Image	CIFAR-10	F-MNIST	Intel Image
Diff. CE	0.46	0.51	0.38	3.57	7.3	0.14
$L_\infty(W)$	2.1	2.3	2.2	1.4	1.5	1.2
LHS	0.11	0.11	0.09	0.81	2.43	0.06
AS score	0.98	2.11	1.05	15.7	53.2	2.75

[2] `https://www.kaggle.com/datasets/puneet6060/intel-image-classification`

5.2 Classification Criterion

Here, we focus on empirically analysing Thm. 2, which specifies an upper bound on the representation's sensitivity against ϵ-perturbations, where the condition ensures robustness of the classifier against local perturbations for each datapoint.

In Tab. 3, we observe that indeed: if condition (1) is fulfilled, the classification is always robust, as can be seen from "rob. fulfilled". With that said, we do observe instances of robust classification which do not fulfill Eq. (1), which we proved ensures robustness, as can be seen from comparing "robust accuracy" to "prop. fulfilled". As with Lemma 1, we see that tighter results would be necessary for identifying all transferred robustness. Furthermore, as discussed in Sec. 4, since datapoints fulfilling our bound are even further away from the classification boundary than necessary, we also find that fulfilling the bound is predictive of robustness to stronger attacks, as can be seen in Appx. B.3.

For the scenario of pretraining on ImageNet and transferring to CIFAR-10 or Fashion-MNIST, almost no datapoint from the test set fulfills condition (1). This observation is accompanied by a steep drop from clean accuracy to robust accuracy, which suggests this is likely a consequence of the large distribution shift between ImageNet and the lower-resolution images of CIFAR-10 (32×32) and Fashion-MNIST (28×28) resized to ImageNet resolution (256×256). Note that this distribution shift is evident regardless of how the images are scaled before applying the ImageNet representation function, see Appx. B.5. In Tab. 1, the correlation between distribution shift and loss of robustness is also captured prior to transfer learning by a large AS score, and Appx. B.5 provides further insight into this connection.

Table 3: For the CIFAR-100 and ImageNet pretraining datasets, we report clean & robust accuracy, the proportion of images fulfilling the theoretical bound, and, of the images fulfilling the theoretical bound, the proportion of images that are robust for each of the CIFAR-10, Fashion-MNIST and Intel Image downstream tasks.

	CIFAR-100			ImageNet		
	CIFAR-10	F-MNIST	Intel Image	CIFAR-10	F-MNIST	Intel Image
clean accuracy	70.9%	84%	78.5%	92.4%	87.2%	91.2%
robust accuracy	42.2%	56.2%	56.8%	15.8%	0%	85.3%
prop. fulfilled	13.4%	10.1%	26.3%	0.1%	0%	56.2%
rob. fulfilled	100%	100%	100%	100%	-	100%

6 Adversarial Sensitivity Score

In the prior sections, we empirically validated that our theoretical results and required assumptions are realistic and hold in applications. Note that, in both of our results, the robustness of the representation function determines the robustness that transfers to downstream tasks. Specifically, Lemma 1 is stated using the *adversarial sensitivity (AS) score* of the representation function on a dataset $\mathbb{X}$,

$$\frac{1}{n}\sum_{i=1}^{n} \max_{\|\delta\|<\epsilon} \|g_\theta(x_i+\delta) - g_\theta(x_i)\|_2, \tag{2}$$

to bound the difference between clean and adversarial loss. Before specifying a downstream task, we can compute the representation function's adversarial sensitivity score on a set of unlabeled input images. In the following, we explore how to calibrate expectations on the robustness upon transfer of the representation function via linear probing to a particular downstream task based on the AS score.

6.1 Robustness Transfer via Linear Probing

Here, we explore how the AS score can be used to provide valuable insights on the transferability of robustness with linear probing. In Tab. 4, we find that if the AS score on the unlabeled pretraining and downstream datasets are similarly small, then it is reasonable to expect that robustness will transfer via linear probing. However, if the AS score on the downstream task is much higher, it is reasonable to expect that robustness will not transfer via linear probing.

In Tab. 4, we compare the AS score to the relative difference between clean and robust CE loss on each dataset, i.e., (CE - robust CE)/CE. Indeed, the similar AS scores of CIFAR-100, CIFAR-10, and Intel Image upon pretraining on CIFAR-100 correspond to a successful transfer of robustness, whereas the dissimilar adversarial sensitivity scores of ImageNet, Fashion-MNIST, and CIFAR-10 upon pretraining on ImageNet reflect the striking loss of robustness upon transfer. One explanation for a larger AS score is a shift in distribution between pretraining and downstream data. This connection is confirmed in an additional empirical analysis in Appx. B.5.

Table 4: For the CIFAR-100 and ImageNet pretraining datasets, we report the AS score (Eq. (2)), clean & robust cross-entropy loss, and the relative difference between said losses for the pretraining dataset, as well as the CIFAR-10, Fashion-MNIST and Intel Image downstream tasks.

Pretraining	Downstream	AS score	clean /robust CE	relative dif.
CIFAR-100	CIFAR-100	1.0	2.71 / 3.27	0.21
	CIFAR-10	0.98	0.82 / 1.55	0.9
	Fashion-MNIST	2.11	0.45 / 1.25	1.8
	Intel Image	1.05	0.61 / 1.08	0.76
ImageNet	ImageNet	1.27	1.38 / 1.64	0.19
	CIFAR-10	15.7	0.22 / 4.36	18.8
	Fashion-MNIST	53.2	0.36 / 9.76	26.1
	Intel Image	2.75	0.25 / 0.39	0.57

6.2 Beyond Linear Probing?

In the prior section, we observe a correlation between AS score and the difference between clean and robust CE, i.e., the AS score is predictive of robustness transfer via linear probing. However, linear probing is not the only transfer learning approach, what can we say about alternatives?

In finetuning, the representation function is changed, and thus also the AS score changes post-transfer. Thus, we can expect the AS score computed pre-transfer is less predictive of the downstream robustness for finetuning than for linear probing. To gain a better understanding, we study the difference between clean/robust accuracy for linear probing (LP), full finetuning (FT) as well as linear probing then finetuning (LP-FT) [35], where LP-FT was proposed to combine the benefits of the two approaches. The main results can be found in Tab. 5, and additional details about the experiment are provided in Appx. B.4.

Tab. 5 shows the performance of each transfer learning method applied to the model robustly pretrained on CIFAR-100 or ImageNet. On CIFAR-100, LP preserves the most robust accuracy, whereas LP-FT performs better on ImageNet. Note that the AS score for LP is highlighted, as this is the score we can obtain prior to transfer learning, which allows us to calibrate expectations on downstream performance.

We find no correlation between LP and FT performance, and do not observe a predictive relationship between AS score pre-transfer and robustness transfer via FT. Instead, we observe successful robustness transfer via FT and LP-FT for cases with high AS score pre-transfer. This provides us with a first indication that the adversarial sensitivity score may point to scenarios where FT preserves more robustness downstream than LP.

7 Discussion

Significance: If the representation does not change at all in response to ϵ-perturbations, it may seem immediately clear that any predictor operating on said representation is robust. However, if the representation function *is* variant to ϵ-perturbations, how much robustness is lost as a result is dependent on the nature of the predictor function under consideration and the distribution shift in the data from pretraining to target task. In this work, we consider a linear predictor, and formally compute the exact effect this function can have on the test-time robustness.

Comparison of the theorems: The theoretical statement in Thm. 1 is closely related to the one in Thm. 2, as a more robust predictor to adversarial attacks leads to a smaller difference in the loss

Table 5: Clean accuracy (Acc) and robust accuracy (RAcc) for the three transfer learning approaches using the model robustly pretrained on CIFAR-100.

CIFAR-100	CIFAR-10			Fashion-MNIST			Intel Image		
Method	Acc	RAcc	AS	Acc	RAcc	AS	Acc	RAcc	AS
FT	**94.2%**	3.9%	8.93	**95.1%**	40%	6.68	**90.1%**	7.5%	6.86
LP	70.9%	**42.2%**	**0.98**	84%	**56.2%**	**2.11**	78.5%	**56.8%**	**1.05**
LP-FT	93.9%	1.4%	6.44	95%	36%	4.27	89.2%	27.1%	3.02
ImageNet	CIFAR-10			Fashion-MNIST			Intel Image		
Method	Acc	RAcc	AS	Acc	RAcc	AS	Acc	RAcc	AS
FT	**98.6%**	38.4%	12.6	**95.7%**	63.5%	9.0	**93.9%**	85.8%	3.8
LP	92.4%	15.8%	**15.7**	87.2%	0%	**53.2**	91.2%	85.3%	**2.7**
LP-FT	98.5%	**44.5%**	9.8	95.6%	**73.9%**	5.1	**93.9%**	**85.9%**	3.9

between clean and adversarial examples, and vice versa. Thm. 1 provides a bound on the robustness based on the difference of the loss function, while Thm. 2 only considers the specific form of the classifier and thus information about and restrictions on the used loss function is unnecessary. However, studying the loss function makes Thm. 1 valid in a much broader set of scenarios, and further provides reasoning on why using weight norm regularization during linear probing improves robustness transfer, see Sec. 5.1.
The bound in Thm. 2, on the other hand, provides insights on how robustness of a classifier can be guaranteed. In particular, one could use our results to derive robustness guarantees post-transfer from attacks on the representation function pre-transfer.

Clarification: Note that we consider a model f as a composition of a linear layer and a representation function, i.e., $f(x) = Wg_\theta(x)$, but do not specify how the parameters of f are learned. As our motivating example, the representation function g_θ is optimized on a pretraining task and subsequently W is trained w.r.t. a downstream task. However, one could also approach the downstream task with the classical strategy of training the model end-to-end instead, i.e., optimizing $\tilde{W}$ and $\tilde{\theta}$ simultaneously in $f_{\tilde{W},\tilde{\theta}} = \tilde{W}g_{\tilde{\theta}}$, assuming the same architecture as before. The parameters θ, W and $\tilde{\theta}, \tilde{W}$ obtained from the two strategies may be very different. For example, the representation function g_θ may not be as useful for linearly separating the dataset into c classes as $g_{\tilde{\theta}}$, which was optimized specifically for this task. This observation is described as a task mismatch between pretraining and the supervised downstream task.
Our contributed theoretical results demonstrate that we can study robustness by studying the learned functions themselves, irregardless of how said functions were optimized in the first place. With that said, this does not mean that the aforementioned task mismatch does not affect the model's robustness. Thm. 1's bound will loosen if, e.g., g_θ does not enable linear separation of the dataset into c classes, and if the AS score of the representation function increases due to this task mismatch or distribution shift in the data. We provide a study on the impact of the distribution shift in the data on the AS score in Appx. B.5.

Finally, while the linear probing protocol is standard practice [10, 11], alternative methodology exists for leveraging a representation function for a specified downstream task. For one, we could consider a model f as a composition of a nonlinear function h and a representation function, i.e. $f(x) = h \circ g_\theta(x)$, a scenario our theoretical statements do not account for. Furthermore, alternative approaches can include updating the parameters of the representation function g_θ, e.g., full finetuning. As suggested by the results in Sec. 6.2, while a high AS score may not be predictive of robustness transfer for finetuning, it may instead indicate when finetuning is required to transfer robustness. Overall, the theoretical dependencies between the robustness of the representations and transfer learning alternatives to linear probing present interesting opportunities for future analysis.

8 Conclusion

We prove two theorems which demonstrate that the robustness of a linear predictor on downstream tasks can be bound by the robustness of its underlying representation. We provide a result applicable to various loss functions and downstream tasks, as well as a more direct result for classification. Our

theoretical insights can be used for analysis when representations are transferred between domains, and sets the stage for carefully studying the robustness of adapting pretrained models.

Acknowledgements

We thank the ICLR 2022 Diversity, Equity & Inclusion Co-Chairs, Rosanne Liu and Krystal Maughan, for starting the CoSubmitting Summer (CSS) program, through which this collaboration began. We thank the ML Collective community for the ongoing support and feedback of this research. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting YS. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A.

References

[1] Sravanti Addepalli, Samyak Jain, et al. Efficient and effective augmentation strategy for adversarial training. *Advances in Neural Information Processing Systems*, 35:1488–1501, 2022.

[2] Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 45(3):1150–1158, 2018.

[3] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

[4] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[5] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[13] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with lipschitz and convex loss functions. *Probability Theory and related fields*, 176(3-4):897–940, 2020.

[14] Kurtland Chua, Qi Lei, and Jason D Lee. How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[15] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[16] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[17] Jia Deng, Wei Dong, Richard 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*, pages 248–255. Ieee, 2009.

[18] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020.

[19] Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

[20] Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[21] Stanislav Fort. Adversarial examples for the openai clip in its zero-shot classification regime and their semantic generalization, Jan 2021.

[22] Stanislav Fort. Adversarial vulnerability of powerful near out-of-distribution detection. *arXiv preprint arXiv:2201.07012*, 2022.

[23] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[25] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in Neural Information Processing Systems*, 30, 2017.

[26] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HJz6tiCqYm`.

[27] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[28] De-An Huang, Zhiding Yu, and Anima Anandkumar. Transferable unsupervised robust representation learning, 2021.

[29] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[30] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in Neural Information Processing Systems*, 33:16199–16210, 2020.

[31] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2, 2019.

[32] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020.

[33] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[34] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[35] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

[36] Alexander Levine, Sahil Singla, and Soheil Feizi. Certifiably robust interpretation in deep learning. *CoRR*, abs/1905.12105, 2019.

[37] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *Advances in neural information processing systems*, 32, 2019.

[38] Dongyue Li and Hongyang Zhang. Improved regularization and robustness for fine-tuning in neural networks. *Advances in Neural Information Processing Systems*, 34:27249–27262, 2021.

[39] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Skq89Scxx`.

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[41] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[42] Matthew E Peters, Sebastian Ruder, and Noah A Smith. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, 2019.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[44] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

[45] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017.

[46] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

[47] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

[48] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

[49] Gal Shachaf, Alon Brutzkus, and Amir Globerson. A theoretical analysis of fine-tuning with linear teachers. *Advances in Neural Information Processing Systems*, 34, 2021.

[50] Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. In *International Conference on Learning Representations*, 2020.

[51] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.

[52] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852–7862, 2020.

[53] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[54] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

[55] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

[56] Sen Wu, Hongyang R. Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020.

[57] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[58] Yutaro Yamada and Mayu Otani. Does robustness on imagenet transfer to downstream tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9215–9224, 2022.

[59] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

[60] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[61] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

[62] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

[63] Jiaming Zhang, Jitao Sang, Qi Yi, Yunfan Yang, Huiwen Dong, and Jian Yu. Imagenet pre-training also transfers non-robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3436–3444, 2023.

A Proofs and Auxiliary Results

We state two auxiliary results first and subsequently provide the full proofs for Thm. 1 and Thm. 2 in Appx. A.1 and Appx. A.2.

Lemma 2. *For a matrix $W \in \mathbb{R}^{c\times r}$ and a vector $g \in \mathbb{R}^r$, the following bounds hold,*

1. $\|Wg\|_2 \leq \|W\|_2\|g\|_2$,
2. $\|Wg\|_1 \leq \sum_{i=1}^c \|W_i\|_2\|g\|_2$,
3. $\|Wg\|_\infty \leq \max_i \|W_i\|_2\|g\|_2$.

Proof.

1. $\|Wg\|_2 = \sqrt{\sum_{i=1}^c (W_i g)^2} \leq \sqrt{\sum_{i=1}^c (\|W_i\|_2\|g\|_2)^2} = \|W\|_2\|g\|_2$, where the second step follows by Cauchy-Schwartz inequality.
2. $\|Wg\|_1 = \sum_{i=1}^c |W_i g| \leq \sum_{i=1}^c \|W_i\|_2\|g\|_2 = \sum_{i=1}^c \|W_i\|_2\|g\|_2$, where the second step follows again by Cauchy-Schwartz inequality.
3. $\|Wg\|_\infty = \max_i |W_i g| \leq \max_i \|W_i\|_2\|g\|_2$, where the second step follows by Cauchy-Schwartz inequality.

□

The following Lemma proves a Lipschitz condition for the cross-entropy-softmax loss.

Lemma 3. *We define the cross-entropy loss with softmax function for a vector $f \in \mathbb{R}^c$ and a class index $y \in \{1,\dots,c\}$, i.e.,*

$$\ell(f,y) := -\log\left(p_y(f)\right) = -\log\left(\frac{\exp\{f_y\}}{\sum_{i=1}^c \exp\{f_i\}}\right).$$

Then, the following Lipschitz condition holds for $\ell(f,y)$,

$$|\ell(a,y) - \ell(b,y)| \leq \; 2\,\|a-b\|_\infty,$$

for any vectors $a, b \in \mathbb{R}^c$ and an arbitrary class index $y \in \{1,\dots,c\}$.

Proof.

$$\begin{aligned}
|\ell(a,y) - \ell(b,y)| &= \; |-\log\Big(\frac{\exp\{a_y\}}{\sum_{j=1}^c \exp\{a_j\}}\Big) + \log\Big(\frac{\exp\{b_y\}}{\sum_{j=1}^c \exp\{b_j\}}\Big)| \\
&= \; |\log\Big(\exp\{b_y - a_y\}\Big) - \log\Big(\frac{\sum_{j=1}^c \exp\{b_j\}}{\sum_{j=1}^c \exp\{a_j\}}\Big)| \\
&= \; |\big(b_y - a_y\big) - \log\Big(\sum_{j=1}^c \frac{\exp\{a_j\}}{\sum_{l=1}^c \exp\{a_l\}} \exp\{b_j - a_j\}\Big)| \\
&\leq \; |\big(b_y - a_y\big) - \sum_{j=1}^c p_j(a)\big(b_j - a_j\big)|,
\end{aligned}$$

where $p_j(a) := \frac{\exp\{a_j\}}{\sum_{l=1}^c \exp\{a_l\}}$ denotes the soft-max output of the vector a. The inequality in the last line follows by Jensen's inequality, since $-\log$ is convex and the soft-max probabilities p_j sum up to one.

$$|\big(b_y - a_y\big) - \sum_{j=1}^c p_j(a)\big(b_j - a_j\big)| \leq \; |b_y - a_y| + \sum_{j=1}^c p_j(a)|b_j - a_j| \leq \; 2\,\|b-a\|_\infty,$$

since both summands in the middle can be bounded by the supremum norm. □

A.1 Proof of Theorem 1

Proof of Theorem 1. We use Hoeffding's inequality on the independent random variables $D_1 \ldots, D_n$, which are defined as

$$D_i := \max_{\|\delta\|<\epsilon} \ell(f(x_i+\delta), y_i) - \ell(f(x_i), y_i),$$

based on independently drawn data points with probability distribution $\mathcal{P}(\mathcal{X})$, where the observations are $x_i \in \mathcal{X}$ with corresponding labels y_i. We can use the upper bound on ℓ to obtain that $0 \leq D_i \leq C_2$ and conclude that

$$\begin{aligned} & \mathbb{P}\left(\Big| \sum_{i=1}^n D_i - n\mathbb{E}[D] \Big| \geq t \right) \leq 2 \cdot \exp\left(\frac{-2t^2}{nC_2^2} \right) \\ \Rightarrow \quad & \mathbb{P}\left(\Big| \frac{1}{n} \sum_{i=1}^n D_i - \mathbb{E}[D] \Big| \leq C_2 \sqrt{\frac{\log(\rho/2)}{-2n}} \right) \geq 1-\rho. \end{aligned}$$

Thus, with probability of at least $1-\rho$ it holds that

$$\begin{aligned} \mathbb{E}[D] &= \Big| \mathcal{L}_{\text{adv}}(f) - \mathcal{L}(f) \Big| = \Big| \mathbb{E}_{(x,y)} \left[\max_{\|\delta\|<\epsilon} \ell(f(x+\delta), y) - \ell(f(x), y) \right] \Big| \\ &\leq \Big| \frac{1}{n} \sum_{i=1}^n \max_{\|\delta\|<\epsilon} \ell(f(x_i+\delta), y_i) - \ell(f(x_i), y_i) \Big| + C_2 \sqrt{\frac{\log(\rho/2)}{-2n}}. \end{aligned}$$

We can further bound the first term on the right-hand side, since the loss function $\ell(r, y)$ is C_1-Lipschitz in $\|\cdot\|_\alpha$ for $r \in f(\mathcal{X})$ and $f(x) = Wg_\theta(x)$.

$$\begin{aligned} & \Big| \frac{1}{n} \sum_{i=1}^n \max_{\|\delta\|<\epsilon} \ell(Wg_\theta(x_i+\delta), y_i) - \ell(Wg_\theta(x_i), y_i) \Big| \\ &= \Big| \frac{1}{n} \sum_{i=1}^n \big| \ell(Wg_\theta(x_i+\delta_i), y_i) - \ell(Wg_\theta(x_i), y_i) \big| \Big| \\ &\leq C_1 \frac{1}{n} \sum_{i=1}^n \|Wg_\theta(x_i+\delta_i) - Wg_\theta(x_i)\|_\alpha, \end{aligned}$$

where $\delta_1, \ldots, \delta_n \in \mathcal{X}$ with $\|\delta_i\| < \epsilon$ are chosen appropriately. In a next step, we apply Lemma 2 and the definition of L_α in the theorem statement,

$$C_1 \frac{1}{n} \sum_{i=1}^n \|Wg_\theta(x_i+\delta_i) - Wg_\theta(x_i)\|_\alpha \;\leq\; L_\alpha(W)\, C_1 \frac{1}{n} \sum_{i=1}^n \|g_\theta(x_i+\delta_i) - g_\theta(x_i)\|_2.$$

The δ_i's might not be the same that maximize the distance between the representations but their effect can be upper bounded by the maximum. Combining the derived bounds and using that $g_\theta(x_i+\delta_i) \leq \max_{\|\delta\|<\epsilon} g_\theta(x_i+\delta)$, we obtain our result.

$$\begin{aligned} \mathcal{L}_{\text{adv}}(f) - \mathcal{L}(f) &= \Big| \mathcal{L}_{\text{adv}}(f) - \mathcal{L}(f) \Big| \\ &\leq\; L_\alpha(W)\, C_1 \frac{1}{n} \sum_{i=1}^n \max_{\|\delta\|<\epsilon} \|g_\theta(x_i+\delta) - g_\theta(x_i)\|_2 + C_2 \sqrt{\frac{\log(\rho/2)}{-2n}}. \end{aligned}$$

□

A.2 Proof of Theorem 2

Proof of Theorem 2. We want to proof that the classifier $c(x)$ is robust in $x \in \mathcal{X}$, i.e., it holds that $c(x+\delta) = c(x)$ for all $\delta \in \mathbb{R}^d$ with $\|\delta\| < \epsilon$. The classifier predicts class y, i.e., $c(x) = y$, if the y-th value of the model $f(x)$ is larger than all other entries, i.e., $f(x)_y > f(x)_j$ for all $j \neq y$. We

denote the y-th row of the linear layer's matrix W by W_y, thus $f(x)_y = W_y g_\theta(x)$.
For an input x with $c(x) = y$, we want to prove that $c(x+\delta) = y$ for any δ with $\|\delta\| < \epsilon$., i.e.,

$$f(x+\delta)_y > f(x+\delta)_j, \text{ if } f(x)_y > f(x)_j.$$

It holds that

$$f(x+\delta)_y = f(x+\delta)_j + (W_j(g_\theta(x) - g_\theta(x+\delta))) + (W_y g_\theta(x) - W_j g_\theta(x)) + (W_y(g_\theta(x+\delta) - g_\theta(x))).$$

Thus, for $f(x+\delta)_y > f(x+\delta)_j$ to be true, it is necessary that

$$\begin{aligned} & (W_j(g_\theta(x) - g_\theta(x+\delta))) + (W_y g_\theta(x) - W_j g_\theta(x)) + (W_y(g_\theta(x+\delta) - g_\theta(x))) \geq 0 \\ \iff \quad & W_y g_\theta(x) - W_j g_\theta(x) \geq (W_y - W_j)(g_\theta(x) - g_\theta(x+\delta))). \end{aligned} \tag{3}$$

By Cauchy-Schwartz inequality follows that

$$(W_y - W_j)(g_\theta(x) - g_\theta(x+\delta))) \leq |(W_y - W_j)(g_\theta(x) - g_\theta(x+\delta))| \leq \|W_y - W_j\|_2 \cdot \|g_\theta(x+\delta) - g_\theta(x)\|_2.$$

Since further $W_y g_\theta(x) - W_j g_\theta(x) = f(x)_y - f(x)_j$, the condition of equation (3) holds for any $j \neq y$ if

$$f(x)_y - f(x)_j = |f(x)_y - f(x)_j| \geq \|W_y - W_j\|_2 \cdot \|g_\theta(x+\delta) - g_\theta(x)\|_2.$$

For $c(x+\delta) = c(x)$ to be true, the above condition needs to be true for all $j \neq y$ and this holds if

$$\min_{j \neq y} \frac{|f(x)_y - f(x)_j|}{\|W_y - W_j\|_2} \geq \max_{\|\delta\| < \epsilon} \|g_\theta(x+\delta) - g_\theta(x)\|_2.$$

□

B Additional Experimental Results

We provide additional information on the datasets used and the employed hyperparameters in Appx. B.1. A further application of Thm. 1 is presented in Appx. B.2, where robustness transfer is analysed for a regression task. In Appx. B.3, we extend the analysis from Sec. 6 by evaluating the robustness against stronger attacks. In Appx. B.4, an empirical study compares the robustness transfer depending on the chosen transfer learning method. Finally, in Appx. B.5, we investigate the effect of a shift in distribution between pretraining and downstream task data on the transferred robustness and the AS score.

B.1 Experimental Setup for Section 5

For our experiments, we consider two models robustly pretrained on CIFAR-100 [34] and ImageNet [17]. CIFAR-100 consists of 32x32 RGB images with 100 different classes. ImageNet-1K consists of 256x256 RGB images of 1,000 different classes.
We investigate the performance of robustness transfer on CIFAR-10 [34], Fashion-MNIST [57] and Intel Image Classification[3]. The CIFAR-10 dataset consists of 32x32 RGB images in 10 classes. Intel Image Classification consists of 150x150 RGB images of 6 natural scenes (buildings, forest, glacier, mountain, sea, and street). Fashion-MNIST is a dataset of Zalando's article images, where each example is a 28x28 grayscale image, associated with a label from 10 classes.

We use two models from RobustBench [16], namely robust models trained on CIFAR-100 and ImageNet. For CIFAR-100, we use a WideResNet-34-10 (≈46M parameters) architecture [61] which is adversarially trained using a combination of simple and complex augmentations with separate batch normalization layers [1]. For ImageNet, we use a WideResNet-50-2 (≈66M parameters) architecture [61] which is is also adversarially trained [47, 40].

For the model pretrained on CIFAR-100, run 20 epochs of linear probing (LP) using a batch size of 128, and resize all input images to 32x32. For the model pretrained on ImageNet, we run 10 epochs of LP using a batch size of 32, and resize all input images to 256x256. We reduce the batch size to compensate for the increase in input resolution. For LP, we set the learning rate using a cosine

[3] `https://www.kaggle.com/datasets/puneet6060/intel-image-classification`

Layer	Layer Type	Activation and Normalization
1	Convolutional layer (c=6, k=5)	ReLu, Average Pooling, and Batch Norm
2	Convolutional layer (c=16, k=5)	ReLu, Average Pooling, and Batch Norm
3	Linear Layer (out=120)	ReLu and Batch Norm
4	Linear Layer (out=84)	ReLu and Batch Norm
5	Linear Layer (out=84)	ReLu and Batch Norm
"head"	Linear Layer (out=1)	-

Table 6: The table presents the model architecture used in experiments on MSE loss and the dSprites dataset. The convolutional layers are provided with number of output channels and kernel size, i.e., (c=channels, k=kernel_size), where as the linear layers are listed together with the size of their output vector.

annealing schedule [39], with an initial learning rate of 0.01, and use stochastic gradient descent with a momentum of 0.9 on the cross-entropy loss.

To compute LHS and RHS, we must attack the cross-entropy loss on f as well as the MSE on g_θ. In both cases, we employ L_∞-PGD attacks implemented in the foolbox package [45]. As hyperparameters for the L_∞-PGD attack, we choose 20 steps and a relative step size of 0.7, since this setting yields highest adversarial sensitivity (AS) scores over all the tasks. The attack strength ϵ is set to $\epsilon = 8/255$ for attacking the CIFAR-100 pretrained model and $\epsilon = 1/255$ for the ImageNet model. Thus, we use the same ϵ on CIFAR-100 as used during the RobustBench pre-training and reduce the ϵ on ImageNet due to the constant pixel blocks introduced after resizing CIFAR-10 and Fashion-MNIST images from 32x32 and 28x28 to 256x256, respectively.

The experiments were run on a single GeForce RTX 3080 GPU and for each downstream task transfer learning or theory evaluation took between 30 minutes to 3 hours.

B.2 Regression - MSE

Here, we empirically evaluate Lemma 1 on regression tasks using the mean squared error (MSE) as the scoring criterion. As per Sec. 3.1, we can apply Lemma 1 with Lipschitz constant $C = 1$ and mapping function $L_2(W)$.

We consider the dSprites [41] dataset, which consists of 64x64 grayscale images of 2D shapes procedurally generated from 6 factors of variation (FoVs) which fully specify each sprite image: color, shape, scale, orientation, x-position (posX) and y-position (posY). The dataset consists of 737,280 images, of which we use an 80-20 train-test split. Here, we pre-train for the regression task of decoding an FoV from an image, i.e., we optimise the network f to minimize MSE with the exact orientation of the sprite. For LP, we consider 4 downstream tasks, each corresponds to decoding a unique FoV, while one matches the task used for pre-training.

The model architecture is summarized in Tab. 6. We consider 3 different representation functions, network before training (with default kaiming uniform initialization [24]), after training, and after adversarial training. For the robust model, we utilize L_∞-PGD attacks implemented in the foolbox package [45] with 20 steps and a relative step size of 1.0. The attack strength ϵ is set to $\epsilon = 8/255$. The model was trained for 10 epochs with a batch size of 512, learning rate of $1\mathrm{e}{-3}$ with a cosine annealing schedule [39], and stochastic gradient descent with a momentum of 0.9.

In Tab. 7, we present the left-hand side (LHS), the rescaled difference between clean and robust cross-entropy (CE) loss, and the right-hand side (RHS), the representation function's AS score, of Lemma 1. To obtain the scores for LHS and RHS, we separately attack f as well as the MSE on g_θ, where in both cases the objective uses the MSE. For these attacks, we again utilize L_∞-PGD, but also L_2-PGD attacks, both based on the foolbox [45] implementation, with 50 steps and a relative step size of 1.0. The attack strength ϵ is set to $\epsilon = 16/255$ for both the attack types. As can be seen, LHS is indeed always upper bounded by RHS. Note that, for each pretraining and attack type combination, the RHS is the same over all latent factors, i.e. in each row, because the same images and the same representation function are used.

Pretraining	Attack Type		Scale	Orientation	posX	posY
Adversarial	L_∞-PGD	LHS	0.10	0.09	0.02	0.02
		RHS	2.29	2.29	2.29	2.29
Standard	L_∞-PGD	LHS	3.32	1.20	2.02	2.95
		RHS	23.80	23.80	23.80	23.80
Random	L_∞-PGD	LHS	2.33	2.29	2.43	1.84
		RHS	29.25	29.25	29.25	29.25
Adversarial	L_2-PGD	LHS	0.004	0.003	0.001	0.001
		RHS	0.15	0.15	0.15	0.15
Standard	L_2-PGD	LHS	0.33	0.22	0.17	0.19
		RHS	3.15	3.15	3.15	3.15
Random	L_2-PGD	LHS	0.19	0.21	0.14	0.12
		RHS	1.69	1.69	1.69	1.69

Table 7: Left-hand side (LHS) and right-hand side (RHS) of Lemma 1 using a model pretrained on the *orientation* latent factor of dSprites.

B.3 Classification Criterion - Attack Strength

Here, we study the hypothesis stated in Sec. 5.2 that datapoints fulfilling the bound in Thm. 2 are still robust against an increased attack strength. We repeat the experiments from Sec. 5.2 for the model pretrained on CIFAR-100, where the attack strength is $\epsilon = 8/255$, since, for this case, a subset of each dataset fulfilled the equation. Additionally, we compute the robust accuracy against stronger attacks employing double and three times the attack strength, i.e., using $2\epsilon = 16/255$ and $3\epsilon = 24/255$.

In Tab. 8, we observe that indeed most of the datapoints fulfilling the bound in Eq. (1) are still robust against the doubled attack strength, which is suggested by the robust accuracy for 2ϵ being larger than the proportion that fulfilled the bound. Most of these datapoints, approximately 50%, are even robust against 3ϵ. Further analysis is needed to better understand the observed robust accuracy scores and to derive a precise connection to the bound from Thm. 2.

		CIFAR-100		
	attack strength	CIFAR-10	Fashion-MNIST	Intel Image
clean accuracy	0	70.9%	84%	78.5%
robust accuracy	ϵ	42.2%	56.2%	56.8%
prop. fulfilled	ϵ	11.3%	4.6%	27.1%
robust accuracy	2ϵ	18.4%	28.5%	31.4%
robust accuracy	3ϵ	6.6%	13.1%	13.4%

Table 8: For the CIFAR-100 pretraining dataset, we report clean accuracy, the proportion of images fulfilling the theoretical bound, and robust accuracy for several attack strengths for each of the CIFAR-10, Fashion-MNIST and Intel Image downstream tasks.

B.4 Comparison Study between LP, FT and LP-FT

Here, we compare robustness transfer on classification tasks for three transfer learning techniques, linear probing, full finetuning, and linear probing then finetuning [35]. As explained before, linear probing (LP) refers to freezing all pretrained weights and only optimizing the linear classification head. In full finetuning (FT) all pretrained parameters are updated during transfer learning. Linear probing then finetuning (LP-FT) combines LP and FT sequentially, as after LP is done, FT is applied to all parameters.

For the model pretrained on CIFAR-100, we choose a batch size of 128, and run 20 epochs for FT, LP and 10 epochs of FT after LP for LP-FT. For the model pretrained on ImageNet, we choose a batch size of 32 to compensate for the increase in input resolution to 256x256. We run 10 epochs for FT and LP and 5 epochs for LP-FT. In both settings, we use half as many epochs for finetuning

after linear probing in LP-FT, as in [35]. For all methods we apply a cosine annealing learning rate schedule [39], with an initial learning rate of 0.01 for LP on CIFAR-100, and 0.001 for all other scenarios. We use stochastic gradient descent with a momentum of 0.9 on the cross-entropy loss for all experiments.
We evaluate the adversarial robustness on the test set after training by employing the L_∞-PGD attack provided by the foolbox package [45], with 20 steps and a relative step size of 0.7. As the attack strength, we choose $\epsilon = 8/255$ for the model pretrained on CIFAR-100. Due to the increase in input resolution for the ImageNet model, we apply $\epsilon = 1/255$.

Tab. 5 in Sec. 6.2 shows the performance of each transfer learning method applied to the model robustly pretrained on CIFAR-100. For all datasets, LP preserves the most robust accuracy, while performing worst on natural examples. FT and LP-FT show similar natural and robust accuracy. Only for the Intel Image dataset do we observe a significant difference (10%) in the robustness of LP-FT and FT. In contrast, when pretraining on ImageNet, LP has the worst clean and robust accuracy. Instead, LP-FT has the highest robust accuracy for all datasets, while FT achieves the highest clean accuracy. Although FT has the highest clean accuracy for all datasets, the difference between FT and LP-FT is marginal.

Additionally, Tab. 9 provides confirmation that our theory indeed holds for the other transfer learning approaches as well. The difference is that for FT and LP-FT the AS score changes during transfer learning and thus the score computed prior transfer learning is no longer predictive of the transferred robustness. Still, it is an interesting observation that indeed the AS score shrinks during LP-FT when it improves robustness transfer compared to LP.

CIFAR-100		CIFAR-10			Fashion-MNIST			Intel Image	
Method	ΔCE	$L_\infty(W)$	AS	ΔCE	$L_\infty(W)$	AS	ΔCE	$L_\infty(W)$	AS
FT	8.7	1.31	8.93	4.49	1.26	6.68	5.62	1.1	6.86
LP	0.73	4.4	0.98	0.8	4.59	2.11	0.47	3.05	1.05
LP-FT	14.1	4.42	6.44	5.98	4.6	4.27	3.06	3.07	3.02
ImageNet		CIFAR-10			Fashion-MNIST			Intel Image	
Method	ΔCE	$L_\infty(W)$	AS	ΔCE	$L_\infty(W)$	AS	ΔCE	$L_\infty(W)$	AS
FT	4.39	1.1	12.6	2.28	1.1	9.0	0.32	0.94	3.8
LP	4.14	1.86	15.7	9.4	2.02	53.2	0.14	1.39	2.7
LP-FT	3.18	1.9	9.8	0.92	2.06	5.1	0.37	1.0	3.9

Table 9: Absolute difference between clean and robust CE loss (ΔCE), $L_\infty(W)$ from Thm. 1 and the adversarial sensitivity score (AS) for the three transfer learning approaches using the model robustly pretrained on CIFAR-100 or ImageNet.

B.5 Distribution Shift Experiments

In this section, we analyse the impact of distribution shift between pretraining and downstream data on the adversarial sensitivity score and thereby on robustness transfer via linear probing. First, we consider the distribution shift between CIFAR-10 and ImageNet and, in particular, its dependence on how the smaller CIFAR-10 images are resized before applying the representation function. Second, we employ ideas from prior work [26][4] to create increasingly corrupted versions of CIFAR-100 to simulate varying shifts in distribution.

In our experiments in Sec. 5, we resize the 32x32 CIFAR-10 input images to 256x256 for the ImageNet model, and observe a strong distribution shift by a severely increased AS score. Here, we evaluate if increasing the image size before applying the ResNet model causes this shift, or if it is inherent in the data. In Tab. 10, we repeat our linear probing experiment for the CIFAR-10 to ImageNet case on differently rescaled input images. The results indicate that reducing the increase in the input image size improves robustness transfer, at the cost of a drop in clean performance. However, even for a smaller input image size, the AS score remains increased indicating a tangible distribution shift remains.

[4] `https://github.com/hendrycks/robustness/blob/master/ImageNet-C/create_c`

input image	clean /robust CE	clean /robust acc	prop. fulfilled	LHS	RHS
256x256	0.22 / 4.36	92% / 16%	0.1%	1.14	15.7
128x128	0.41 / 1.31	86% / 57%	0.2%	0.28	12.7
64x64	1.04 / 1.49	64% / 48%	0.1%	0.14	8.2

Table 10: The table contains clean and robust CE loss and accuracy, the proportion of images fulfilling the theoretical bound in Eq. (1), LHS and RHS of Lemma 1, when performing linear probing from an ImageNet pre-trained model based on differently resized CIFAR-10 images.

In the following, we consider an increasingly corrupted version of CIFAR-100 to simulate distribution shift following prior work [26]. In particular, we use the Gaussian noise (Tab. 12) and Zoom blur (Tab. 13) corruption types with increasing severity to simulate an increasing distribution shift. The experimental results from both distortion methods confirm that an increasing shift between the data distributions leads to an increase in the AS score and poorer robustness transfer via linear probing. The different corruption strength are used as suggested in [26] and are stated in Tab. 11.

Approach	C1	C2	C5
Gaussian Noise	0.04,	0.06	0.10
Zoom Blur	[1, 1.06]	[1, 1.11]	[1, 1.26]

Table 11: Definition of the corruption strength considered in our experiments. The Gaussian noise corruption severity is defined by the scale, i.e. the variance of the Gaussian noise distribution. The Zoom blur is defined via the interval of zoom factors. For each 0.01 increment in the interval, a zoomed-in image is created, and all zoomed-in images are summed to create a blurred image.

Metric	CIFAR-100	CIFAR-100-C1	CIFAR-100-C2	CIFAR-100-C5
Acc.	67.3%	66.5%	65.3%	63.6%
Robust Acc.	46.8%	44.5%	43.3%	40.9%
CE	1.28	1.32	1.37	1.44
robust CE	2.1	2.21	2.3	2.44
difference CE	0.82	0.89	0.93	1.0
$L_\infty(W)$	3.31	3.34	3.36	3.47
LHS (Lem.1)	0.124	0.133	0.138	0.144
AS score	0.835	0.892	0.915	0.949

Table 12: Various metrics for transfer learning from a model robustly pretrained on CIFAR-100 to an increasingly corrupted CIFAR-100 dataset using Gaussian noise. C1 indicates severity level 1 of the corruption is applied.

Metric	CIFAR-100	CIFAR-100-C1	CIFAR-100-C2	CIFAR-100-C5
Acc.	67.3%	65.4%	65.7%	63.9%
Robust Acc.	46.8%	41.2%	40.3%	35.5%
CE	1.28	1.36	1.36	1.43
robust CE	2.1	2.39	2.42	2.73
difference CE	0.82	1.03	1.06	1.3
$L_\infty(W)$	3.31	3.36	3.35	3.42
LHS (Lem.1)	0.124	0.153	0.158	0.19
AS score	0.835	1.001	1.031	1.195

Table 13: Various metrics for transfer learning from a model robustly pretrained on CIFAR-100 to an increasingly corrupted CIFAR-100 dataset using Zoom blur. C1 indicates severity level 1 of the corruption is applied.

Optimization of Inverted Double-Element Airfoil in Ground Effect using Improved HHO and Kriging Surrogate Model

Paras Singh[1*], Arun Ravindranath[1], Aryan Tyagi[1], Aryaman Rao[2], Raj Kumar Singh[1]

[1] Department of Mechanical Engineering, Delhi Technological University, New Delhi, India

[2] Department of Electrical Engineering, Delhi Technological University, New Delhi, India

[*] parass2802@gmail.com

ABSTRACT

In the automotive industry, multi-element wings have been used to improve the aerodynamics of race cars. Multi-element wings can enhance a vehicle's handling and stability by reducing drag and increasing downforce, allowing it to corner more effectively and achieve higher speeds. Performance gains by utilizing the ground effect are highly sensitive to the wing setup. This study focuses on identifying the optimum design parameters for the airfoil to achieve the desired downforce and drag performance. The design parameters chosen are ride height, flap overlap, flap angle, and flap gap (the spacing between the flap and the main airfoil). These parameters are optimized for three different use cases: high downforce, low drag, and a setup with the highest airfoil efficiency. The force coefficient and flow field data were gathered using two-dimensional (2D) Reynolds Averaged Navier Stokes (RANS) simulations, with the turbulent flow modeled using the k-ω Shear Stress Transport (SST) turbulence model. The Improved Harris Hawks Optimization (HHO) algorithm was used to obtain the optimal configuration of the double-element and the resulting designs showed a significant improvement in downforce and drag performance compared to the baseline designs. Improved HHO was further compared with other state-of-the-art algorithms for assessing the algorithm's performance for a problem with highly non-linear behavior, where it was able to demonstrate its ability to obtain the optimal solutions more efficiently.

Keywords: Computational Fluid Dynamics; Aerodynamics; Multi-Element Airfoils; Ground Effect; Harris Hawks Optimization; Surrogate Model

1 Introduction

Historically, aerodynamics typically played a significant role in aircraft development, where it was an essential aspect of the design process. In contrast, it did not usually hold as much importance in the development of road vehicles, where factors like aesthetics often took precedence. However, racing cars were an exception; they strategically utilized airflow to enhance cooling and performance [1]. In general, a car's aerodynamic performance can be enhanced in two ways: by reducing drag, which leads to higher straight-line speeds, or by increasing negative lift (downforce), which improves the car's grip for faster cornering. Improving downforce often increases drag, and vice versa, necessitating a balance between the two to optimize performance based on the specific track layout [2].

The majority of aerodynamic development on open-wheel racing cars has been on improving downforce, thereby improving performance in zones with limited traction such as high-speed corners. This led to inverted airfoil wings having found use in improving the aerodynamics of race cars, the renowned Lotus 49B being one of the first cars to showcase inverted wings in 1967 [3]. By reducing drag and increasing downforce, inverted wings were found to enhance a vehicle's handling and stability, allowing it to corner more effectively and achieve higher speeds. By 1984, McLaren developed multi-element wings for the MP4/2 which led to inverted multi-element wings in ground effect being used within the automotive industry extensively in the years that followed [4].

When positioned near a ground plane, an inverted wing creates a more intricate flow pattern in contrast to the free-stream scenario. In the case of multi-element wings in free-stream conditions, the resulting flow pattern is predominantly influenced by viscous effects and includes areas with streamlines with pronounced curvature, the merging of various boundary layers, and the convergence of wake regions [5]. Hence, the interaction between a multi-element airfoil's flow field and the ground effect conditions results in a significantly intricate, mutually influencing flow pattern primarily characterized by viscous effects.

The study of multi-element flows has presented a longstanding challenge to researchers, both in experimental and computational domains. In 1978, Smith [6] presented a proof stating that double-element airfoils are better than single-element airfoils, thus by the nature of the development, proving that multi-element airfoils are better than lone airfoils in certain conditions. Twenty years later, Ranzenbach et al. [7] conducted investigations on an inverted airfoil (NACA 63_2-215 Mod B) using computational and experimental methods and compared the results with road conditions. They carried out Reynolds-averaged Navier-Stokes or RANS simulations while considering a moving ground at a Reynolds number (Re) of $1.5x10^6$ based on the airfoil's chord length. The grid used consisted of only $3x10^4$ cells, considerably coarser than current investigations which usually use grids in the magnitude of $6x10^5$ cells.

In the years following Ranzenbach, Zerihan [8], Zerihan and Zhang [9], and Zhang and Zerihan [10-12] conducted various experiments on a 1:1.25 scale model of the Tyrrell 026 car's front wing, which competed in the 1998 Formula One championship. These experiments have arguably become the primary reference for researchers seeking validation in the study of single and multi-element wings operating in ground effect. The evaluations were conducted using laser Doppler anemometry and particle image velocimetry, along with the collection of force balance data to assess both drag and downforce. Mahon and Zhang [13,14] used these results to conduct computational investigations into a double-element airfoil in ground effect and concluded that the main element of the airfoil was responsible for a major part of the downforce generated while the flap contributed towards most of the drag generated by the wing. It was also deduced that the overall thickness of the wake created by the wing increased with a reduction in ride height due to the spreading rate of the wake beneath the main element vastly increasing as the ride height was reduced.

Conventionally, design engineers relied manually on design processes with the help of their intellect and expertise. However, over the past years, the complexity of these design processes has increased attributable to intricate factors such as design constraints, multi-body dynamics, and non-linearity. This inconvenience led to the adoption of classical optimization techniques to accurately estimate the optimal parameters employing the gradients of various objective functions in the field of design mechanics. Although the classical methods had their perks, they suffered from getting trapped in the

local optima of the objective and their time complexity was quite high, making them unfeasible for real-world applications. The introduction of metaheuristics optimization gained immense popularity in the domain of mechanical engineering spanning from aerospace or aircraft design [17] to automotive industries [18]. Metaheuristics overcame premature convergence without the assistance of objective function gradients which can be time-consuming. This popularity prompted us to incorporate Improved HHO to acquire the optimal parametric configuration for the double-element wing for improved drag and downforce performance.

Despite these experiments, there still remain gaps in knowledge regarding the optimal performance of multi-element inverted airfoils in ground effect. The intricate aerodynamic interactions, the requirement to optimize ground clearance, the difficulties with computational modeling, the difficulties of design optimization, and the requirement for experimental validation all contribute to the complexity of double-element airfoil optimization in the case of ground effect. To increase the setup's overall effectiveness, it becomes necessary to optimize the various characteristics of the multi-element wing for diverse usage scenarios. This study focuses on computationally identifying the optimum design parameters for the double-element airfoil to achieve the desired downforce and drag performance.

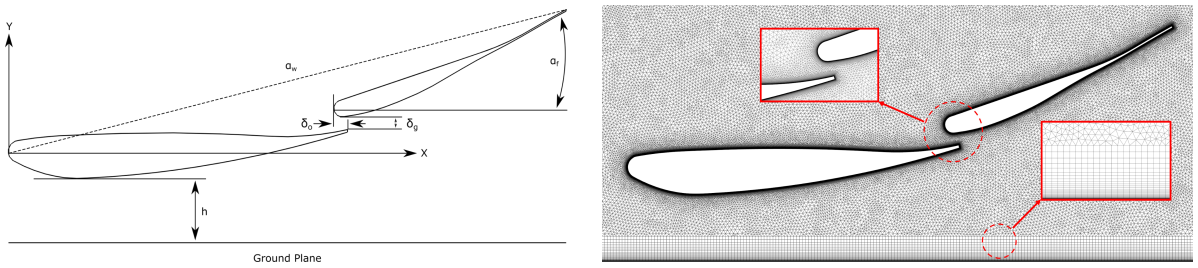
2 Methodology

2.1 Airfoil Layout and Parameterization

The optimization parameters chosen are ride height, flap overlap, flap angle, and flap gap. These parameters are optimized for three different use cases: high downforce, low drag, and a setup with the highest airfoil efficiency. This study uses the data provided by Zhang and Zerihan [10]. The model used is the front wing of the racecar Tyrrell Racing built for the 1998 Formula One Championship (Tyrrell 026) at a 1:1.25 scale, which in turn used a modified version of the airfoil created by NASA, here referred to as the GA(W) LS (1)-0413 airfoil. This model was employed due to the extensiveness of the experiments, along with data for various aerodynamic forces like downforce and drag, and pressure measurements along the span and chord for a variety of angles of attack and ride heights [17]. For the present study, the angle of incidence of the main element is held constant at $\alpha_m = 1°$. The ride

height h lies within the range of 20 mm to 200 mm and the flap angle α_f lies between 0 to 25°. The range for flap overlap δ_o is taken to be 30 mm from both sides of the tip of the main element while the range of 9 mm to 40 mm is considered as the flap gap δ_g range. The schematic of the airfoil configuration is shown in Fig.1 (a).

For modeling the flow physics of the problem accurately a high-quality unstructured tri mesh was used with progressively finer cells near the airfoil and the moving ground. To replicate the actual conditions inside the wind tunnel, a moving ground was placed beneath the airfoil with inflation layers to model boundary layer effects accurately. The initial cell spacing, parallel to the wall, inside the ground and airfoil boundary layer blocks, was determined to guarantee that $y^+ \approx 1$. The ground and airfoil surfaces were represented as solid walls with a no-slip constraint. The translational velocity of the ground surface was set equal to the freestream to accurately represent the ground effect. The mesh around the airfoils and the moving ground is presented in Fig.1 (b).



(a) (b)

Fig 1: (a) Definition of terms and coordinate system, (b) Mesh near the airfoils and the moving ground

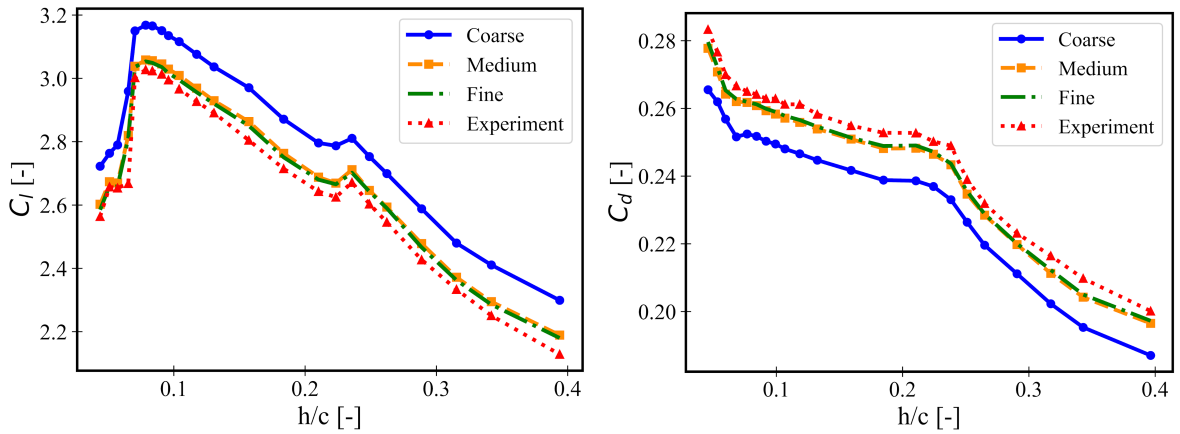
2.2 Setup and Boundary Conditions

Steady-state RANS simulations were conducted using the k-ω Shear Stress Transport (SST) turbulence model for closure. The inlet was set to a freestream velocity of 30 m/s, with a Reynolds number based on the airfoil's chord length held at 7.86 × 10^5. The airfoil was subjected to a no-slip boundary condition, and a pressure outlet condition was applied at the exit. The operational pressure was defined as 101325 Pa, with both inlet and outlet set to a gauge pressure of 0 Pa. Given that the Mach number for the simulation was well below 0.3, an incompressible pressure-based solver was employed. The fluid density (ρ) was kept constant at 1.225 kg/m^3 and dynamic viscosity (μ) was maintained at

1.7894×10^{-5} kg/m.s. Non-dimensionalized force coefficients (C_l and C_d) were derived using a reference area, A = 1.6 m^2, and a reference length, l = 1 m. The pressure-velocity coupling utilized the COUPLED algorithm, and gradient discretization was carried out using a least square cell-based scheme. Spatial discretization for momentum, pressure, specific dissipation rate, and turbulent kinetic energy employed a second-order linear upwind scheme. To ensure convergence, the simulation underwent 1000 iterations for all cases, until C_l and C_d monitors, along with the residuals, reached the predetermined limit of 1×10^{-5}.

2.3 Grid Independence Test and Model Validation

To examine the sensitivity of the numerical solution on the grid size, a grid independence test was conducted. For the test, three different grids were constructed, namely Coarse, Medium, and Fine, with their cell counts being 2×10^5, 4×10^5, and 8×10^5 respectively. In addition, the numerical results were validated against experimental data [4]. The results of the study are summarized in Fig.2. It can be observed that both the fine and medium grids work well at modeling the behavior of C_l and C_d for the double-element airfoil. Also, the deviation between the values of C_l and C_d for the fine and medium grids is quite small and thus we can infer that grid convergence has been achieved. Hence, to save computational resources, the medium grid will be utilized for further parts of the study.



(a) (b)

Fig 2: Comparison of (a) C_l and (b) C_d for the double-element airfoil computed using three different grid resolutions (Coarse, Medium, and Fine) with experimental data [4]

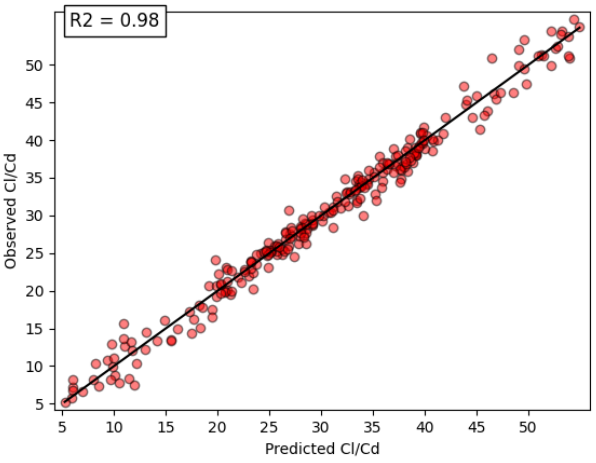


Fig 3: Validation of Kriging Surrogate Model

2.4 Optimization Methods

The optimization methodology involved the integration of the Kriging surrogate model with the Improved Harris Hawks Optimization algorithm [7]. To construct the surrogate model, a training dataset was obtained through RANS simulations of the double-element airfoil. Sobol sampling method was employed to generate 200 samples for the training data. The independent variables, as discussed above, were ride height, flap overlap, flap angle, and gap. The objective function was chosen appropriately for each setup: minimization of drag coefficient for the low drag setup, maximization of lift coefficient for high downforce setup, and maximization of C_l/C_d for maximum airfoil efficiency setup. The validation of the Kriging surrogate model performed using 200 additional unseen data points is shown in Fig.3.

The trained surrogate model served as a valuable tool to approximate the complex aerodynamic behavior of the airfoil, facilitating efficient exploration of the design space. To determine the optimal set of parameters, the Improved Harris Hawks Optimization algorithm was employed. This evolutionary algorithm, inspired by the hunting strategies of Harris's Hawks, leveraged a combination of exploration and exploitation techniques to iteratively search for the global optimum. The Improved HHO oscillates between two phases: exploration and exploitation phase. In the exploration mechanism, the hawks rely on their keen eyesight to spot and follow potential prey. Yet, there are instances where their prey remains hidden from view. During these times, the hawks exhibit patience, diligently monitoring vacant areas to locate their quarry. Once the prey is spotted, they transition into

their hunting phase, strategizing their descent for the capture. Typically, the prey, often a rabbit, represents the optimal candidate solution, while the hawks themselves symbolize the other potential solutions.

$$E = 2E_0(1 - \frac{t}{T}) \quad (1)$$

$$X(t+1) = X_{rand}(t) - r_1 | X_{rand}(t) - 2r_2X(t), \text{ if } q \geq 0.5 \quad (2)$$

Depending on the escaping energy E parameter as shown in Equation 1, the hawks effortlessly shift from exploration to exploitation phase. E_0 indicates the initial energy state in the equation. 1. The exploration focuses on following the prey until it has run out of its escaping energy and then finally encircling and attacking the prey, while the exploitation aims at ambushing the prey with a surprise pounce once it is exhausted. The equations (2-3) represent the updation procedure for hawks in the exploration phase and similarly, the hawks attack their prey or achieve ideal solutions through different modes in the exploitation phase. In the equation, X(t) represents the solution at time t, X_{rabbit} denotes the best or optimal solution and X_m (Eq. 4) signifies the mean solution of hawks. The parameters r_1, r_2, r_3, r_4 denote random values between 0 and 1, LB and UB represent lower and upper bounds of the optimization function. The Improved HHO effectively addresses the issue that traditional HHO often encounters which is becoming trapped in local optima when dealing with complex objective functions. This issue is resolved by incorporating functions such as a Levy-Flight, which enables the hawks to thoroughly navigate the search space, and effectively find optimal solutions.

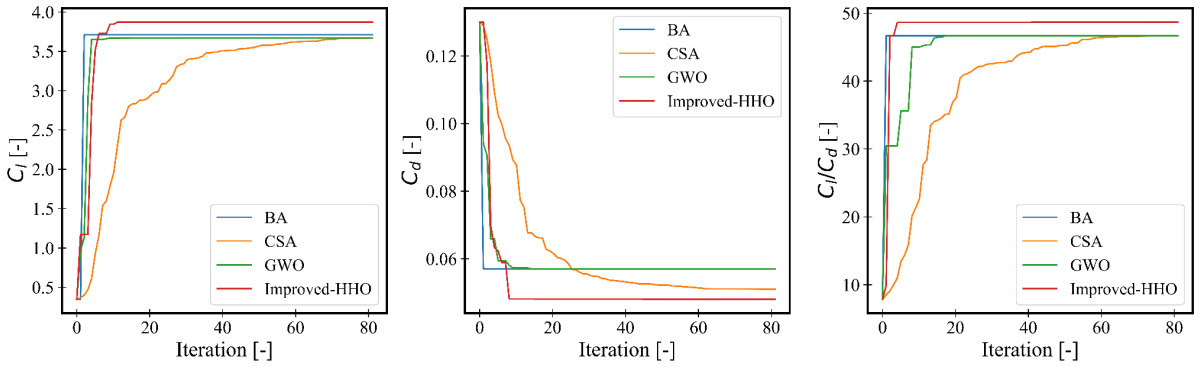
$$X(t+1) = (X_{rabbit}(t) - X_m(t)) - r_3(LB + r_4(UB - LB)), \text{if } q < 0.5 \quad (3)$$

$$X_m(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t) \quad (4)$$

3 Results and Discussion

3.1 Performance against state-of-the-art optimization algorithms

Table 2 displays the result of Improved HHO as compared to different baseline algorithms such as Bat algorithm [19], Cuckoo Search Algorithm [20] and Grey Wolf Optimisation [21]. It can be clearly observed from Figure 4 that Improved HHO outperforms the other algorithms by achieving the best fitness value for all of the three cases.



(a) (b) (c)

Fig 4. Comparison of metaheuristic algorithms for (a) maximizing C_l (b) minimizing C_d (c) maximizing C_l/C_d

3.2 Optimum parameters

The results of the study are summarized in Table 3. Here Setup A (h= 142 mm, δ_o= -2 mm, δ_g= 18 mm, α_f= 15.7°) refers to the initialized design point for the optimization algorithm, with the values of objective function being as follows - (a) C_l= 2.83, (b) C_d= 0.086, (c) C_l/C_d= 32.91. Setup B (h=20 mm, δ_o= -12 mm, δ_g= 13.8 mm, α_f= 22.5°) is the high downforce setup, Setup C (h=182 mm, δ_o= 23 mm, δ_g= 33.2 mm, α_f= 6.3°) is the low drag setup and Setup D (h=56 mm, δ_o= 8.7 mm, δ_g= 21.9 mm, α_f= 9.5°) is the setup with highest airfoil efficiency. From the table, it can be observed that the configurations obtained by the Improved HHO algorithm lead to significant improvements in the objective function.

Table 3: Comparison of force coefficients between Setup A, Setup B, Setup C and Setup D

Case	Objective Function [-]	Change in comparison to Setup A [%]
Setup B	C_l= 3.87	36.75
Setup C	C_d= 0.048	44.19
Setup D	C_l/C_d= 48.67	47.89

3.3 Velocity Distribution

For understanding the flow features leading to the increments in performance, the velocity distribution around the double element airfoil is presented in Fig. 5 for Setup A, B, C, and D. For downforce production, the pressure differential must be in the vertically downward direction. From the velocity distribution, the strongest high-velocity region is observed for Setup B followed by Setup D, then Setup A, and at last Setup C. The incoming air is accelerated due to the convexity of the airfoil's suction surface and also due to the change in cross-section formed between the airfoil's suction surface and the moving ground. This cross-section first decreases till the lowest point of the main element's suction surface is reached and then it begins to increase on moving towards the leeward side. Hence, the peak velocity is observed close to the lowest point of the main element's suction surface. Due to this, the downforce production is highest for Setup B followed by Setup D, and then A, and at last Setup C. Furthermore, a stronger wake region is also observed for the same order. The wake region is a region of slow-moving air that pulls the object towards the downstream direction. Due to the large angle attack for Setup B, the air has to traverse a sharper curvature to stay attached throughout the length of the airfoil. However as observed from Fig. 5 the incoming air does not have the required energy to traverse such sharp curvature and due to this, it separates leading to a stronger separation region for Setup B and hence the highest overall drag force. In contrast, due to the low angle of attack for the Setup C, the incoming air is able to easily follow the path along the main element and the flap leading to a much weaker wake and thus contributing to its lowest C_d among all the Setups presented.

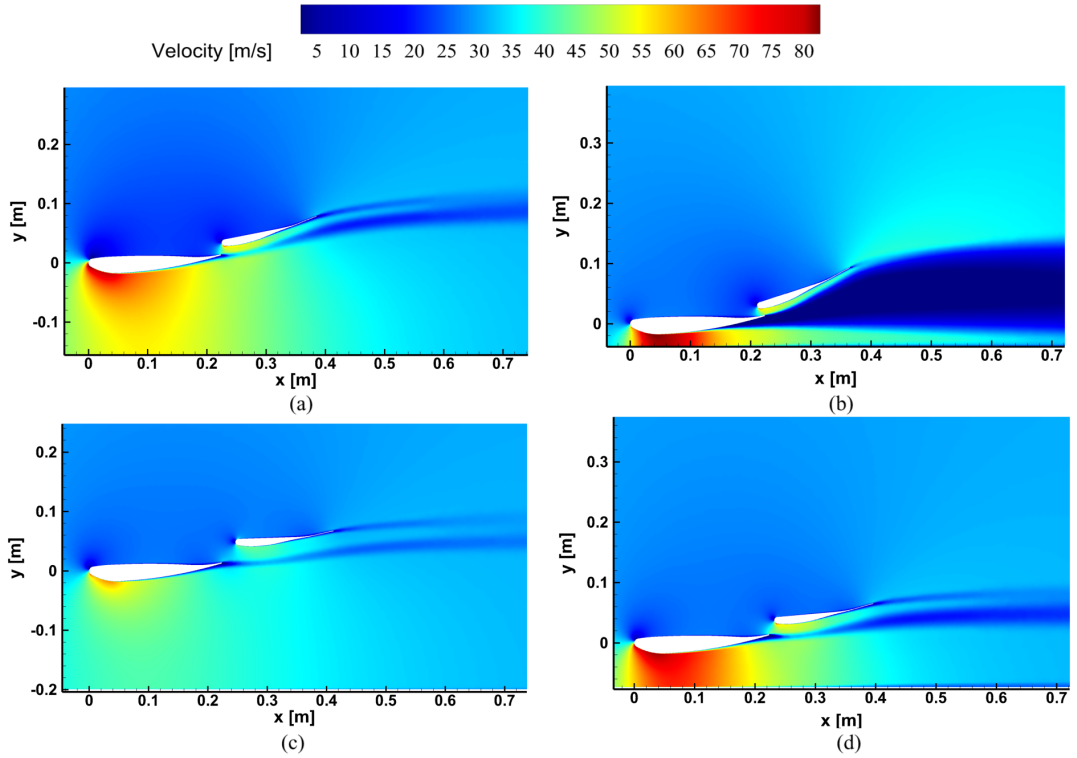


Fig 5: Velocity distribution around (a) Setup A, (b) Setup B, (c) Setup C, and (d) Setup D

3.4 Pressure Distribution

The distribution of pressure around the main element and the flap for different setups are presented in Fig. 5. The largest suction pressure is observed for Setup B due to the venturi effect as discussed earlier. This leads to a greater pressure differential between the pressure and the suction side of the double-element airfoil which in turn contributes to the highest negative lift coefficient for Setup B. Setup D has the next higher pressure differential, followed by Setup A and then Setup C. This is also evident from the coefficient of pressure distribution plot shown in Fig. 7 (a). From this plot, it can be inferred that the maximum pressure differential is generated by the main element and thus it is the major contributor towards the overall downforce production which in turn is in line with previous works [11,12,13,14].

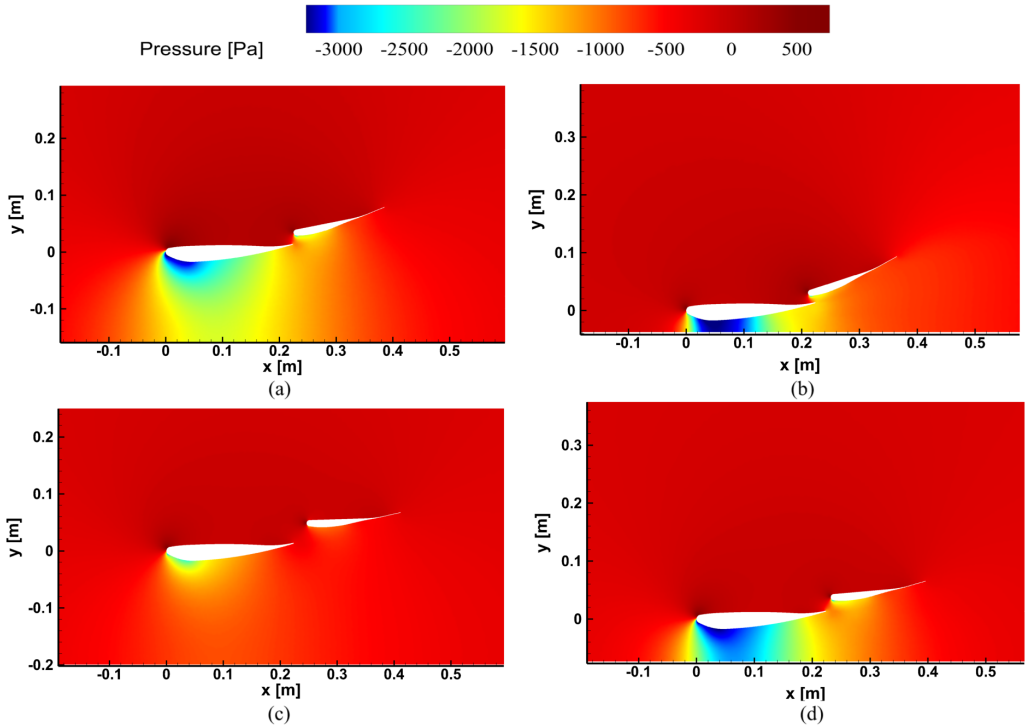
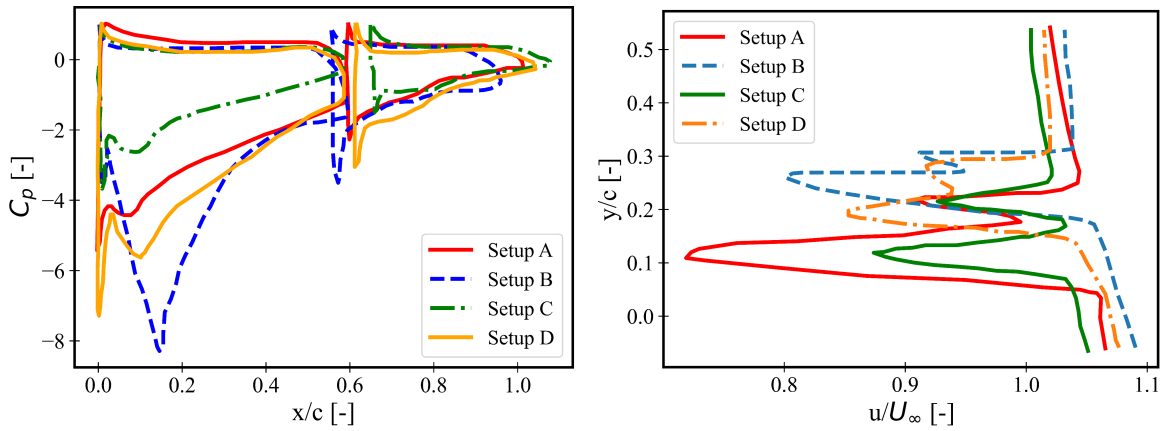


Fig 6: Pressure distribution around (a) Setup A, (b) Setup B, (c) Setup C, and (d) Setup D

3.5 Wake velocity Distribution

The wake velocity profiles at a distance of 2%c from the leading edge of the main element in the downstream direction for the different setups are presented in Fig. 7 (b). The strongest upwash effect is created by Setup B, thus leading to its highest downforce generation capability. Furthermore, the highest momentum deficit is there for Setup A which is the initialized design point fed into the optimization algorithm. The lowest momentum deficit in the wake is for Setup C, thus justifying its lowest C_d. Finally, Setup D has the second smallest momentum deficit in the wake and the second highest overall pressure differential in the downward direction and also the strength of the up wash effect. Thus it has the proper balance of both effects, leading to the highest airfoil efficiency (C_l/C_d).



(a) (b)

Fig 7: (a) Pressure distribution along the surface of the main element and the flap, and (b) Wake velocity profiles at x=2%c downstream of the leading edge of the main element for different setups

4 Conclusion

This study presented a multidisciplinary optimization methodology for the design of an inverted double-element airfoil in ground effect for applications in high-performance vehicles. The Improved Harris Hawks Optimization algorithm was employed along with the Kriging surrogate model to obtain the most optimal wing design for three cases. The observations showed that the proposed method led to an improvement of at least 35% in all three cases. The ability of the Improved HHO algorithm and its speed of convergence were further compared with more widely used algorithms such as Particle Swarm Optimization, Genetic Algorithm, Cuckoo Search Algorithm, and standard HHO algorithm.

References

[1] Slaughter, J., Moxey, D. & Sherwin, S. Large Eddy Simulation of an Inverted Multi-element Wing in Ground Effect. Flow Turbulence Combust 110, 917–944 (2023).

[2] Heyder-Bruckner, Jacques. The aerodynamics of an inverted wing and a rotating wheel in ground effect. Diss. University of Southampton, 2011.

[3] Seljak, Gregor. "Race car aerodynamics." University of Ljubljana Faculty of mathematics and physics, Department of physics (2008).

[4] Agathangelou, B., and Gascoyne, M., 1998, “Aerodynamic Design Considerations of a Formula 1 Racing Car,” SAE Publication 980399.

[5] Mahon, S., and Zhang, X. (April 9, 2006). "Computational Analysis of an Inverted Double-Element Airfoil in Ground Effect." ASME. J. Fluids Eng. November 2006; 128(6): 1172–1180.

[6] Smith, A. M. O., 1975, “High-Lift Aerodynamics,” J. Aircr., 12-6, pp. 501– 530.

[7] Ranzenbach, R., Barlow, J. B., and Diaz, R. H., 1997, “Multi-Element Airfoil in Ground Effect—An Experimental and Computational Study,” AIAA Pap. 97–2238.

[8] Zerihan, J., “An Investigation into the Aerodynamics of Wings in Ground Effect,” Ph.D. Thesis, Univ. of Southampton, School of Engineering Sciences, UK, 2001.

[9] Zerihan,J., and Zhang, X., “Aerodynamics of a Single Element Wing in Ground Effect,” Journal of Aircraft, Vol. 37, No. 6, 2000, pp. 1058– 1064.

[10] Zhang, X., and Zerihan, J., “Aerodynamics of a Double Element Wing in Ground Effect,” AIAA Paper 2002-0834, 2002.

[11]Zhang, X., and Zerihan, J., “Off-Surface Aerodynamic Measurements of a Wing in Ground Effect,” Journal of Aircraft, Vol. 40, No. 4, 2003, pp. 716–725.

[12] Zhang, X., and Zerihan,J., “Edge Vortices of a Double Element Wing in Ground Effect,” Journal of Aircraft, Vol. 41, No. 5, 2004, pp. 1127– 1137.

[13] Mahon, S., and Zhang, X., “Computational Analysis of an Inverted Double-Element Aerofoil in Ground Effect,” Journal of Fluids Engineering, Vol. 128, No. 6, 2006, pp. 1172–1180.

[14] Mahon, S., and Zhang, X., “Computational Analysis of Pressure and Wake Characteristics on an Aerofoil in Ground Effect,” Journal of Fluids Engineering, Vol. 127, No. 2, 2005, pp. 290–298.

[15] Doig, Graham and Tracie Jacqueline Barber. “Considerations for Numerical Modeling of Inverted Wings in Ground Effect.” AIAA Journal 49 (2011): 2330-2333.

[16] Harun Gezici, Haydar Livatyali, An improved Harris Hawks Optimization algorithm for continuous and discrete optimization problems, Engineering Applications of Artificial Intelligence, Volume 113, 2022.

[17] Fei C, Liu H, Zhu Z, An L, Li S, Lu C (2020) Whole-process design and experimental validation of landing gear lower drag stay with global/local linked driven optimization strategy. Chin J Aeronaut 34:318–328

[18] Millo F, Arya P, Mallamo F (2018) Optimization of automotive diesel engine calibration using genetic algorithm techniques. Energy 158:807–819

[19] Yang, Xin-She, and Xingshi He. "Bat algorithm: literature review and applications." International Journal of Bio-inspired Computation 5.3 (2013): 141-149.

[20] Xin-She Yang, & Suash Deb. (2010). Cuckoo Search via Levy Flights.

[21] Mirjalili, Seyedali, Seyed Mohammad Mirjalili, and Andrew Lewis. "Grey wolf optimizer." Advances in engineering software 69 (2014): 46-61.

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Optimizing Construction Waste Recycling: Strategies, Technologies, and Environmental Impacts

Yonatan Abera (aberas2010@gmail.com)
Dilla University

Article

Keywords: Construction waste recycling, strategies, technologies, environmental impacts, sustainable waste management, case studies

Posted Date: October 31st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3497942/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Optimizing Construction Waste Recycling: Strategies, Technologies, and Environmental Impacts

Yonatan Ayele Abera
Lecturer, Department of Construction Technology and Management, Dilla University
PhD, Candidate, Department of Civil Engineering, Delhi Technological University
Email: yonatanayele@du.edu.et

Abstract:

The increased worldwide building activity has resulted in an increase in trash output, posing serious environmental issues. As a result, the construction sector is increasingly concentrating on improving waste recycling processes in order to reduce its environmental imprint. This research article goes into the complex world of building waste recycling, analyzing tactics, technology, and the environmental consequences. The paper evaluates important recycling technologies such as source separation, on-site processing, and off-site recycling centers using a comprehensive literature analysis. The research investigates the significance of developing technologies in recycling efficiency, such as robotic sorting systems, enhanced material separation techniques, and digital platforms. The environmental ramifications are investigated, including reduced landfill consumption, energy savings, and lower greenhouse gas emissions. Real-world case studies shed light on successful garbage recycling initiatives, emphasizing the economic, environmental, and societal advantages. The article finishes by addressing construction waste recycling issues and possibilities, highlighting the necessity of stakeholder engagement, legislative assistance, and innovative technology in supporting sustainable waste management practices.

Keywords: Construction waste recycling, strategies, technologies, environmental impacts, sustainable waste management, case studies.

1. Introduction:

The worldwide construction sector, considered as a pillar of economic development and urbanization, makes major contributions to the physical environment that supports human societies. However, this business has significant environmental implications, with building and demolition waste (CDW) emerging as a major problem(Thomas & Harilal, 2016). The increasing rate of urbanization and infrastructural development has resulted in an increase in CDW creation, putting enormous strain on landfills, depleting natural resources, and contributing to pollution. In the face of escalating waste management difficulties and calls for more responsible resource consumption, the unsustainable "take-make-dispose" linear approach has become increasingly untenable(Chen et al., 2013).

In response to these issues, the notion of building waste recycling has gained importance as a critical technique for reducing the construction industry's ecological imprint. Recycling waste materials not only keeps them out of landfills, but it also allows for resource recovery, energy savings, and less environmental deterioration(Hachemi et al., 2023). The circular economy principles, which stress the continuous use and reuse of resources, are at the heart of construction waste recycling

This research article digs into the complex world of construction waste recycling, with the goal of providing a thorough knowledge of the tactics, technology, and environmental implications related with this activity. Exploring possibilities to enhance recycling operations becomes critical for encouraging sustainable development as CDW remains a global challenge(Tay et al., 2003). This study attempts to contribute to the body of knowledge regarding waste management in the construction industry by exploring the essential components of construction waste recycling, including strategies, technical advances, case studies, and problems(Walker & Pavía, 2014).

The sections that follow will look into the many aspects of construction waste recycling. It will look at several recycling systems, such as source separation, on-site processing, and off-site recycling centers. Innovative waste separation, material recovery, and digital tracking technologies will be investigated in the context of improving recycling procedures(Yaras et al., 2019). The paper will also discuss the environmental consequences of construction waste recycling, highlighting

how reduced landfill utilization, energy conservation, and lower greenhouse gas emissions contribute to the achievement of sustainable development goals(Maraveas, 2020).

(Adilhodzhaev et al., 2018).

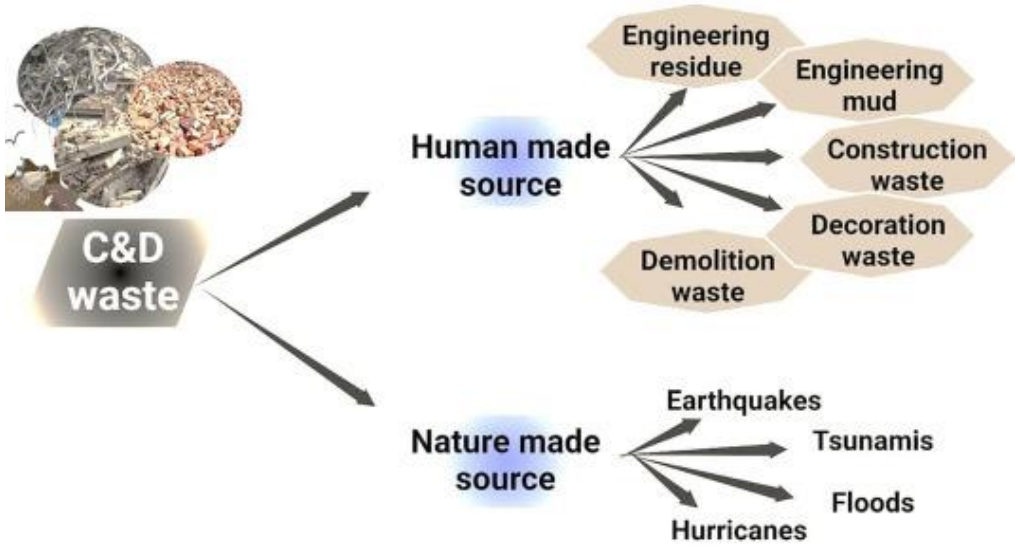


Figure 1: Sources of C&D waste

Real-world case studies will be given to demonstrate successful trash recycling initiatives as the construction sector converts to more responsible waste management methods (See fig. 1). These case studies will show the economic, environmental, and societal advantages of implementing recycling solutions in a variety of construction projects(H. Liu et al., 2016). Furthermore, the presentation will explore the problems that have been encountered in implementing construction waste recycling, such as legislative impediments, stakeholder engagement, and technical acceptance. It will also highlight the prospects for cooperation, innovation, and policy development that may encourage mainstream recycling practices(Ricciardi et al., 2021).

Finally, including construction waste recycling is a critical step toward reducing the environmental effect of the building sector. This article intends to shed light on the potential of recycling to encourage sustainable growth and resource conservation in the construction sector by researching methods, technologies, and environmental consequences(Sadek et al., 2013). It is feasible to pave the way for a more circular and responsible building sector that coincides with larger environmental goals through information sharing and stakeholder involvement(Xiong et al., 2023).

2. Literature Review

The building sector contributes significantly to environmental deterioration by producing large volumes of trash that pose considerable obstacles to sustainable development. Concrete, wood, metal, plastic, and glass are some of the materials that are wasted during building, restoration, and demolition activities(Ricciardi et al., 2021). Construction waste improperly disposed of in landfills pollutes the environment and depletes resources. As a result, there is a pressing need to maximize construction waste recycling through efficient techniques and technology to reduce its environmental impact(Sierra-Beltran et al., 2014).

Source reduction, which entails limiting trash creation on the building site, is one of the key techniques for maximizing construction waste recycling. This may be accomplished with careful planning and efficient design, with a focus on recyclable and reused materials. Architects and engineers may greatly minimize overall waste created during construction projects by supporting the adoption of sustainable construction methods(Sadek et al., 2013). Furthermore, excellent communication and coordination among stakeholders, such as architects, engineers, contractors, and waste management businesses, are critical in source reduction. Making sure that all parties involved understand the necessity of reducing trash output can lead to more environmentally conscious building practices(Sutcu et al., 2016).

Furthermore, on-site sorting and separation processes are critical for maximizing construction waste recycling. On-site sorting entails separating different types of waste materials at the source to make recycling and reuse easier(Han et al., 2020). To promote effective on-site sorting, several technologies such as mobile sorting equipment and conveyor systems can be used. Construction businesses may greatly enhance recycling rates and minimize the pressure on landfills by isolating materials such as concrete, bricks, and metals. Furthermore, offering enough training and awareness programs to construction workers on correct sorting processes will improve the efficacy of on-site waste management activities(Yan et al., 2008).

Recycling technology advances in recent years have played a critical role in maximizing building waste recycling. Mechanical recycling is one such approach, in which waste materials are treated using equipment to generate recovered aggregates for building uses(Yaras et al., 2019). Crushed concrete and asphalt, for example, may be recycled to make base materials for road building. This

procedure not only saves natural resources, but it also minimizes the demand for fresh aggregates, reducing the environmental effect of quarrying activities(Roessler et al., 2016). Thermal methods, such as incineration and pyrolysis, have also been investigated for recycling building and demolition waste. These techniques include the controlled burning of waste items in order to create heat or energy. While these technologies have the potential to minimize the amount of garbage delivered to landfills, the environmental emissions connected with these processes must be carefully considered to ensure their long-term viability(Zhang et al., 2021).

The installation of effective recycling facilities and infrastructure is another critical part of improving construction waste recycling. The availability of recycling facilities outfitted with cutting-edge gear and qualified workers is critical for processing and converting building debris into useful materials(H. Liu et al., 2016). These facilities can handle a variety of waste kinds, including concrete, wood, and metal, and use specific processes to recycle them successfully. Furthermore, the creation of a strong transportation network is essential for the effective collection and transfer of building debris to recycling facilities. Proper logistics and coordination between building sites and recycling plants may greatly cut transportation costs while also reducing the environmental impact of trash transportation(P. Liu et al., 2017).

Aside from technological considerations, economic incentives and legal frameworks play an important role in improving construction waste recycling. Governments and regulatory organizations may encourage recycling projects by providing tax breaks and subsidies to building enterprises that use environmentally friendly procedures(Adazabra et al., 2017). Financial incentives can motivate enterprises to invest in recycling technology and infrastructure, making recycling building debris economically viable. Furthermore, strict rules and policies requiring a particular amount of construction debris to be recycled might encourage industry-wide efforts toward sustainable waste management techniques. Governments may guarantee that construction businesses prioritize recycling and invest in the required infrastructure by ensuring compliance with recycling objectives(Xiong et al., 2023).

Construction waste recycling has a wide range of environmental consequences that go beyond trash reduction. The protection of natural resources is one of the key environmental benefits of recycling building debris(Zhao & Rasoulinezhad, 2023). The need for raw resources is minimized

by reusing materials such as concrete and steel, safeguarding natural habitats and ecosystems. Furthermore, recycling building debris aids in energy conservation since the creation of recovered materials uses less energy than the production of new materials from virgin resources. This energy savings helps to reduce greenhouse gas emissions, which helps to mitigate climate change and its environmental implications(Maderuelo-Sanz et al., 2022).

Furthermore, building waste recycling helps to reduce landfill consumption, which reduces the demand on limited land resources. Landfills are a major source of pollution in the environment, generating hazardous gases and leachate that damage soil and water(Alim et al., 2023). Recycling activities alleviate these negative environmental consequences and create healthier ecosystems by diverting building debris from landfills. Furthermore, recycling construction debris decreases the environmental impact of raw material extraction, processing, and shipping. As a result, the overall environmental damage caused by these activities is reduced, resulting in a more sustainable approach to resource usage(Xu et al., 2022).

To summarize, maximizing construction waste recycling is critical for attaining sustainable development and limiting the building industry's environmental consequences. Source reduction, on-site sorting, and improved recycling technology, together with the development of recycling infrastructure and supportive legislative frameworks, are critical for fostering effective waste management practices(Elheddad et al., 2021). The environmental advantages of construction waste recycling, such as resource conservation, energy savings, decreased landfill usage, and pollution reduction, emphasize the need of continuing efforts in this regard. By using these techniques and technology, the construction sector may move toward a circular economy in which waste is reduced, resources are preserved, and environmental sustainability is promoted(Nagase, 2020).

3. Materials and Methods

This study's Materials and Methods section describes the entire strategy adopted to maximize construction waste recycling, which includes a variety of techniques, technologies, and environmental analyses. The goal of this research is to promote sustainable practices in the construction sector by reducing waste output, increasing resource recovery, and reducing environmental consequences.

The first stage was a thorough examination of the construction waste stream (See fig. 2). This involved identifying the types and quantities of items that are commonly dumped on building sites. Waste streams were separated into concrete, wood, metals, plastics, and inert materials, allowing for more focused recycling operations(de Brito & Robles, 2010). To increase recycling rates, building sites were urged to separate waste materials at the source by providing clearly labeled disposal receptacles for various materials. Workers were instructed to separate garbage into designated containers, assuring material quality preservation and reducing contamination (Zhao & Rasoulinezhad, 2023).

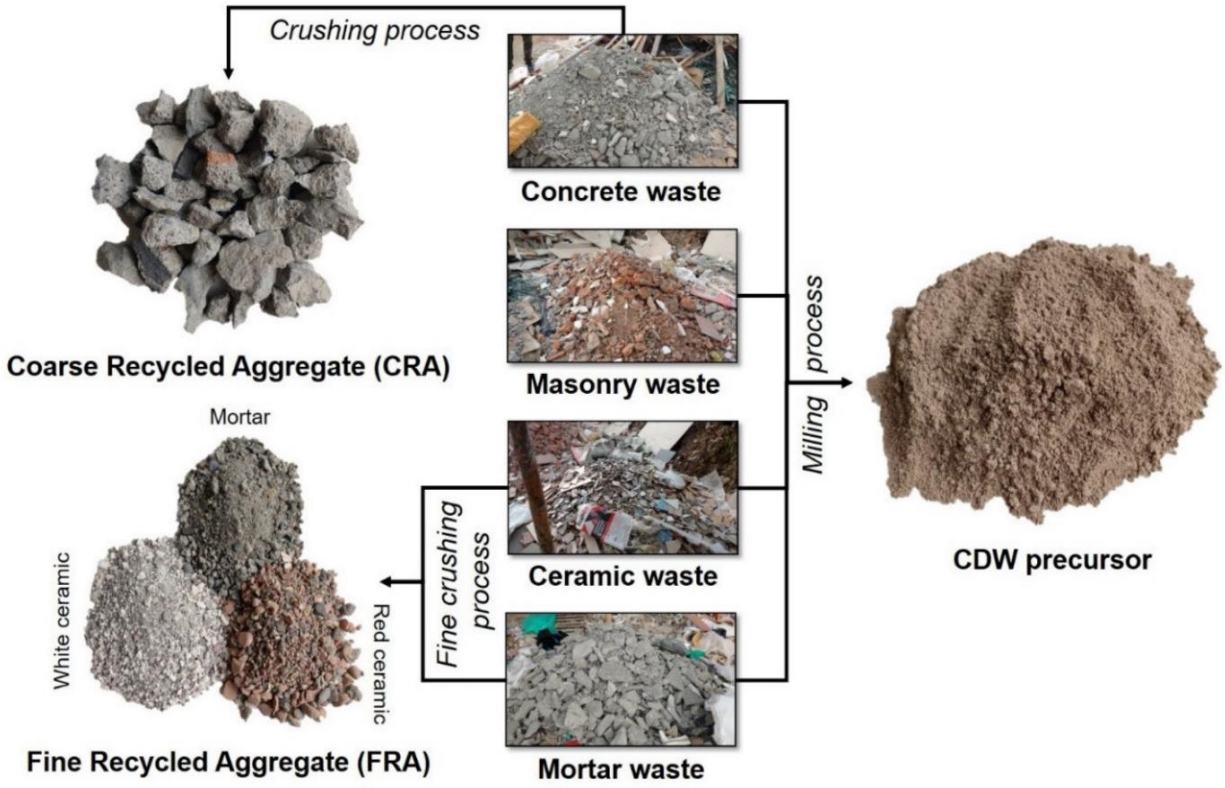


Figure 2: Production flowchart of Recycled aggregate from C&D waste

To successfully recycle various waste materials, many cutting-edge recycling technologies were used. Crushed and screened concrete debris was used to create recycled aggregates for use in new building projects. Wood waste was chopped and processed into wood chips or pellets, which may be used in landscaping or for biomass energy generation(Fapohunda et al., 2017). Magnets and

eddy current separators were used to separate metals, while plastics were cleaned, shredded, and processed into reusable pellets.

In addition to traditional recycling processes, new resource recovery techniques were used. Inert materials such as bricks and ceramics, for example, were crushed into fine powders and utilized as supplemental raw materials in concrete production, lowering the need for virgin resources and lowering carbon emissions (See fig. 3). A thorough quality control regimen was created to assure the quality and compatibility of recycled materials. Strength, durability, and particle size distribution of recycled aggregates were all tested. Materials that did not satisfy the stipulated specifications were either reprocessed or discarded(P. Liu et al., 2017).

The environmental implications of building waste recycling were thoroughly evaluated. This included assessing the reduction in landfill utilization, the energy savings resulting from the use of recycled materials, and the reduction in greenhouse gas emissions connected with waste disposal. To estimate the entire environmental benefits of recycling processes, life cycle analyses were undertaken.

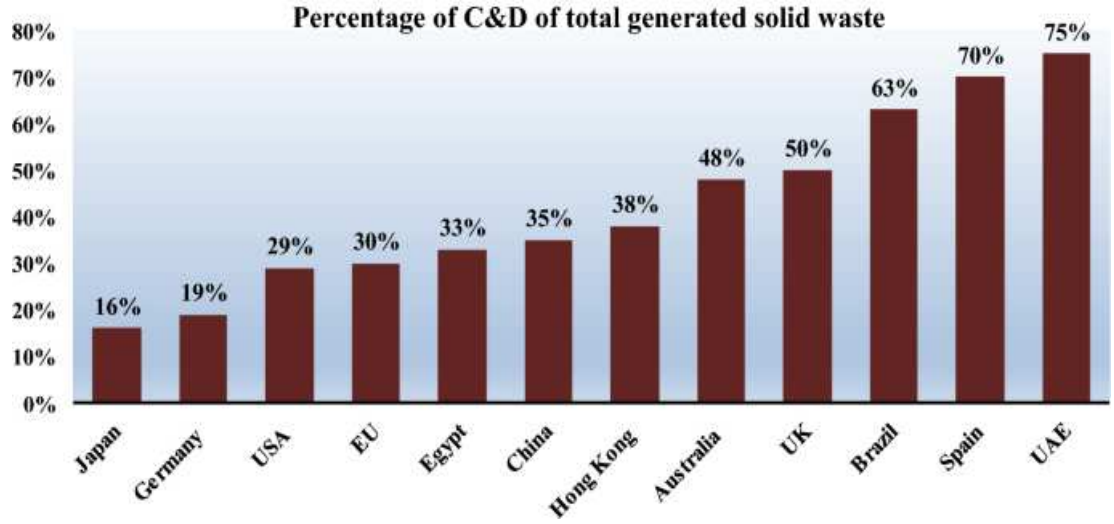


Figure 3: Percentage of C&D waste generation nation wide

To analyze the economic viability of recycling activities, a complete cost-benefit analysis was performed. This research took into account initial expenditures in recycling technology, operational expenses, revenue from recycled material sales, and prospective savings from lower

trash disposal prices. Economic viability was required to show to construction stakeholders that recycling operations make financial sense(Hu et al., 2012). In addition to recycling procedures, technical improvements such as construction waste tracking software and digital waste management systems were employed. These advances enabled real-time trash generation monitoring, enhanced logistics, and expedited waste treatment operations.

The research also looked at the influence of rules and policies in encouraging construction waste recycling. An assessment of local and national waste disposal and recycling rules was done in order to identify areas where policy adjustments may further promote recycling activities. Engagement with construction industry stakeholders, such as contractors, builders, waste management businesses, and government agencies, was a critical component of our research. The goal of the collaboration and discourse was to establish a common commitment to sustainable waste management methods in the building industry.

In summary, this study reflects a multidimensional attempt to maximize construction waste recycling. We hope to boost recycling rates while simultaneously promoting sustainable practices throughout the construction sector by defining waste streams, deploying sophisticated recycling technology, analyzing environmental and economic implications, and engaging stakeholders. These strategies and approaches are critical steps toward developing a more circular and ecologically responsible construction sector, which will eventually reduce its ecological impact and contribute to a more sustainable future.

4. Results and Discussions

4.1. Strategies for Construction Waste Recycling:

Effective construction waste management necessitates a multimodal approach that emphasizes waste reduction, promotes resource recovery, and minimizes environmental consequences. A crucial method is source separation, which includes classifying waste materials at the place of creation. Source separation improves the quality of recyclable resources by encouraging employees to isolate items such as concrete, wood, metals, and plastics (See table 1). This makes them more suited for reprocessing.

Table 1: Common Types of Construction Waste

Type of Waste	Description
Concrete and Bricks	Demolished concrete and bricks
Wood	Old lumber, plywood, and pallets
Metals	Scrap metal, steel, and aluminum
Plastic	Plastic pipes, containers, and packaging
Asphalt	Road and roofing materials
Glass	Broken glass windows and fixtures
Gypsum	Drywall and plasterboard waste
Insulation	Fiberglass and foam insulation waste

Another important method is on-site processing, which involves converting waste materials into reusable goods right on the building site. Concrete, for example, may be crushed and utilized as aggregate in the creation of fresh concrete, reducing the requirement for virgin materials. On-site processing not only decreases transportation costs, but it also promotes localized material cycles, which contributes to more environmentally friendly construction practices(Yu et al., 2015).

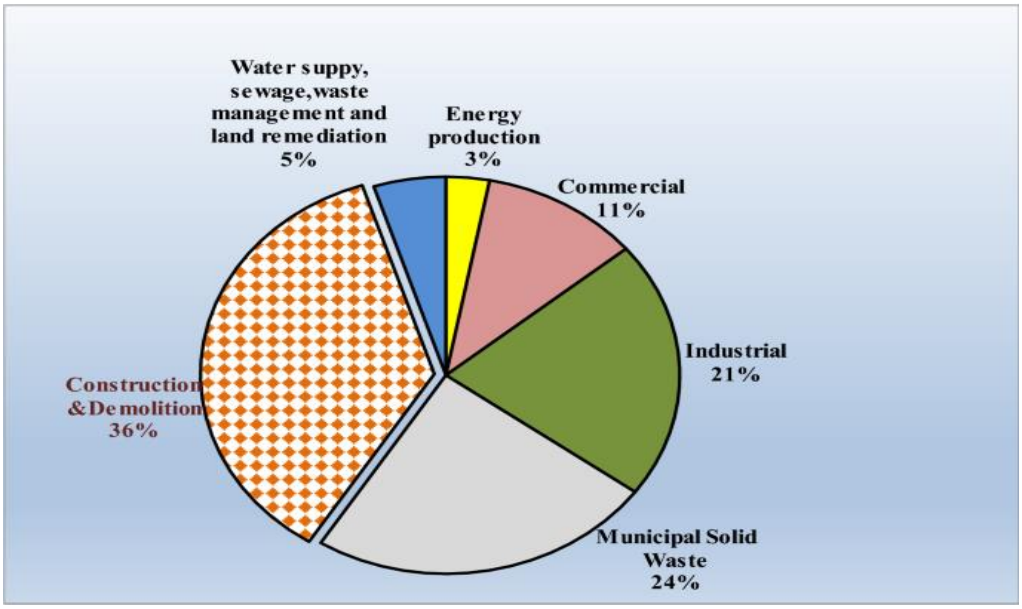


Figure 4: Broad categories waste generation

Off-site recycling centers are critical to improving construction waste recycling. These specialist facilities are outfitted with innovative sorting and processing technology that can handle a variety of waste streams efficiently (See fig. 4). Materials gathered from building sites are transferred to

these facilities, where they are sorted, cleaned, and reprocessed. Crushed concrete, recycled aggregates, and recovered metals, for example, may be reintegrated into new building projects, completing the loop and supporting circular material flows (See table 2).

Table 2: Construction Waste Recycling Strategies

Strategy	Description
Source Separation	Sorting waste at the construction site to separate recyclables from non-recyclables.
On-Site Recycling	Processing and reusing materials on-site, such as crushing concrete for reuse as aggregate.
Construction Waste Audits	Regular waste audits to identify opportunities for improvement and track progress.
Sustainable Procurement	Purchasing eco-friendly materials and products to minimize waste generation.
Waste-to-Energy	Converting non-recyclable waste into energy through incineration or other methods.
Material Exchange Programs	Facilitating the exchange of surplus materials between construction sites.
Donation and Reuse Programs	Donating usable materials to charities or organizations, promoting reuse.
Public Awareness Campaigns	Raising awareness among construction workers about the importance of recycling and waste reduction.

Furthermore, waste-to-energy conversion technologies have the potential to provide an alternate method of handling non-recyclable garbage. Waste is converted into heat and power using advanced thermal technologies such as waste incineration with energy recovery(Hunag et al., 2015). While this strategy does not adhere to the circular economy principles of direct material reuse, it can help to reduce landfill dependency and balance nonrenewable energy usage (See fig. 5).

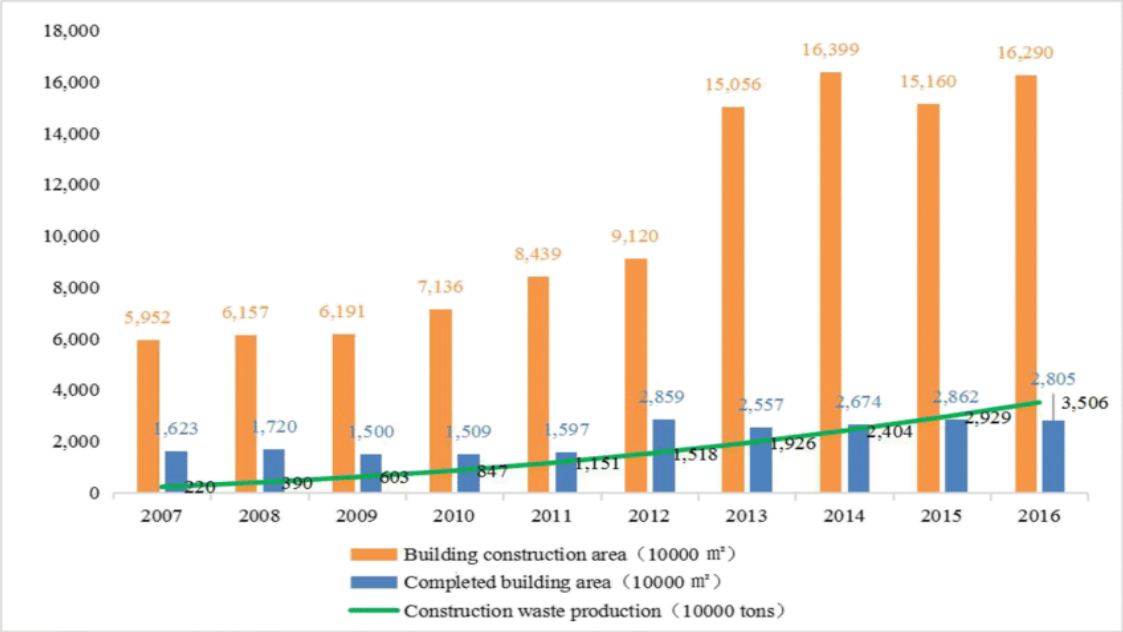


Figure 5: Comparison of C&D waste volume with construction area

In essence, a comprehensive framework for construction waste recycling is presented by a mix of source separation, on-site processing, off-site recycling centers, and waste-to-energy conversion schemes. The building industry can drastically reduce the amount of trash transported to landfills, save natural resources, and contribute to a more sustainable and circular built environment by implementing these measures.

4.2. Technological Advancements:

In the aim of enhancing construction waste recycling, technological advances are playing a critical role in modernizing recycling operations' efficiency, accuracy, and sustainability. Robotic sorting systems powered by artificial intelligence and sensors are redefining garbage separation. These automated systems can quickly recognize and sort various materials depending on their properties, improving sorting accuracy and lowering contamination. Robotic sorting systems not only speed up the recycling process, but they also eliminate dependency on human labor, boosting worker safety and productivity (See table 3).

Table 3: Key Recycling Technologies

Technology	Description

Concrete Crushers	Machines that crush concrete waste into reusable aggregate.
Wood Shredders	Equipment that shreds wood waste for reuse in particleboard or biomass energy.
Metal Separators	Devices that separate ferrous and non-ferrous metals for recycling.
Plastic Grinders	Machines that grind plastic waste into pellets for manufacturing new products.
Asphalt Recyclers	Equipment that rejuvenates old asphalt for use in new road construction.
Glass Crushers	Machines that crush glass waste into cullet for remelting into new glass products.
Gypsum Recyclers	Systems for recycling gypsum drywall into new drywall or agricultural amendments.
Insulation Reclaimers	Equipment that recovers materials from insulation waste for reuse.

Advanced material separation processes expand recycling plants' capacities even more. Light sensors are used in optical sorting technologies to recognize and separate different types of materials depending on their spectral qualities. Magnetic separation, on the other hand, uses materials' magnetic characteristics to remove ferrous metals from mixed waste streams. These cutting-edge procedures allow for the precise extraction of precious elements, improving the quality of recovered goods and encouraging resource efficiency(Tang et al., 2016).

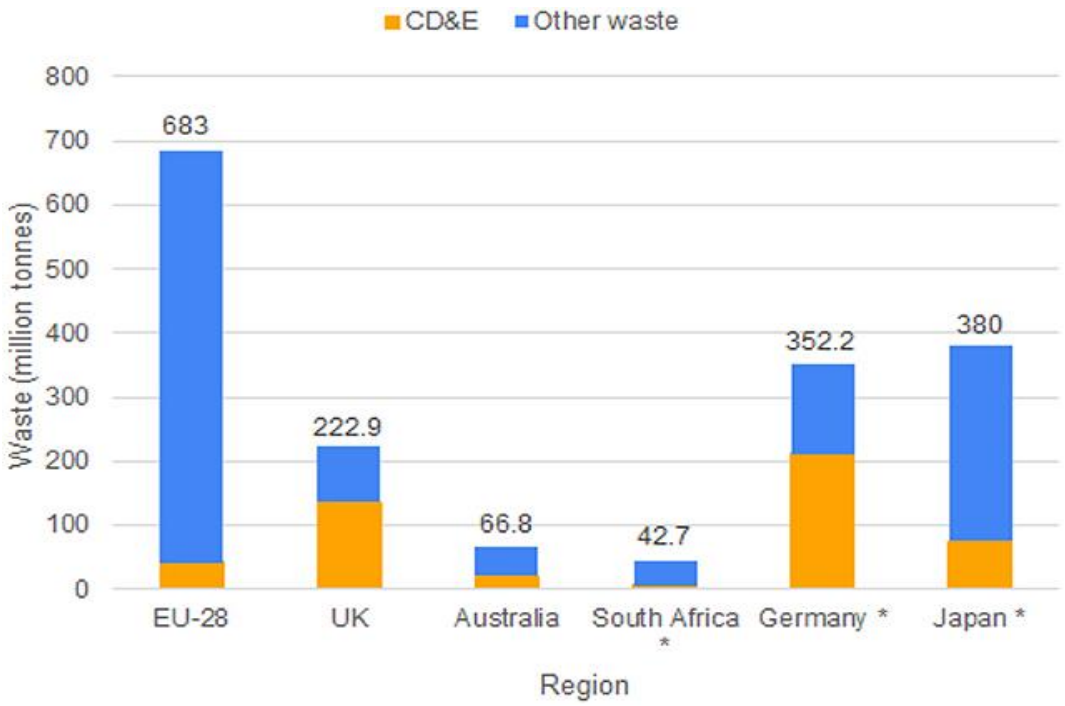


Figure 6: Volume of C&D waste generated in different countries

Digital platforms and applications are connecting stakeholders and recycling processes. Construction businesses may use waste monitoring systems to track garbage creation, collection, and disposal in real time (See fig.6). Transparency improves accountability, guarantees compliance with waste management standards, and enables data-driven decision making. Furthermore, digital networks can improve communication between building sites and recycling facilities, hence streamlining trash transportation and processing logistics (See table 3).

Furthermore, new materials are being created with recycling in mind. Biodegradable plastics, for example, provide a long-term alternative to regular plastics while lowering the environmental impact of non-biodegradable materials. These materials can be included into building goods to provide functional qualities as well as end-of-life recyclability.

Finally, technological advances are accelerating a paradigm change in building waste recycling techniques. Robotic sorting technologies, enhanced material separation techniques, and digital platforms are boosting recycling efficiency, material quality, and waste management transparency. These technologies are critical facilitators in the building industry's move to more sustainable and circular construction processes, aligning with larger environmental aims and fostering a more responsible attitude to waste management.

4.3. Environmental Impacts:

Adoption of building waste recycling technologies results in a variety of major environmental advantages that go beyond trash reduction. The decrease of landfill consumption is one of the key environmental benefits. Diverting garbage from landfills not only saves important land resources, but it also reduces the risk of soil and water contamination from landfill leachate. Construction waste recycling helps to healthier ecosystems and the preservation of natural habitats by reducing dependency on landfill disposal.

Recycling also saves energy and lowers greenhouse gas emissions. The extraction, transportation, and processing of raw materials all need significant amounts of energy. Recycling materials such as concrete, metals, and wood greatly reduces the energy consumption involved with the extraction and processing of virgin resources (See table 4). This decrease in energy usage results in decreased

greenhouse gas emissions, which helps to contribute to global efforts to combat climate change and reduce carbon footprints.

Table 4: Environmental Impact of Construction Waste Recycling

Environmental Impact	Description
Reduced Landfill Waste	Diverting construction waste from landfills reduces the strain on landfill capacity.
Energy Savings	Recycling materials typically requires less energy than producing new materials.
Resource Conservation	Recycling conserves natural resources, such as timber, minerals, and fossil fuels.
Lower Carbon Emissions	Reduced energy consumption and transportation result in lower carbon emissions.
Improved Air and Water Quality	Recycling reduces pollution associated with waste disposal and extraction of raw materials.
Habitat Preservation	Preserving natural habitats by reducing the need for resource extraction.
Economic Benefits	Job creation and cost savings associated with recycling contribute to economic well-being.
Community Engagement and Education	Promotes community involvement and environmental awareness.

Furthermore, the use of construction waste recycling accords with circular economy ideas, supporting a more sustainable approach to resource management. Recycling extends the lives of resources, reducing the demand for new materials to be mined and made. As a result, the environmental effect of resource extraction and processing is reduced, preserving limited resources and limiting ecosystem damage(Adilhodzhaev et al., 2018).

Furthermore, recycling can help to close material loops, reduce demand for virgin resources, and reduce environmental deterioration associated with resource exploitation. Recycling materials in new building projects saves raw resources and energy, resulting in a lower total environmental effect.

In short, the environmental effects of construction waste recycling are felt across various domains, including trash reduction, energy saving, carbon emission reduction, and resource preservation. The construction industry plays a critical role in promoting a more sustainable and ecologically

responsible built environment by adopting recycling practices, aligning with global sustainability targets and contributing to a more harmonious relationship between human activities and the natural world.

4.4. Economic Viability

The economic feasibility of construction and demolition (C&D) waste recycling is a critical concern in the context of sustainable waste management and resource conservation. Recycling C&D trash may provide significant economic advantages to both industry stakeholders and society as a whole. This section delves into the economics of C&D waste recycling, shining light on its possible benefits(Fairbairn et al., 2010).

To begin with, recycling C&D waste makes a compelling economic rationale by lowering disposal costs. Due to tipping fees, transportation expenses, and regulatory compliance, traditional trash disposal options such as landfilling can be prohibitively expensive. Recycling diverts a considerable part of garbage away from landfills, lowering disposal costs. Construction businesses, in particular, may save significantly by reducing landfill costs and maximizing the use of recovered materials.

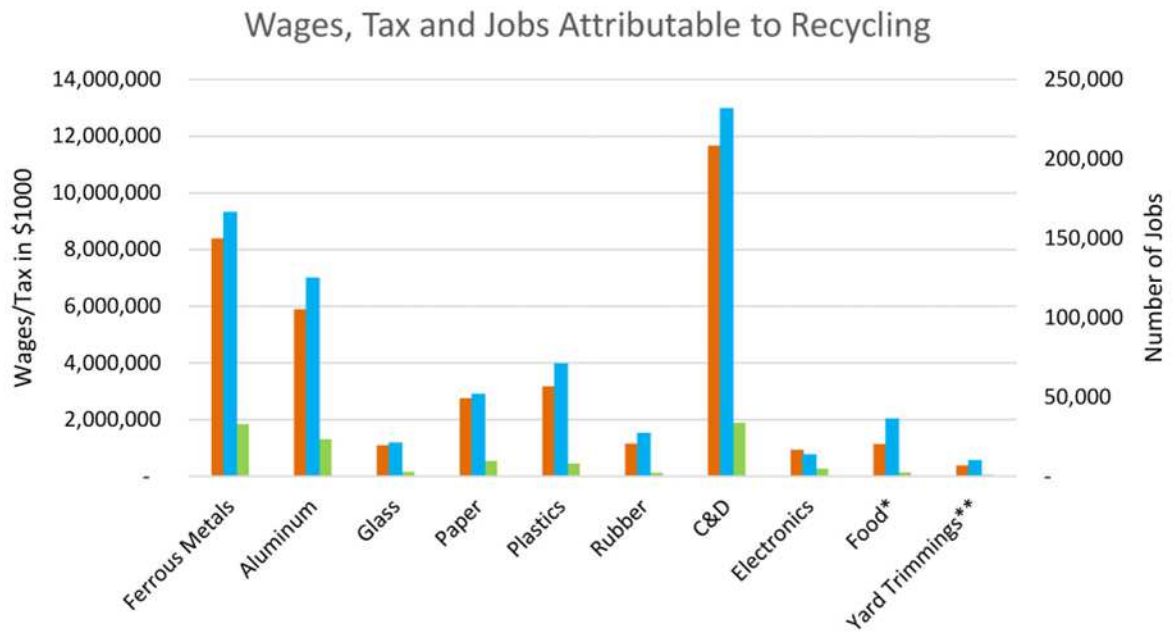


Figure 7: Economic value of recycling C&D waste

Furthermore, the sale of recovered products from C&D waste recycling might produce revenue streams. Concrete, wood, metals, and plastics are examples of typical materials found in C&D waste streams that have intrinsic economic value when processed and marketed as recovered goods (Ruggerio, 2021). For example, recycled concrete aggregates may be used in new building projects instead of virgin aggregates, lowering material procurement costs while tapping into a profitable market. Metals salvaged from demolished buildings can be sold to scrap yards, creating revenue and contributing to the circular economy (See fig. 7).

Another important factor of economic feasibility is the possibility for employment development from C&D waste recycling (See table 5). The recycling business, which includes the processing, transportation, and marketing of recovered materials, has the potential to provide jobs in local areas. These positions include a wide variety of skill levels, from labor-intensive sorting and processing to managerial and administrative positions. As a result, by providing jobs and developing economic resilience, C&D waste recycling can support economic growth.

Table 5: Financial Benefits of Construction Waste Recycling

Project	Cost Savings (USD)	Revenue Generation (USD)	Net Financial Benefit (USD)
Green Building Complex	$250,000	$100,000	$350,000
Eco-Friendly Renovation	$80,000	$25,000	$105,000
Sustainable Demolition	$150,000	$40,000	$190,000
LEED-Certified Building	$180,000	$70,000	$250,000
Zero Waste Initiative	$500,000	$300,000	$800,000

Furthermore, recycling C&D trash is consistent with environmental objectives and can improve a company's reputation and market competitiveness. Construction businesses that demonstrate a commitment to responsible waste management and resource conservation can attract environmentally concerned clients and earn a competitive edge as environmental consciousness

grows. This might lead to more project prospects and better market positioning, which would eventually contribute to long-term economic stability(Dong et al., 2023).

Figure 8: Recycling market Trend price

Recycling C&D waste not only provides immediate economic advantages, but it also reduces the economic risks associated with resource scarcity and price volatility. Construction enterprises may lessen their reliance on virgin resources, which are prone to market swings and supply chain disruptions, by using recycled materials (See fig. 8). This resource diversity can protect enterprises from the economic effects of shortages and price increases, contributing to increased economic resilience.

Furthermore, recycling C&D waste aids regulatory compliance and may result in cash incentives. Many areas have adopted waste diversion objectives and recycling incentives, which can result in financial rewards for construction firms that reach or surpass these standards. These incentives may include tax credits, grants, or refunds, which would strengthen the economic case for recycling even more.

While there are evident economic benefits to recycling C&D waste, various constraints must be considered, such as initial expenditures in recycling infrastructure and equipment, operational

expenses, and market volatility. Nonetheless, with careful planning, smart alliances, and technology adoption, these obstacles may typically be avoided.

Finally, the economic feasibility of C&D waste recycling activities is a diverse and appealing proposal. Recycling C&D trash provides substantial economic advantages to construction industry stakeholders by lowering disposal costs, increasing income, creating employment, improving reputation, and limiting economic risks. Furthermore, it promotes regulatory compliance and corresponds with environmental goals, thereby unlocking further financial incentives. Recycling C&D trash will continue to play an important role in attaining economic, environmental, and social sustainability as the construction sector implements responsible waste management methods.

4.5. Case Studies:

Examining real-world case studies gives useful insights into the practical execution and measurable advantages of building waste recycling solutions. These examples demonstrate the various situations in which recycling strategies can be used, highlighting their potential to contribute to environmental, economic, and social sustainability.

The use of source separation and on-site processing procedures resulted in significant waste reduction in a commercial building project. The initiative cut trash transportation costs, conserved resources, and lessened the pressure on local landfills by isolating materials during construction and processing them on-site (See table 6). Furthermore, recycled materials were used in the building of new structures, demonstrating how circularity may be included into the project lifespan.

Table 6: Case Studies in Construction Waste Recycling

Project Name	Location	Description	Recycling Rate (%)	Environmental Impact
Green Building Complex	City A	A multi-building project with a focus on waste reduction and recycling.	85%	Reduced landfill waste, energy savings
Eco-Friendly Renovation	City B	A renovation project that reused existing materials and recycled extensively.	92%	Resource conservation, cost savings

Sustainable Demolition	City C	A demolition project that prioritized salvaging and recycling materials.	88%	Reduced waste, habitat preservation
LEED-Certified Building	City D	A LEED-certified construction project with a strong recycling plan.	90%	Lower emissions, community engagement
Zero Waste Initiative	City E	A city-wide initiative to achieve zero construction waste to landfills.	98%	Reduced landfill waste, economic benefits

The success of off-site recycling centers is demonstrated by a residential construction project. The initiative diverted a considerable amount of its trash from landfills by collaborating with a specialist recycling plant. The recycling facility quickly collected and processed materials including concrete, metals, and wood, resulting in recovered items that were subsequently used in future building projects. This method not only decreased garbage disposal expenses, but it also encouraged collaboration within the recycling ecosystem.

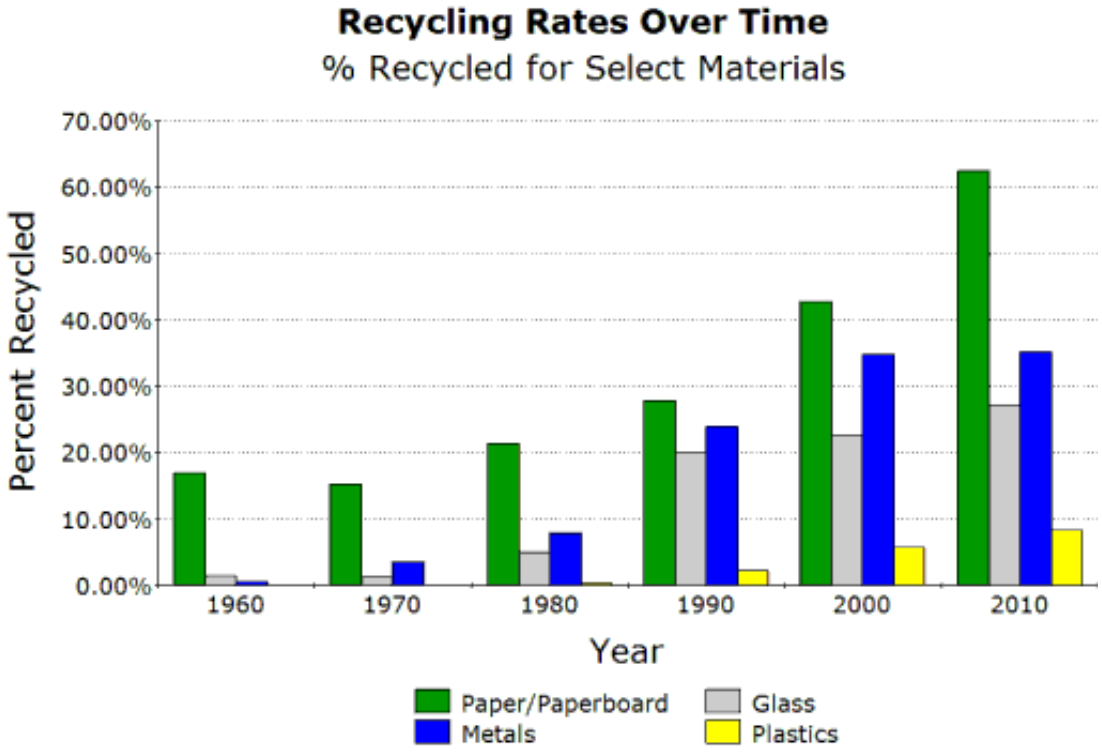


Figure 9: Recycling rate of selected waste materials over time

Construction waste recycling procedures were used in the context of a heritage building restoration to retain historical relevance while limiting environmental effect. The restoration project was able to conform with circular economy concepts by salvaging and reusing original architectural components, as well as incorporating modern technology for material identification. The symbiotic link between preservation, recycling, and sustainable building is highlighted in this example.

In addition, a large-scale infrastructure project demonstrated the economic potential of construction waste recycling by streamlining waste separation and recycling processes, lowering waste management expenditures and the demand for virgin materials. These reductions helped to reduce total costs and improved the project's competitiveness (See fig. 9).

Finally, these case studies show the flexibility and success of construction waste recycling systems in a variety of circumstances. Recycling strategies provide a variety of benefits, including trash reduction, cost savings, resource conservation, and sustainability, in anything from commercial building to historical restoration and infrastructure projects. These examples highlight the transformational potential of recycling methods in the construction sector, emphasizing the necessity of stakeholder engagement, creativity, and the incorporation of circular ideas to promote sustainable growth.

4.6. Challenges and Opportunities:

While building waste recycling offers a viable path for long-term growth, it is not without problems. One of the most significant barriers is a lack of awareness and education among stakeholders regarding the benefits and practicality of recycling methods. To overcome this difficulty, extensive educational initiatives are needed to educate all stakeholders, from construction workers and contractors to project managers and policymakers, about the necessity of recycling and its potential benefits.

Regulatory constraints and fragmented waste management rules also make extensive recycling adoption difficult. Inconsistent rules between areas may impede the smooth implementation of recycling processes, especially in cross-border initiatives. Construction waste recycling may be made more feasible and successful by harmonizing waste management rules and setting clear standards(de Vries et al., 2015).

Furthermore, for recycling operations to be optimized, technical adoption and infrastructural expansion are required. In order to integrate new technologies such as robotic sorting systems and digital tracking platforms, initial expenditures and training are required (See table 7). To overcome the reticence to embrace these technologies, industry-wide collaboration and assistance from organizations and institutions that promote innovation are required.

Table 7: Challenges in Construction Waste Recycling

Challenge	Description
Contamination	Mixing of recyclable and non-recyclable materials in waste streams.
Lack of Awareness	Limited knowledge among construction workers and contractors about recycling practices.
Space Constraints	Limited space on construction sites for waste storage and recycling facilities.
Regulatory Compliance	Adherence to recycling regulations and standards can be complex and costly.
Transportation Issues	Transporting waste materials to recycling facilities can be logistically challenging.
Market Demand	The demand for recycled construction materials can fluctuate, affecting recycling viability.
Cost of Recycling	Initial investments in recycling infrastructure may be substantial.
Waste Sorting Complexity	Sorting and processing construction waste can be labor-intensive and time-consuming.

Financial factors can also influence the adoption of recycling measures. The initial expenses of establishing recycling facilities or acquiring specialist equipment may dissuade certain stakeholders, particularly those involved in small-scale projects with limited finances. Aligning economic incentives with sustainability goals, such as grants, subsidies, or tax breaks, can stimulate the adoption of recycling systems, offering a win-win situation for the environment as well as project finances (See table 8).

However, there are countless potential for constructive transformation and progress amid these challenges. The growing demand for trained recycling specialists, together with the advent of a circular economy, creates opportunities for job creation and economic growth. Furthermore,

collaborative efforts between construction firms, waste management facilities, and recycling centers may establish a symbiotic connection that promotes waste reduction and resource recovery.

Table 8: Opportunities in Construction Waste Recycling

Opportunity	Description
Technology Advancements	Innovative recycling technologies can improve efficiency and reduce costs.
Green Building Practices	Encouraging green building practices that prioritize waste reduction and recycling.
Circular Economy Adoption	Promoting a circular economy approach to construction materials and waste.
Education and Training	Providing training and education to construction industry professionals on recycling.
Policy Support	Government policies and incentives that promote recycling and sustainable practices.
Collaboration and Partnerships	Collaborating with recycling facilities, non-profits, and local communities.
Research and Development	Continued research to identify new materials that are easier to recycle.
Market Expansion	Diversifying markets for recycled construction materials to increase demand.

Technological advancements, such as material identification and tracking systems, offer opportunity to expedite recycling operations and increase efficiency. Stakeholders may improve trash collection, transportation, and processing by developing digital platforms that connect construction sites with recycling facilities, reducing inefficiencies and increasing transparency.

Finally, the constraints of building waste recycling are exceeded by the promise for good change, creativity, and sustainable development. The construction sector can alter waste management practices and contribute to a circular economy by raising awareness, removing legislative hurdles, investing in technology, and aligning economic incentives. Recycling procedures not only decrease the environmental effect of building, but they also pave the way for a more responsible and resource-efficient built environment.

5. Conclusion:

Construction waste recycling has risen to the top of the worldwide construction sector as a result of the requirement for sustainable growth. The complicated interaction of environmental concerns, scarcity of resources, and desire for more responsible behaviors has resulted in a paradigm change in waste management solutions. This research article deconstructed the diverse arena of construction waste recycling, deconstructing its techniques, technical advances, environmental repercussions, case studies, obstacles, and potential.

The incorporation of building waste recycling technologies has enormous potential to transform the construction industry's destiny. Stakeholders may considerably reduce waste volumes, improve resource efficiency, and pave the road for a circular economy in the industry by implementing source separation, on-site processing, and off-site recycling center techniques. Technological advancements, ranging from robotic sorting systems to digital trash monitoring platforms, provide effective tools for streamlining recycling operations, improving material quality, and increasing transparency.

The environmental consequences of building waste recycling are felt throughout ecosystems and societies. Recycling solutions contribute to a more sustainable built environment by minimizing landfill usage, conserving energy, and reducing greenhouse gas emissions. Real-world case studies have shed light on the practical benefits of recycling, which range from cost savings and resource conservation to increased community participation and decreased environmental costs.

However, problems remain, ranging from a lack of knowledge and legislative impediments to financial concerns and technical uptake. Addressing these difficulties will need stakeholder engagement, novel regulatory frameworks, and industry-wide commitment to sustainable practices. These difficulties are matched, however, by potential for job creation, economic growth, technological innovation, and the alignment of economic incentives with environmental aims.

Finally, building waste recycling is a shining example of sustainability in the construction business. The transition to a circular economy requires collaborative efforts, commitment, and creativity. The construction sector may drastically decrease its environmental impact, promote responsible resource management, and contribute to a more harmonious connection between human activity

and the earth if stakeholders' band together to embrace recycling techniques. The building sector not only produces structures, but also a more sustainable future for future generations, by including recycling practices.

Acknowledgments:

The author would like to appreciate the Department of Civil Engineering, DTU, for their helpful initiatives during the current study work's practical laboratory work. I'd also want to thank everyone who helped make site visits to the Burari C&D waste recycling facilities, Shastri Park C&D waste recycling plant, and Mundika C&D waste recycling facility in New Delhi possible.

Data Availability:

Supporting data for this research is included in the manuscript.

Funding:

There is no external funding received. The researcher has covered all the financial issues.

Conflict of interest*:*

The author declares that there is no conflict of interest.

References:

Adazabra, A. N., Viruthagiri, G., & Shanmugam, N. (2017). Infrared analysis of clay bricks incorporated with spent shea waste from the shea butter industry. *Journal of Environmental Management*, *191*, 66–74. https://doi.org/10.1016/j.jenvman.2017.01.006

Adilhodzhaev, A. I., Mahamataliev, I. M., & Shaumarov, S. S. (2018). Theoretical aspects of structural and simulation modeling of the macrostructure of composite building materials. *Nauchno-Tekhnicheskiy Vestnik Bryanskogo Gosudarstvennogo Universiteta*, *4*(3), 312–320. https://doi.org/10.22281/2413-9920-2018-04-03-312-320

Alim, M. A., Jahan, S., Rahman, A., Rahman, M. A., Liebman, M., Garner, B., Griffith, R., Griffith, M., & Tao, Z. (2023). Experimental investigation of a multilayer detention roof for stormwater management. *Journal of Cleaner Production*, *395*. https://doi.org/10.1016/J.JCLEPRO.2023.136413

Chen, S. H., Wang, H. Y., & Jhou, J. W. (2013). Investigating the properties of lightweight concrete containing high contents of recycled green building materials. *Construction and Building Materials*, *48*, 98–103. https://doi.org/10.1016/j.conbuildmat.2013.06.040

de Brito, J., & Robles, R. (2010). Recycled aggregate concrete (RAC) methodology for estimating its long-term properties. *Indian Journal of Engineering and Materials Sciences*, *17*(6), 449–462. https://doi.org/10.1063/1.4953383

de Vries, R., Hassell, R., & Player, J. (2015). Geotechnical design considerations for Dugald River - from slot to sequence. *Underground Design Methods*. http://www.mendeley.com/research/geotechnical-design-considerations-dugald-river-slot-sequence

Dong, X., Zhang, X., Zhang, C., & Bi, C. (2023). Building sustainability education for green recovery in the energy resource sector: A cross country analysis. *Resources Policy*, *81*. https://doi.org/10.1016/j.resourpol.2023.103385

Elheddad, M., Benjasak, C., Deljavan, R., Alharthi, M., & Almabrok, J. M. (2021). The effect of the Fourth Industrial Revolution on the environment: The relationship between electronic finance and pollution in OECD countries. *Technological Forecasting and Social Change*, *163*. https://doi.org/10.1016/j.techfore.2020.120485

Fairbairn, E. M. R., Americano, B. B., Cordeiro, G. C., Paula, T. P., Toledo Filho, R. D., & Silvoso, M. M. (2010). Cement replacement by sugar cane bagasse ash: CO2 emissions reduction and potential for carbon credits. *Journal of Environmental Management*, *91*(9), 1864–1871. https://doi.org/10.1016/j.jenvman.2010.04.008

Fapohunda, C., Akinbile, B., & Shittu, A. (2017). Structure and properties of mortar and concrete with rice husk ash as partial replacement of ordinary Portland cement – A review.

International Journal of Sustainable Built Environment, *6*(2), 675–692. https://doi.org/10.1016/j.ijsbe.2017.07.004

Hachemi, S., Khattab, M., & Benzetta, H. (2023). Enhancing the performance of concrete after exposure to high temperature by coarse and fine waste fire brick: An experimental study. *Construction and Building Materials*, *368*. https://doi.org/10.1016/j.conbuildmat.2023.130356

Han, S., Jang, I., Choi, E. K., Park, W., Yi, C., & Chung, N. (2020). Bacterial Self-Healing Performance of Coated Expanded Clay in Concrete. *Journal of Environmental Engineering*, *146*(7). https://doi.org/10.1061/(ASCE)EE.1943-7870.0001713

Hu, S.-C., Hu, S.-H., Fu, Y.-P., & Sie, S.-F. (2012). Lowered Temperature Resource Recycling of Paper Sludge Using a Co-Melting Technology. *Bioresources*, *7*(3), 2766–2783.

Hunag, L. J., Wang, H. Y., & Wang, S. Y. (2015). A study of the durability of recycled green building materials in lightweight aggregate concrete. *Construction and Building Materials*, *96*, 353–359. https://doi.org/10.1016/j.conbuildmat.2015.08.018

Liu, H., Wang, X., Jiao, Y., & Sha, T. (2016). Experimental investigation of the mechanical and durability properties of crumb rubber concrete. *Materials*, *9*(3), 1–12. https://doi.org/10.3390/ma9030172

Liu, P., Farzana, R., Rajarao, R., & Sahajwalla, V. (2017). Lightweight expanded aggregates from the mixture of waste automotive plastics and clay. *Construction and Building Materials*. https://doi.org/10.1016/j.conbuildmat.2017.04.009

Maderuelo-Sanz, R., García-Cobos, F. J., Sánchez-Delgado, F. J., Mota-López, M. I., Meneses-Rodríguez, J. M., Romero-Casado, A., Acedo-Fuentes, P., & López-Ramos, L. (2022). Mechanical, thermal and acoustical evaluation of biocomposites made of agricultural waste for ceiling tiles. *Applied Acoustics*, *191*. https://doi.org/10.1016/j.apacoust.2022.108689

Maraveas, C. (2020). Production of sustainable and biodegradable polymers from agricultural waste. *Polymers*, *12*(5). https://doi.org/10.3390/POLYM12051127

Nagase, A. (2020). Novel application and reused materials for extensive green roof substrates and drainage layers in Japan – Plant growth and moisture uptake implementation -. *Ecological Engineering*, *153*. https://doi.org/10.1016/J.ECOLENG.2020.105898

Ricciardi, P., Belloni, E., Merli, F., & Buratti, C. (2021). Sustainable panels made with industrial and agricultural waste: Thermal and environmental critical analysis of the experimental results. *Applied Sciences (Switzerland)*, *11*(2), 1–14. https://doi.org/10.3390/APP11020494

Roessler, R., Mpouam, S. E., Muchemwa, T., & Schlecht, E. (2016). Emerging development pathways of urban livestock production in rapidly growing West Africa cities. *Sustainability (Switzerland)*, *8*(11). https://doi.org/10.3390/SU8111199

Ruggerio, C. A. (2021). Sustainability and sustainable development: A review of principles and definitions. *Science of the Total Environment*, *786*. https://doi.org/10.1016/j.scitotenv.2021.147481

Sadek, D. M., El-Sayed, W. S., Heniegal, A. M. A., & Mohamed, A. S. (2013). Utilization of solid wastes in cement bricks for an environmental beneficial. *Annals of Faculty Engineering Hunedoara*. http://www.mendeley.com/research/utilization-solid-wastes-cement-bricks-environmental-beneficial

Sierra-Beltran, M. G., Jonkers, H. M., & Schlangen, E. (2014). Characterization of sustainable bio-based mortar for concrete repair. *Construction and Building Materials*, *67*(PART C), 344–352. https://doi.org/10.1016/J.CONBUILDMAT.2014.01.012

Sutcu, M., Ozturk, S., Yalamac, E., & Gencel, O. (2016). Effect of olive mill waste addition on the properties of porous fired clay bricks using Taguchi method. *Journal of Environmental Management*, *181*, 185–192. https://doi.org/10.1016/j.jenvman.2016.06.023

Tang, P., Florea, M. V. A., & Brouwers, H. J. H. (2016). Recycling Of Waste Materials As Lightweigth Aggregate By Cold Bonding Pelletizing Technique. *Proceedings of the Thirty-Eighth International Conference on Cement Microscopy*. http://www.mendeley.com/research/recycling-waste-materials-lightweigth-aggregate-cold-bonding-pelletizing-technique

Tay, J., Show, K., Hong, S., Chien, C., & Lee, D. (2003). Thermal Stabilization of Iron-Rich Sludge for High Strength Aggregates. *Journal of Materials in Civil Engineering*. https://doi.org/10.1061/ASCE0899-1561200315:6577

Thomas, J., & Harilal, B. (2016). Mechanical properties of cold bonded quarry dust aggregate concrete subjected to elevated temperature. *Construction and Building Materials*. https://doi.org/10.1016/j.conbuildmat.2016.08.093

Walker, R., & Pavía, S. (2014). Moisture transfer and thermal properties of hemp-lime concretes. *Construction and Building Materials*, *64*, 270–276. https://doi.org/10.1016/j.conbuildmat.2014.04.081

Xiong, B., Falliano, D., Restuccia, L., Di Trapani, F., Demartino, C., & Marano, G. C. (2023). High-strain rate compressive behavior of concrete with two different substituted recycled plastic aggregates: Experimental characterization and probabilistic modeling. *Construction and Building Materials*, *368*. https://doi.org/10.1016/j.conbuildmat.2022.130279

Xu, P., Hussain, M., Ye, C., Wang, J., Wang, C., Geng, J., Liu, Y., & Chen, J. (2022). Natural resources, economic policies, energy structure, and ecological footprints' nexus in emerging seven countries. *Resources Policy*, *77*. https://doi.org/10.1016/j.resourpol.2022.102747

Yan, C. Y., Min, Z. R., Ming, Q. Z., Chen, J., Gui, C. Y., & Xue, M. L. (2008). Self-healing polymeric materials using epoxy/mercaptan as the healant. *Macromolecules*, *41*(14), 5197–5202. https://doi.org/10.1021/MA800028D

Yaras, A., Sutcu, M., Gencel, O., & Erdogmus, E. (2019). Use of carbonation sludge in clay based building materials processing for eco-friendly, lightweight and thermal insulation. *Construction and Building Materials*, *224*, 57–65. https://doi.org/10.1016/j.conbuildmat.2019.07.080

Yu, Q. L., Spiesz, P., & Brouwers, H. J. H. (2015). Ultra-lightweight concrete: Conceptual design and performance evaluation. *Cement and Concrete Composites*, *61*, 18–28. https://doi.org/10.1016/j.cemconcomp.2015.04.012

Zhang, L. V., Nehdi, M. L., Suleiman, A. R., Allaf, M. M., Gan, M., Marani, A., & Tuyan, M.

(2021). Crack self-healing in bio-green concrete. *Composites Part B: Engineering*, *227*. https://doi.org/10.1016/J.COMPOSITESB.2021.109397

Zhao, L., & Rasoulinezhad, E. (2023). Role of natural resources utilization efficiency in achieving green economic recovery: Evidence from BRICS countries. *Resources Policy*, *80*. https://doi.org/10.1016/j.resourpol.2022.103164

Optimizing Electric Vehicle Efficiency with Real-Time Telemetry using Machine Learning

Aryaman Rao
dept. of Electrical Engineering
Delhi Technological University
Delhi, India
aryaman26601@gmail.com

Harshit Gupta
dept. of Mechanical Engineering
Delhi Technological University
Delhi, India
hgh1602@gmail.com

Parth Singh
dept. of Mechanical Engineering
Delhi Technological University
Delhi, India
parth55singh@gmail.com

Shivam Mittal
dept. of Applied Mathematics
Delhi Technological University
Delhi, India
shivammittal2124@gmail.com

Utkarsh Singh
dept. of Mechanical Engineering
Delhi Technological University
Delhi, India
utkarsh.an23o5@gmail.com

Dinesh Kumar Vishwakarma
dept. of Information Technology
Delhi Technological University
Delhi, India
dinesh@dtu.ac.in

arXiv:2311.08085v1 [eess.SY] 14 Nov 2023

***Abstract*—In the contemporary world with degrading natural resources, the urgency of energy efficiency has become imperative due to the conservation and environmental safeguarding. Therefore, it's crucial to look for advanced technology to minimize energy consumption. This research focuses on the optimization of battery-electric city style vehicles through the use of a real-time in-car telemetry system that communicates between components through the robust Controller Area Network (CAN) protocol. By harnessing real-time data from various sensors embedded within vehicles, our driving assistance system provides the driver with visual and haptic actionable feedback that guides the driver on using the optimum driving style to minimize power consumed by the vehicle. To develop the pace feedback mechanism for the driver, real-time data is collected through a Shell Eco Marathon Urban Concept vehicle platform and after pre-processing, it is analyzed using the novel machine learning algorithm TEMSL, that outperforms the existing baseline approaches across various performance metrics. This innovative method after numerous experimentation has proven effective in enhancing energy efficiency, guiding the driver along the track, and reducing human errors. The driving-assistance system offers a range of utilities, from cost savings and extended vehicle lifespan to significant contributions to environmental conservation and sustainable driving practices.**

***Index Terms*—Telemetry, Efficiency, Electric Vehicle, Sensor Technology, Machine Learning**

I. Introduction

In recent times, the significance of energy efficiency has grown monumentally, driven by the urgent necessity to address waste management and mitigate pollution for a sustainable environment. This alarming concern extends to various aspects of daily life, including the transportation industry [1]. The pollution emitted by the increasing number of gasoline vehicles present on the road is nearly unavoidable and the amount of greenhouse gases released by running vehicles standstill at a traffic light is an understatement to the impact they have on the environment [2]. Besides, how the car is controlled, particularly the way it accelerates, brakes, and maneuvers can have a significant impact on its overall driving efficiency. Even a minor malfunction in any component of the vehicle can cost its efficiency heavily. While significant research has been conducted on increasing vehicle efficiency [3], there is still scope for further improvement with the advent of artificial intelligence in the automotive industry.

With so many variables at play when driving, it's challenging for the driver to always operate under optimal conditions. That's where a real-time telemetry system comes into play. It can assist the driver by providing important insights about the behaviour and performance changes of the car. Our system utilizes different sensors to gather raw data during various runs and after thorough data analysis, we gain an understanding of the car's behavior, its component interactions, and especially energy consumption.

The Shell Eco-Marathon is an annual global student competition where student teams design and fabricate their vehicles to compete against each other to achieve the highest possible energy efficiency. The Urban Concept class represents a replica of modern-day passenger vehicles and is one of the different categories of vehicles eligible for the competition. , which are predominantly designed to attain exceptional energy-efficiency. As shown in the Fig. 2, our battery-electric Urban Concept vehicle participated in the International Shell Eco-Marathon Asia-Pacific Competition held on October 2022 at Lombok, Indonesia.

Through this paper, we delve into the progression and working of the telemetry system with emphasis on the novelty and research characteristics of the system employed in a battery-powered Urban Concept vehicle. Stress is given on how this evidence-based methodology, fortified by machine learning (ML) integration, can lead to a more efficient racetrack experience. This research work is part of a larger effort to save energy and make urban travel better for a more promising future. Our concept received the

2^{nd} prize in the Data and Telemetry Award, sponsored by Shell and Schmid Elektronik, acknowledging the innovative contribution and advocating the use of data to achieve sustainable mobility.

The subsequent sections are organised as follows. Section II involves literature review and finding out existing research gaps that our work suffices while in Section III, we explain the set up and methodology of our experiment which includes details about hardware, driving strategies, and the onboard systems. In Section IV, we provide information about the various data processing techniques and machine learning models utilized and in Section V, we have present the simulation outcomes and key findings, followed by conclusions in Section VI.

Fig. 1. Test Vehicle

II. Literature Review

With electric vehicles (EVs) taking centre stage in the consumer market, significant advancements have been made in enhancing their efficiency through exterior remodelling or introducing semi-autonomous (upto Level 3) driving capabilities [4]. These strategies utilize diverse vehicle data points in order to achieve any efficacious results [5]. With the current market trend being driver engagement minimization while driving, little research is being done on learning based predictive control strategies.

In the context of forward-looking driving assistance for energy efficiency, systems that provide drivers with insights about speed limits have been developed. This early awareness helps drivers avoid undesired braking losses and offers guidance for energy-efficient driving [6], [7]. However, a common challenge with most modern methods is that they rely on internet access for real-time updates or access to traffic conditions and might not work satisfactorily in areas with poor internet connectivity. Our telemetry system builds upon this by using CAN protocol for communication instead of popular wireless protocols such as Bluetooth, WiFi and LTE [8]–[10]. An interesting area to explore here involves developing driver assistance systems that promote efficient driving. These systems use techniques like shared control, adjusting air-to-fuel ratios, and adapting speed based on the type of terrain [11], [12]. Multiple studies have even tried to classify different driving patterns and environments [13]. Yet, drivers often lack actionable feedback to help them improve driving efficiency.

With the recent advances in Artificial Intelligence over the past few decades, many researchers have been eager to incorporate them into various fields. Some researchers have dived into using neural networks to improve vehicle efficiency, as discussed in [14]. Several studies [15] have been conducted on extracting valuable information using feature engineering on time series problems. Ensemble modelling has become quite popular in predicting patterns in telemetry or time-variant data, as demonstrated in [16].The potential of integrating onboard sensors and real-time data processing is intriguing for the realm of vehicle efficiency. These sensors are vital in delivering critical insights into factors such as the road surface characteristics, traffic density, and gradient variations. These critical insights facilitate the implementation of precise and context-aware energy management strategies [17], [18], ultimately contributing to the improvement in operational efficiency of the vehicle. Harnessing this data empowers vehicles to dynamically fine-tune power distribution, optimize regenerative braking mechanisms, and provide optimal route suggestions. Furthermore, the development of vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication systems [19] presents a promising avenue for further vehicle efficiency optimization. These systems facilitate seamless data exchange on aspects such as speed, location, and driver intent, enabling the deployment of collaborative strategies aimed at minimizing energy-depleting actions such as abrupt lane changes and sudden acceleration. The pace feedback mechanism proposed by us expands upon the existing literature by providing an accessible haptic and visual feedback interface based on robust CAN communication enabling the driver to easily maximize vehicle efficiency by altering their driving style.

III. Data Acquisition

The real-time data acquisition setup for research into the optimum driving strategy consists of various elements such as sensors ranging from accelerometers to joule-meters and a robust network. Various key aspects such as the layout of the track, the driving methods employed, and the inclusion of special equipment for real-time telemetry on the vehicle are discussed. An insight into the methods employed to acquire data is necessary for the reliability and dependability of the mechanism.

A. Test Track

To test the working of the telemetry system, we had to perform a number of test laps in the Urban Concept vehicle mimicking the actual track to be encountered in Shell Eco Marathon. The test track is shown in the Fig.2 highlighted in blue. The route was created inside the Delhi Technological

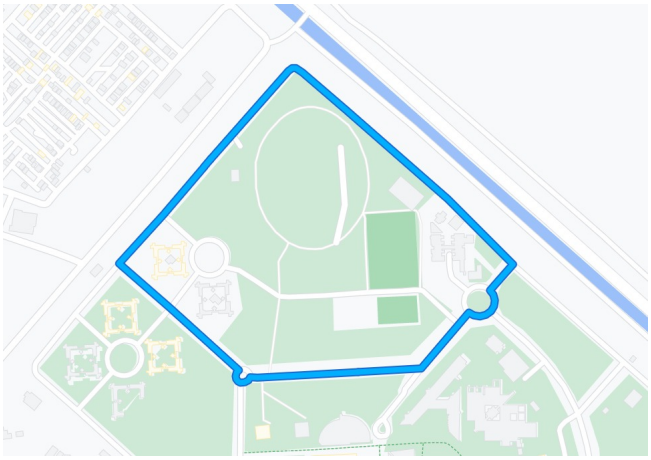


Fig. 2. Test Track

University campus with two objectives in mind.

1) Track Length: The length of the track is chosen as close as possible to the Shell Eco-Marathon tracks, to provide data that will be helpful on track. The average length of a Shell Eco-Marathon track is 1.4 km therefore, the route chosen to be covered was about 1.5 km long.

2) Number of Corners: This is a critical parameter to understand and devise an optimum pattern with the objective of minimizing energy consumption. Our test track had 8 corners over which the vehicle was tested for 24 laps in slightly less than 2 hours wherein driving strategies were changed after every 4 laps. Data was collected every 100 ms for the entire run, amounting to more than 10,000 data points which were split in the ratio of 7:3, comprising of 70% data for training and 30% data for testing.

B. Driving Techniques

Driving behaviors and proper situation-dependent techniques can save energy consumption in a dynamic manner as it affects the rate of energy consumption and gas emission as much as the manufacturing technologies involved. These habits have significant effects on reducing energy consumption in EVs. [20]

1) Acceleration Variation Method (Table I): In EVs, different driving styles can have great impact on the vehicle's battery life, efficiency, and the SOC gradient. Before testing the vehicle some assumptions were made: there should be minimal traffic on the track and acceleration and jerk should only be done on the straight line paths, not around the corners. The different acceleration patterns that were tested are as follows:

- Aggressive: Acceleration at a rapid rate
- Gentle: Acceleration at an optimal rate
- Mild: Acceleration at a slow rate

Several test runs were conducted for different patterns to plot a graph between acceleration and the power required from battery.

TABLE I
ACCELERATION VARIATION METHOD

Driving Style	Extra Power Required (As compared to Gentle)	Energy Consumed (As compared to Gentle)
Aggressive	46.6 %	41.5 %
Mild	30.0 %	15.7 %

TABLE II
BURN AND COAST METHOD

Driving Technique	Efficiency Increase
Full burn and coast	15 %
Mild Burn and coast	8.5 %

2) Burn and Coast Method: (Table II) : It is a very popular driving method wherein the vehicle is accelerated to a particular speed, and the car is allowed to glide by switching off the engine till the vehicle speed drops to a set minimum velocity from where the motor is started again. Since our vehicle is battery-electric, we do not switch off the motor and we ensure to maintain a constant average velocity on straight paths as much as possible. The frequent switching on-and-off of the motor might induce high currents in the motor, consequently affecting it's efficiency, damaging it or harming the driver and the vehicle. Two techniques of burn and coast method have been investigated in the research study namely "Full" burn and coast and "mild" burn and coast. [21]

C. Onboard Telemetry System

Sensors used to collect various data points for understanding the effect of driving techniques on the energy consumption in real-time are mentioned in Table III which are connected to their respective Arduino Uno modules.

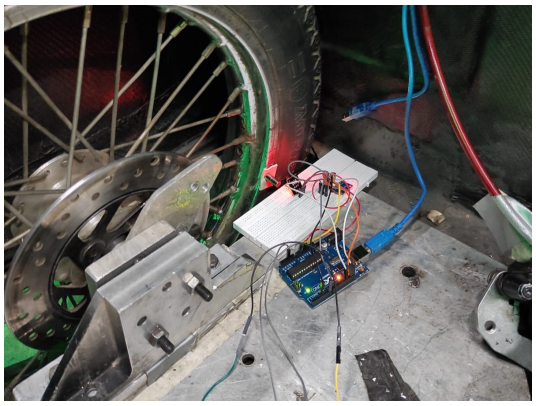


Fig. 3. Raspberry Pi Setup in our Test Vehicle

Fig. 4. Working of the LEDs Pace Feedback on the Steering Wheel

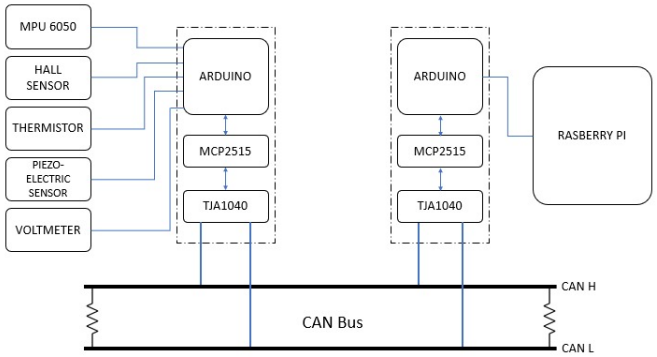


Fig. 5. Block diagram depicting sensor-to-Raspberry Pi communication via CAN bus protocol, intermediated by Arduino for seamless data exchange

The nucleus of the entire telemetry system is a Raspberry Pi 4-4GB RAM board connected to the steering wheel. A 5-inch touch screen is also attached to the steering wheel to allow the driver to interact with the interface. Sensors can be connected directly using the I2C communication to the Raspberry Pi, or through the CAN Protocol as shown in figure 5. Sensors act as the clients and the Pi board acts as the centralized server, hence forming the Star Topology network. The CAN protocol is used to provide a robust communication between a controller and transceiver. MCP2515 acts as the CAN controller capable of transmitting and receiving data at high speeds (1 Mbps) while the TJA1040 creates an interface between the CAN controller and the bus. The combined data is sent over the CAN bus to the Raspberry Pi, where it is separated into individual data points. The steering wheel (Fig. 4) has been programmed to provide both visual and haptic feedback to the driver, using LEDs and two vibration motors under the finger rests; this system has been dubbed as the "Pace Feedback Mechanism". Slow velocity and acceleration are represented by RED lights with low intensity-low frequency haptics, high velocity and acceleration are indicated by GREEN lights with high intensity-high frequency haptics, and the optimal pace is denoted by BLUE with no haptic feedback.

TABLE III
ONBOARD SENSORS FOR DATA ACQUISITION

Sr No.	Sensors	Functionality
1	MPU 6050 [22]	Accelerometer & Gyroscope
2	Hall Sensor [23]	Speedometer & Tachometer
3	NEO-6m Sensor [24]	GPS receiver
4	Piezoelectric Sensor [25]	Pressure
5	Motor-Controller	Current & Voltage from Battery.
6	Joulemeter	Energy Consumed
7	NTC Thermistor [26]	Temperature

IV. PROPOSED METHODOLOGY & EVALUATION

A. Data Mining

Data mining research has made significant strides in temporal data analysis encompassing methods such as data representation, dimensionality reduction, and segmentation. In this study, we have integrated a comprehensive data mining framework, as explicated in the work of [27]. This framework was instrumental in exploiting the power of data mining, that leveraged various ML techniques, as cited by [28], [29], to extract the crucial information and insights from raw datasets. In accordance with the observations made by I. K. Nti [30], the omnipotence of data emphasizes the necessity for preprocessing methodologies that can adeptly harness information and knowledge. In our research, we have systematically partitioned this section into three integral components: Data Preprocessing, Modelling Framework, and Model Evaluation.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{1}$$

Data preprocessing is a fundamental step in data analysis, essential for elevating the quality and reliability of predictive models. This process covers attributes like outlier rectification, management of missing data, and the normalization of features to ensure compatibility. Its importance is evident in real-world applications, where data often suffers from contamination by noise and intricate patterns. Through the application of such techniques, it can curtail the training complexity and significantly augment the efficacy of predictive analysis of telemetry data. A prevailing issue in recent ML models has been that of high dimensionality. To combat this drawback in large datasets, we apply feature extraction techniques to transform the data into a lower dimensional vector, making the model more efficient. We employ Principal Component Analysis (PCA) [31], which is an unsupervised ML algorithm, that reduces dimensionality and improves performance for our experimentation.

TABLE IV
PRACTICE RUN RESULTS

Sr No.	Driving Style	Method	Velocity (km/h)	Current (A)	Efficiency (km/kWh)
1	Acceleration	Aggressive	43.2-10.8	40	62.5
		Gentle	36.0-12.2	32	107.1
		Mild	25.0-10.5	25	78.9
2	Burn & Coast	Full	32.0-15.3	27	150.2
		Mild	23.4-10.8	20	125.5
3	Pace Feedback	**TEMSL Burn & Coast**	28.6-15.1	24	166.7

Feature selection also plays a crucial role in enhancing the performance of multitude of ML models, assisting extracting relevant information while minimizing redundancy. It involves calculating scores for each feature and selecting the top-k features that are most beneficial for the model. Out of the two primary methods explored, namely the Chi-Square method [32] and Minimum Redundancy Maximum Relevance, we opt for the Chi-Square method. The Chi-Square or χ^2 method (Eq. 1) employs a statistical test to assess how features deviate from an expected distribution. In the Eq. 1, O and E represents observed and expected frequency/value respectively. χ^2 assigns a variance-based score to each feature, indicating its relevance relative to other features, effectively selecting the best-k features through a threshold mechanism.

B. Modelling Framework

A deep learning (DL) framework has been applied for our proposed heterogeneous ensemble model [33], [34] to predict the optimal velocity of our vehicle. A heterogeneous ensemble model is a ML approach that combines predictions from multiple diverse base algorithms. This diversity enhances the overall model's performance and robustness by reducing over-fitting. We have implemented our ensemble model (TEMSL), acronym for Tree-based Ensemble Model with stacked LSTM algorithm, using three primary algorithms: Random Forest Regressor, Extra Trees Regressor, and custom Stacked Deep LSTM Network using bootstrap aggregation. The Random Forest excels in handling structured data and complex interactions, Extra Trees Regression is known for incorporating randomization of decision trees to enhance model robustness, and LSTM networks are proficient in modelling sequential data with memory retention. The proposed model is typically compared across three major baseline approaches, as listed below.

- CatBoost Regression: CatBoost is a ML supervised algorithm known for combining ordered boosting trees, which prevents overfitting to deliver enhanced results.
- Stacked Deep LSTM Network: This is a novel DL approach that employs multiple LSTM layers which exceeds in capturing sequential dependencies.
- Linear Regression (LR): LR is the most popular supervised regression algorithm that captures the linear relationship between the features and data labels.

The architecture or framework of our proposed model has been illustrated in the Figure 6. Our proposed method performs well on all performance metrics, surpassing the baseline algorithms utilized in this research study.

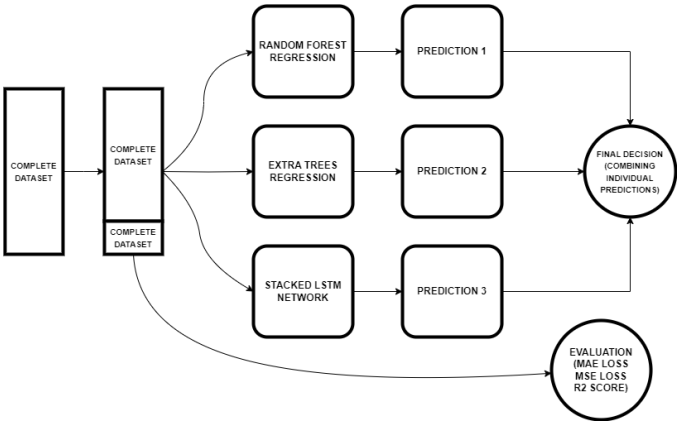


Fig. 6. Model Architecture

C. Model Evaluation

The experimentation process was done using the TensorFlow Keras deep learning library [35] in Python Language. All the algorithms were trained under the exact hyperparameters as the other baseline methods. The dataset comprises of more than 14000 entries, which were finally inhibited to a few relevant features using the feature extraction and selection techniques. The split ratio in the dataset is 0.3, which comprised of 70% data for training and 30% data for testing. After extensive training process, the experimented algorithms were evaluated fairly on the test samples, as shown in Table V, using the following evaluation metrics.

- Mean Absolute Error (MAE): Measures the absolute magnitude of errors between predicted and actual values to provide a judgement of model accuracy, with smaller MAE values indicating better performance.
- Mean Squared Error (MSE): Computes the mean of the squared discrepancies between forecasted and real values. It heavily penalizes bigger mistakes, which makes it susceptible to anomalies.
- R^2 Score (R^2): Quantifies how well a model explains the variance in the data. It ranges between 0 and 1, with higher values indicating a better fit. R^2 helps evaluate the fitness of regression models.

V. RESULTS

Leveraging a dataset collected through rigorous test track practice sessions, we introduced the innovative TEMSL approach, focusing on accurately predicting the optimal vehicle speed to optimize energy efficiency.

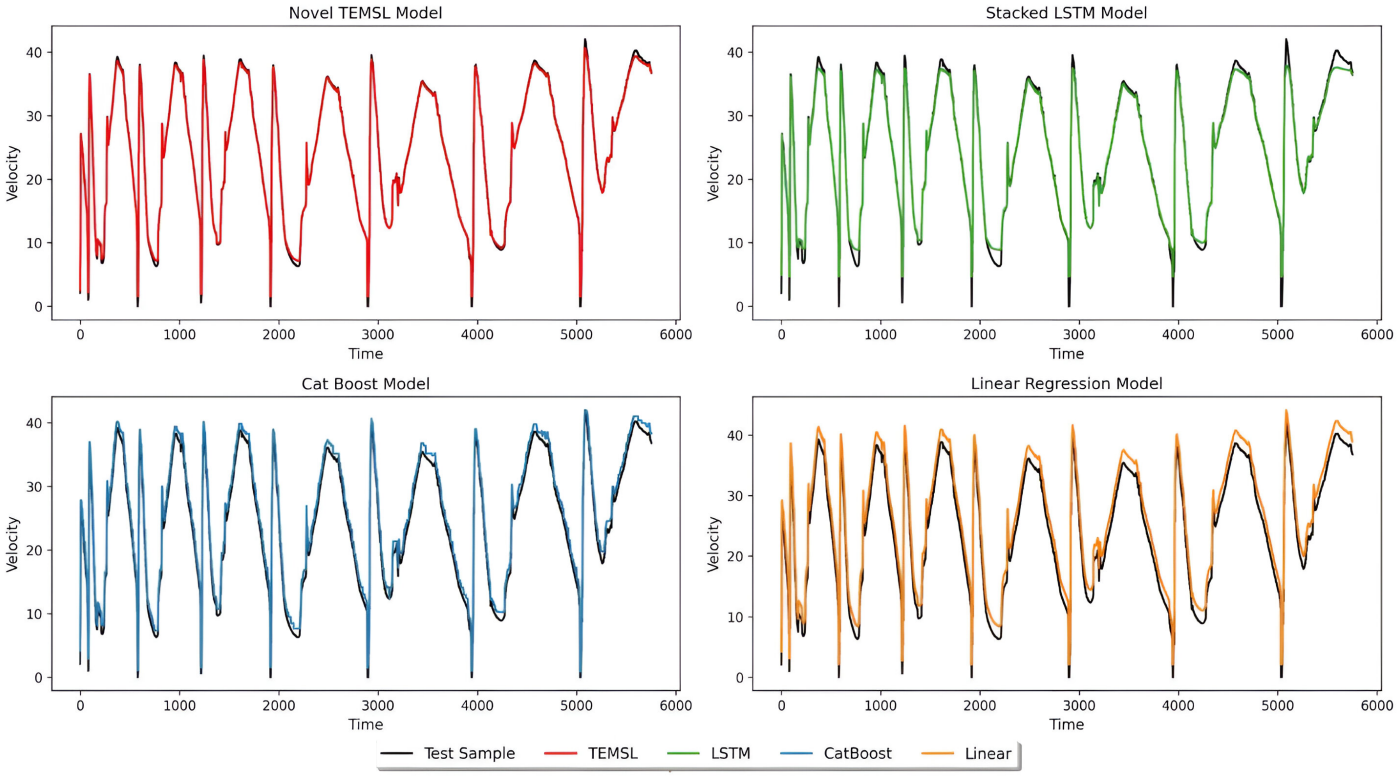


Fig. 7. Graphical Representation of the True and Predicted values using the various baseline algorithms

After meticulous experimentation (Table IV, V), TEMSL outstripped the existing baseline approaches across various performance metrics. Most notably, TEMSL achieved an exceptional validation accuracy (R^2-Score) of 99.85%, affirming its potential to significantly enhance energy efficiency in real-world vehicular applications. The visual representation of the comparison between models is illustrated in Fig. 7. The results from the model seemed favourable to our goal of achieving the maximum energy-efficiency possible, ultimately sending feedback to the driver to precisely follow the chosen strategy.

By installing this model on the Raspberry Pi, which controls the Steering Wheel (Feedback Mechanism) and applying the TEMSL algorithm with the full Burn and Coast driving method, we observed significant improvements in efficiency. After training and testing on the practice run data, the new test data from the final run was fed into the algorithm, and achieved an accuracy of 99.85%. The model compared the predicted value with the actual value. When the velocity exceeded the reference speed by 5 km/h, a green light illuminated, signalling the driver to coast (decelerate) for optimal speed. In the alternate case, the light turned red. If the actual velocity was within the range of ±2.5 km/h of the predicted value, the Pace Feedback turned blue. The best efficiency achieved by the vehicle was 166.7 km/kWh, which was 26.5 km/kWh more than the total Burn & Coast efficiency.

TABLE V
PERFORMANCE METRICS

Algorithms	**MAE**	**MSE**	**R^2 Score**
TEMSL	0.17	0.08	99.52%
Stacked LSTM	0.46	0.56	98.15%
CatBoost	0.71	0.92	96.65%
Linear Regression	3.12	14.18	86.45%

VI. CONCLUSION

Through the process of data acquisition, exploratory data analysis, and computation, the study of the effect of driving techniques on the overall vehicle efficiency was carried out successfully. The onboard real-time telemetry system is actively assisting the driver in adjusting driving patterns for minimum energy consumption. The novel approach, TEMSL, assisted us in accurately selecting the suitable driving approach based on sensor and haptic feedback during track driving. Using the suggested Pace feedback mechanism equipped with TEMSL approach, an increase in energy efficiency of about 10.6% was recorded as compared to the closest alternative technique, thus validating our hypothesis. The current research focuses and analyzes primarily two driving patterns commonly used for minimizing energy consumption. However, since our research is centered on just two driving patterns, further research work

needs to be conducted to impact the energy economy by the use of intelligent systems for various types of driving behaviors. The model also has the potential to contribute to efficiency enhancements on real city roads, beyond our chosen test track. This research can foster a broader societal transition towards cleaner and greener transportation solutions by demonstrating the successful implementation of the mechanism in enhancing the performance and energy efficiency of EVs.

ACKNOWLEDGMENT

This work was supported and funded by Delhi Technological University (DTU), New Delhi, India. Furthermore, we extend our gratitude to Shell Eco-Marathon Committee for providing us an international platform to showcase our research innovations.

REFERENCES

[1] Masoumi, Houshmand E., and Ahmadreza Faghih Imani. "Energy Consumption and Transportation in Developing Countries: Need for Local Scenario-Based Energy Efficiency Plans." Sustainability 7, no. 3 (2015): 2560-2583

[2] B. Bolaji and S. B. Adejuyigbe, "Vehicle Emissions and their Effects on the Natural Environment - a Review," Journal of the Ghana Institution of Engineers, vol. 4, pp. 35-41, 2006.

[3] Canan G. Corlu, Rocio de la Torre, Adrian Serrano-Hernandez, Angel A. Juan, Javier Faulin, "Optimizing Energy Consumption in Transportation: Literature Review, Insights, and Research Opportunities", Energies (Multidisciplinary Digital Publishing Institute)-Vol. 13 (2020)

[4] D. Parekh et al., "A Review on Autonomous Vehicles: Progress, Methods and Challenges," Electronics, vol. 11, no. 14, p. 2162, Jul. 2022

[5] K. Vasudevan, A. P. Das, Sandhya B and Subith P, "Driver drowsiness monitoring by learning vehicle telemetry data," 2017 10th International Conference on Human System Interactions (HSI), Ulsan, Korea (South), 2017, pp. 270-276.

[6] Tobias Bär, Ralf Kohlhaas, J. Marius Zöllner and Kay-Ulrich Scholl, "Anticipatory driving assistance for energy efficient driving," Integrated and Sustainable Transportation System (FISTS), 2011 IEEE Forum , June 2011.

[7] Tianyi Guan and Christian W. Frey, "Fuel efficiency driver assistance system for manufacturer independent solutions," 2012 15th International IEEE Conference on Intelligent Transportation Systems, September 2012.

[8] Aris Triwiyatno, Suroto Munahar, Munadi Muna and Joga Dharma Setiawan, "Application of driving behavior control system using artificial neural network to improve driving comfort by adjusting air-to-fuel ratio", IIUM Engineering Journal 24(2), pp. 337–353, July 2023.

[9] Tianyi Guan and Christian W. Frey, "Model adaptive driver assistance system to increase fuel savings," Vehicular Electronics and Safety (ICVES), 2012 IEEE International Conference, July 2012.

[10] Sandareka Wickramanayake and H.M.N. Dilum Bandara, "Fuel consumption prediction of fleet vehicles using Machine Learning: A comparative study," 2016 Moratuwa Engineering Research Conference (MERCon), April 2016.

[11] Erik Hellström, Maria Ivarsson, Jan Åslund and Lars Nielsen, "Look-ahead control for heavy trucks to minimize trip time and fuel consumption," Advances in Automotive Control, IFAC Proceedings Volumes 40(10), pp. 439–446, August 2007.

[12] Jose del C. Julio-Rodriguez, Alfredo Santana, Carlos A. Rojas-Ruiz, Rogelio Bustamante-Bello and Ricardo A. Ramirez-Mendoza, "Environment Classification Using Machine Learning Methods for Eco-Driving Strategies in Intelligent Vehicles," Applied Sciences 12(11):5578, May 2022.

[13] Javier E. Meseguer, Carlos T. Calafate, Juan Carlos Cano and Pietro Manzoni, "Characterizing the Driving Style Behavior using Artificial Intelligence Techniques," 38th IEEE Conference on Local Computer Networks (LCN 2013), October 2013.

[14] Mosabbir Bhuiyan and Md Ahasan Kabir, "Vehicle Speed Prediction based on Road Status using Machine Learning," Advanced Science, Engineering and Medicine 2(1), pp. 1–9, January 2020.

[15] Y. Li and Y. Pan, "A novel ensemble deep learning model for stock prediction based on stock prices and news," Int J Data Sci Anal, vol. 13, pp. 139–149, 2022.

[16] S. Ravikumar and P. Saraf, "Prediction of Stock Prices using Machine Learning (Regression, Classification) Algorithms," pp. 1-5, 2020.

[17] Ralf Kohlhaas, Thomas Schamm, Dennis Nienhüser and J. Marius Zöllner, "Anticipatory energy saving assistant for approaching slower vehicles," 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), October 2011.

[18] Sandareka Wickramanayake and H.M.N. Dilum Bandara, "Fuel consumption prediction of fleet vehicles using Machine Learning: A comparative study," 2016 Moratuwa Engineering Research Conference (MERCon), April 2016.

[19] K. C. Dey, A. Rayamajhi, M. Chowdhury, P. Bhavsar, and J. Martin, "Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication in a heterogeneous wireless network – Performance evaluation," Transportation Research Part C: Emerging Technologies, vol. 68, pp. 168-184, 2016. ISSN: 0968-090X.

[20] M. B. Younes, "Towards Green Driving: A Review of Efficient Driving Techniques," World Electric Vehicle Journal, vol. 13, no. 6, p. 103, Jun. 2022.

[21] Jeongwoo Lee , "Vehicle Inertia Impact on Fuel Consumption of Conventional and Hybrid Electric Vehicles Using Acceleration and Coast Driving Strategy", Virginia Tech, March 2014.

[22] Syed shujat Ali, "MPU-6050 (Accelerometer and Gyroscopic sensor)," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, July 2021.

[23] Marco Crescentini, Sana Fatima Syeda and Gian Piero Gibiino, "Hall-Effect Current Sensors: Principles of Operation and Implementation Techniques," IEEE Sensors Journal, Vol. 22, Iss. 11, pp. 10137–10151, June 2022.

[24] Syed shujat Ali, "NEO-6M-0-001(GPS module)," National University of Sciences and Technology, Formula Student Team, July 2021.

[25] Li Tianze, Zhang Xia, Jiang Chuan and Hou Luan, "Analysis of the characteristics of piezoelectric sensor and research of its application," 2009 18th IEEE International Symposium on the Applications of Ferroelectrics, pp. 1-4, August 2009

[26] DTutunea Dragos, Ilie Dumitru, George Gherghina and Alexandru Dima, "Evaluation of Thermistors Used for Temperature Measurement in Automotive Industry," Applied Mechanics and Materials, vol. 880, pp. 157–162, March 2018.

[27] S. H. Liao, P. H. Chu, and P. Y. Hsiao, "Review: Data mining techniques and applications - A decade review from 2000 to 2011," Expert Systems with Applications: An International Journal, vol. 39, pp. 11303-11311, 2012.

[28] P. Taylor, N. Griffiths, A. Bhalerao, S. Anand, T. Popham, Z. Xu, and A. Gelencser, "Data Mining for Vehicle Telemetry," Applied Artificial Intelligence, vol. 30, no. 3, pp. 233-256, 2016.

[29] Z. Constantinescu, C. Marinoiu, and M. Vladoiu, "Driving style analysis using data mining techniques," International Journal of Computers Communications & Control, vol. 5, no. 5, pp. 654–663, 2010.

[30] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," J Big Data, vol. 7, p. 20, 2020.

[31] R. Rosipal, M. Girolami, L. Trejo, et al., "Kernel PCA for Feature Extraction and De-Noising in Nonlinear Regression," NCA 10, pp. 231–243, 2001.

[32] S. Thaseen and A. K. Cherukuri, "Intrusion Detection Model Using fusion of Chi-square feature selection and multi class SVM," Journal of King Saud University - Computer and Information Sciences, vol. 29, 2016.

[33] N. Sharma, J. Dev, M. Mangla, V. Wadhwa, S. Mohanty, and D. Kakkar, "A Heterogeneous Ensemble Forecasting Model for Disease Prediction," New Generation Computing, vol. 39, 2021.

[34] M. Re and G. Valentini, "Ensemble methods: A review," 2012.

[35] J. F. J. Joseph, S. Nonsiri, and A. Monsakul, "Correction to: Keras and TensorFlow: A Hands-On Experience," in Advanced Deep Learning for Engineers and Scientists, edited by K. B. Prakash, R. Kannan, S. Alexander, G. R. Kanagachidambaresan, EAI/Springer Innovations in Communication and Computing, Springer, Cham, 2021.

Cognitive Computation
https://doi.org/10.1007/s12559-023-10227-3

Optimizing Sentiment Analysis: A Cognitive Approach with Negation Handling via Mathematical Modelling

Neha Punetha[1] · **Goonjan Jain**[1]

Received: 28 June 2023 / Accepted: 12 November 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Negation handling is a crucial aspect of sentiment analysis, as it presents challenges to accurate sentiment classification by altering polarity and reducing reliability. Traditional lexicon-based approaches often lack adequate techniques for modeling negation and fail to identify the appropriate negation window. Moreover, building machine learning models for negation handling in conversational text data proves difficult due to the intricate syntactic structure of negation. To address these issues, we propose a novel unsupervised cognitive sentiment classification approach. Our research introduces the multi-criteria decision-making (MCDM)–based "Negation Handling of the Text Using the VIKOR Optimization Technique" (NEGVOT) model, which effectively handles negation in sentiment analysis. By employing the decision science method, the NEGVOT model provides a solution for correctly labeling text sentiment in both negation-free and negation-containing texts. Our approach utilizes a lexicon database to obtain context scores of textual comments and integrates emotional scores to achieve accurate sentiment classification. Through experiments conducted on three benchmarked datasets, we demonstrate that the NEGVOT model performs comparably to state-of-the-art models. The NEGVOT model achieves the accuracy of 83%, 85%, and 82% over three datasets. It significantly enhances the accuracy of sentiment orientation tagging by effectively handling sentences with negation. We employ statistical analysis to support the relevance of our findings. The NEGVOT paradigm ensures logical and consistent results while exhibiting a strong generalization capacity, enabling sentiment classification for texts containing negations. This study makes notable contributions to the advancement of unsupervised techniques and provides a robust framework for handling negation in sentiment classification tasks.

Keywords Negation handling · MCDM · Social media response · VIKOR

Introduction

Sentiment analysis is the process of identifying, extracting, and quantifying human emotions and opinions from text data and is becoming an increasingly important tool in humanities and social communication research. It plays a crucial role in social and human science discourses because it reveals individuals' and societies' emotions, attitudes, and beliefs. In these disciplines, evaluating the meaning and analyzing cultural and social situations is crucial to paying attention to the emotions conveyed in texts, speeches, and other ways people interact [1]. It can be used as a valuable tool that can be used to analyze public attitudes and emotions towards various social and political issues and who seek a deeper understanding of how emotion and sentiment impact human interactions and society's culture [2].

As social media usage increased, Facebook, Twitter, Instagram, and YouTube were extensively researched, increasing Internet comments, reviews, and ratings. Analyzing massive data available online and manually, sentiment analysis is more challenging. While sentiment analysis is a powerful tool for understanding a text's emotional tone, several challenges can reduce its efficiency and accuracy. Some of the challenges of sentiment analysis include ambiguity of language, sarcasm, irony, negation, and modifiers; the accuracy of sentiment analysis heavily depends on the quality and quantity of data used to analyze the models. If the data used is biased or too small, the results may not be reliable. People express their

✉ Goonjan Jain
jaingoonjan@outlook.com

Neha Punetha
nehapunetha80@gmail.com

[1] Department of Applied Mathematics, Delhi Technological University, Delhi, India

Published online: 25 November 2023

sentiments differently, including non-verbal cues such as emojis, images, or videos. Sentiment analysis tools may not always be able to analyze these variations accurately. These challenges highlight the need for continual improvement and adaptation of sentiment analysis tools to ensure they remain effective and accurate in different contexts and languages. In this article, we focus on dealing with the negation-handling challenges of sentiment analysis [3].

Negation handling is an important area of research in sentiment analysis because it can significantly improve the accuracy of sentiment analysis results. Negation is a phenomenon of natural language in which the meaning of statements is inverted. The polarity of a word is altered, and the text's tone is shifted since positive sentences are typically transformed into negative ones. An essential part of sentiment analysis in NLP, negation handling is also one of the most challenging issues in NLP for opinion mining. Negation refers to using words like "not" or "never" that can reverse the meaning of a sentence or phrase. For example, "I love this product" and "I do not love this product" have opposite sentiments [4]. If negation is not appropriately handled in sentiment analysis, it can lead to inaccurate results, as the sentiment of a sentence can be misclassified. However, recent research has shown that syntactic parsing, sentiment shifters, and machine learning algorithms can improve negation handling, leading to better results in sentiment analysis [5]. Another approach to negation handling is to use sentiment shifters, words that can modify the sentiment of a sentence. For example, in the sentence "I hate this product less than I used to," the phrase "less than I used to" acts as a sentiment shifter, indicating that the sentiment is less negative than before. By identifying sentiment shifters, sentiment analysis tools can better account for the nuances of sentiment in a sentence [6]. Below are a few examples of the text that contain the negations (not, don't, couldn't) [7].

- You are not a lazy student.
- I don't take the trash out.
- Alex couldn't get unhealthy food.
- "I do not hate this product".

Overall, research in negation handling can lead to better sentiment analysis results by improving the accuracy and reliability of sentiment analysis tools. By identifying and accounting for negation in a sentence, sentiment analysis tools can provide more accurate and nuanced insights into the sentiment of a piece of text, which can help businesses and organizations make more informed decisions. By incorporating negation handling into sentiment analysis, cognitive computation systems can enhance their ability to understand complex language constructs and provide more accurate results. So the major contributions of this cognitive framework are given below.

Contribution

1. This study presents a sentiment analysis tagger capable of effectively handling sentences with and without negation in the text. The research evaluates the feasibility of a multiple-criteria decision-making (MCDM) model called "Negation Handling of the Text Using the VIKOR Optimization Technique" (NEGVOT) for sentiment analysis. The NEGVOT model enables accurate sentiment classification in both negation-containing and negation-free sentences, providing a comprehensive solution for sentiment analysis tasks.
2. The proposed approach employs a mathematical optimization technique, eliminating the need for pre-training in sentiment tagging. Importantly, it is designed to be domain-independent, enabling its application to datasets from diverse domains without requiring modifications.
3. The proposed approach achieves a high level of accuracy while optimizing implementation time and space requirements. Comparative evaluations demonstrate a significant 2–3% improvement in accuracy compared to other mathematical models, resulting in more precise sentiment orientation prediction. Furthermore, the NEGVOT model exhibits both reliability and adaptability, making it well-suited for sentiment analysis tasks in low-resource languages and easily extendable to such linguistic contexts.

Framework

The following summarizes this paper's structure: the "Related Work" section provides a summary of prior research. The "Methodology" section contains the preliminary associated with the methodology. The "Result and Discussion" section discusses the performance of the NEGVOT model for negation handling and compares it to that of previous approaches on three datasets. Statistical significance was also used to validate the results. The "Discussion" and "Conclusion" sections summarize significant findings, analyze errors, recommend further research, and conclude the study.

Related Work

An early effort to deal with negation modifiers was made by Das and Chen, who added "_n" as a negation marker to the feeling word. For instance, the term-document vectors for "not good" have "good n" as its corresponding value. Although negation modifiers were considered, the number of word vectors roughly doubled.

Taboada et al. [8] propose a lexicon-based method. The proposed model integrates intensification and negation and

uses dictionaries of words annotated with their semantic orientation (polarity and strength). Also, we detail how dictionaries are built and how we use Mechanical Turk to ensure their accuracy and consistency. Limsopatham et al. [9] propose a term representation for negated language in medical records that considers query term dependence. The TREC Medical Records 2011 search task evaluates our negation handling technique. The existing training set was augmented by flipping the data. The original and inverted training sets were used to fine-tune the classification model.

In another study, Xia et al. [10] suggest a dual training algorithm that utilizes original and inverted training reviews in pairs to create a sentiment classifier and a dual prediction algorithm that classifies the test reviews by considering both sides of a single review. The outcomes show that DSA works well for supervised sentiment classification. To better develop a sentiment classifier, we offer a dual training approach that uses pairs of original and reversed training reviews and a dual prediction algorithm that considers both sides of a review when classifying a test review. These findings prove that DSA is superior to other supervised methods for classifying emotion. Xia et al. [11] introduced the framework of document-level sentiment categorization; we suggest a three-stage cascade approach to handle the polarity shift problem. Multiple experimental findings show that our technology surpasses competitors in detecting and removing polarity shifts.

An evaluation and proposal of a model for managing negation modifiers for sentiment classification were made in this paper. The suggested negation modifier handling methodology was tested on a dataset comprised of movie scripts. After processing negation modifiers, there was a marginal improvement in precision [12]. Amalia et al. [13] used rules based on the syntax of negation in Indonesian to establish the scope of negation. SVM classifier with RBF kernel outperforms the other studies when syntactic rules are adapted and "NEG_" tags are used.

In one of the studies, Ljajić and Marovac et al. [14] analyzed how negation affects the mood of tweets written in Serbian. To reverse polarity, the linguistic principles that govern such shifts are considered. When our rules for processing negation were combined with either a lexicon-based strategy or machine learning approaches, a statistically significant relative improvement of up to 31.16% was reached. Gupta and Joshi [15]'s algorithm to improve Twitter categorization notably sought posts with ambiguous negation.

Mao and Li [16] introduce an innovative gating mechanism designed to connect MTL towers. The performance is assessed through aspect-based sentiment analysis and sequential metaphor identification tasks.

An LSTM-based deep neural network model is proposed for negation processing by Singh and Paul [17], with the model itself learning the negation features from the labeled input training dataset. The models were trained and tested using the Conan Doyle narrative corpus, including negation annotations. In particular, this research aims to recommend better content that considers semantics by employing a negation-based dynamic discriminative interest-based preference learning strategy [18]. Lal and Kamath [19] introduce an efficient procedure for managing negations that the author has constructed on a collection of IMDB movie reviews based on first sentiment word (FSW) antonymy in the WordNet. Another study by Mao et al. [20] conducted a comprehensive empirical study on prompt-based sentiment analysis and emotion identification. The goal was to scrutinize the biases of pre-trained language models (PLMs) within the domain of affective computing.

How the Current Investigation Is Better Than the Research Established in the Domain of Mathematical Optimization

Previous studies in the realm of sentiment analysis have predominantly relied on various machine learning techniques, many of which necessitate extensive pre-training [21]. However, there has been a notable gap in research addressing the challenges inherent to sentiment analysis, particularly the complex issue of negation handling. This study introduces innovative unsupervised techniques for sentiment classification, with a primary focus on effectively addressing the challenge posed by negation in textual data. This model can automatically identify and categorize sentiments within textual data. These models are efficient for sentiment tagging for binary class classification but cannot overcome the problem of sentimental challenges like negation handling, irony detection, and sarcasm detection. In the current study, we address the problem of negation handling. The proposed model is MCDM based targets the sentiment analysis of the text. In this study, we use context score and emotion score as parameters for the NEGVOT model to categorize sentiment in text. So the NEGVOT model can be used with any type of written text. The NEGVOT model presents a valuable approach to sentiment analysis in the text without the need for labeled examples. This model can automatically identify and categorize sentiments within textual data. What makes this model particularly advantageous is that it does not rely on pre-existing labeled datasets, offering flexibility and adaptability across diverse domains and languages. By utilizing the NEGVOT model, researchers can explore underlying sentiment patterns and detect anomalies in text. Furthermore, this model proves to be useful in sentiment analysis tasks where labeled data is scarce or not available,

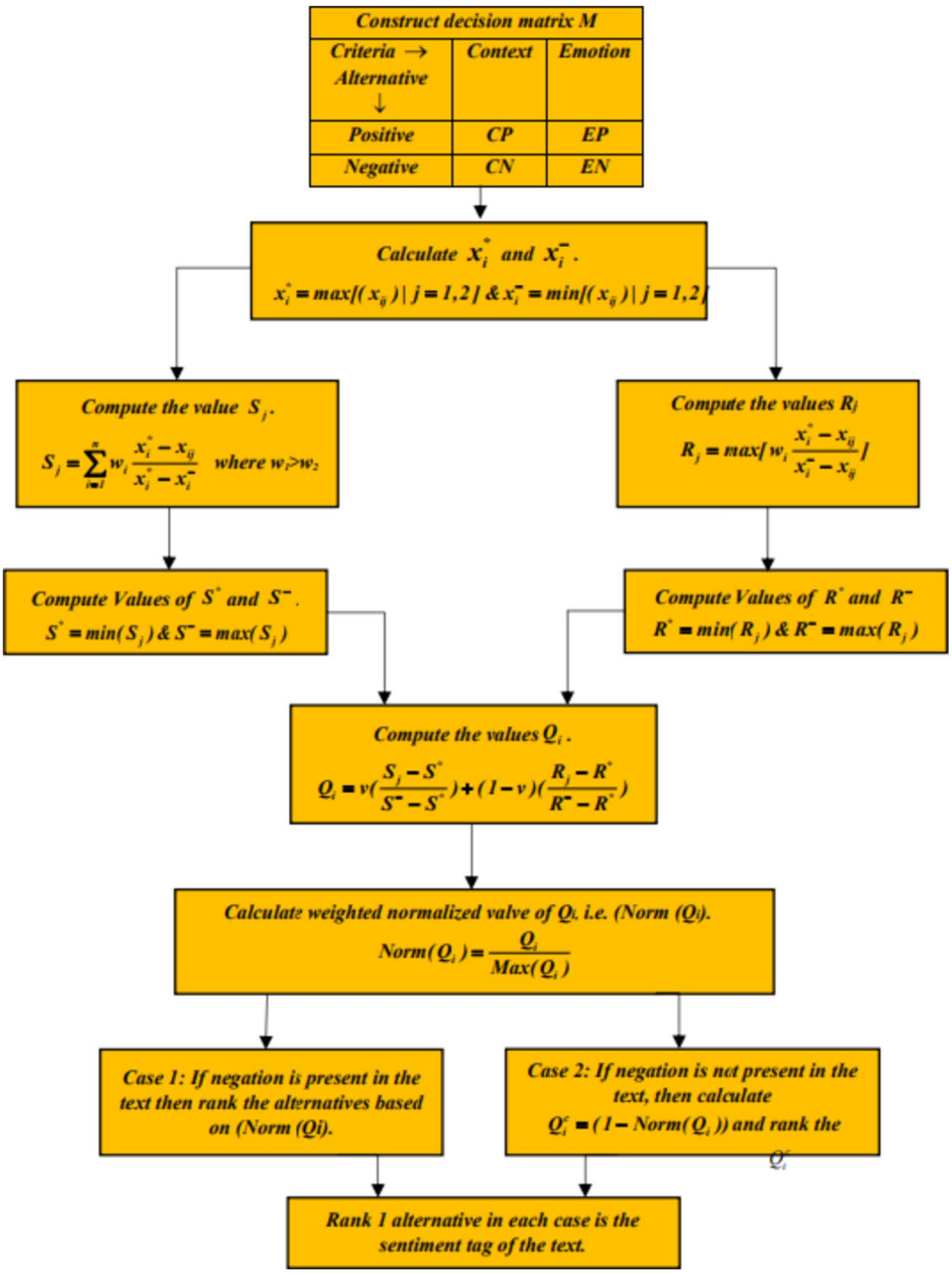


Fig. 1 Diagram of the VIKOR method

thereby providing a cost-effective and efficient solution for comprehending and interpreting sentiments in textual content. However, it is important to mention that none of the aforementioned research efforts has developed an unsupervised sentiment orientation tagger specifically designed for analyzing written text that can resolve the problem of negation handling. The NEGVOT model works in both types of text having negation and text not having negation. The accuracy archived by the NEGVOT model is better than the existing mathematical.

Methodology

We proposed the model named "Negation Handling of the Text Using the VIKOR Optimization Technique" (NEGVOT) model in the methodology. Figure 1 shows the pipeline of the proposed methodology. We calculate the context and emotion scores associated with these parameters below the equations, and then, we supply the scores of these parameters to the NEGVOT model and achieve the desired target. The "Proposed VIKOR Technique" section introduces the

preliminaries of the VIKOR method we used in the NEGVOT model.

Proposed VIKOR Technique

The VIKOR technique is a method of MCDM. The word VIKOR first emerged in 1990 [22] and is derived from the Serbian phrase "Multi-criteria Optimization and Compromise Solution," which is pronounced, "VIKOR." It was first devised by Serafim Opricovic to address choice issues with competing and non-commensurable (different units) criteria, assuming that compromise is allowed for conflict resolution, the decision-maker desires the closest answer to the ideal, and alternatives are evaluated based on all defined criteria. VIKOR assesses alternatives and selects the solution known as a compromise that comes closest to being ideal. The following is how the MCDM problem is stated in Fig. 1. MCDM problems typically involve a finite discrete set of alternatives with each alternative being evaluated according to several criteria. As shown in Fig. 1, alternatives are positive, negative, and neutral, and the criteria are context and emotion as illustrated in Fig. 1. The values corresponding to these positive, negative, and neutral alternatives and across context and emotion criteria are shown in Fig. 1. The first step is to create a decision matrix as shown in Fig. 1. We first collect the alternatives and criteria. Then, we evaluate x_i^* and x_i^-. Then, we calculate S_j,R_j, S^*, S^-, R^*, R^-, *(Norm (Qi)*, and Q_i^c as shown in the flowchart in Fig. 1. After evaluating *(Norm (Qi)* and Q_i^c, we get two cases and an alternative with a rank sentiment tag of the review in either of the cases.

Proposed Model Pipeline

The proposed methodology, as depicted in Fig. 1, comprises the following steps: Initially, we collected data from three different sources. In the subsequent step, the data undergoes a cleaning process involving preprocessing, tokenization, lemmatization, and stopword removal. This ensures that the data is in an appropriate format for further analysis. Once the dataset is cleaned, it is divided into two parts using the Negex Python library. One part contains statements with negations, while the other part consists of sentences without negations. This division allows for a separate analysis of the impact of negations on subsequent steps. Next, the context and emotion scores are calculated for both parts of the dataset. These scores provide valuable insights into the overall sentiment and context of the statements, facilitating a more in-depth analysis. Then, we calculate the context and emotion scores of both parts. Next, we applied the NEGVOT model and evaluated *(Norm (Qi)*. Here, we

Input: A sequence of words (W) in a review R.
Output: Normalised Positive Context Score ($PCS^{\otimes}$), Negative Context Score ($NCS^{\otimes}$), and Neutral Context Score ($OCS^{\otimes}$).
1: Initialise $PCS = 0$, $NCS = 0$ and $OCS = 0$ for each reviews.
2: Retrieve synsets for the word. Synsets=SentiWordNet.lookup(word).
3: Retrieve positive score negative score and neutral score for the synset.

$$PS = synset.\ positiveScore$$
$$NS = synset.\ NegativeScore$$
$$OS = (1 - (PS + NS))$$

4: Update positive context score *(PCS)* and negative context score *(NCS)*.

$$PCS \leftarrow PCS + PS\ ;\ NCS \leftarrow NCS + NS \text{ and } OCS \leftarrow OCS + OS$$

5: Normalized positive context score ($PCS^{\otimes}$) and negative context score ($NCS^{\otimes}$) and neutral context score ($OCS^{\otimes}$).

$$(PCS^{\otimes}) = \frac{PCS}{|W|} \parallel (NCS^{\otimes}) = \frac{NCS}{|W|} \parallel (OCS^{\otimes}) = \frac{OCS}{|W|}$$

Algorithm 1 Evaluate the Context Score of Reviews

get two cases. Case 1 is text containing negation for which we evaluate *(Norm (Qi)*. Case 2 is the text that is negation-free for which we evaluate Q_i^c. After evaluating *(Norm (Qi)* and Q_i^c, we rank the three alternatives; the alternative with rank 1 is the tag to that review. Then, to verify the results, we compare the results of the NEGVOT model with the annotated dataset. This step ensures the model's predictions align with the expected classifications, providing confidence in its performance. Furthermore, the performance of the NEGVOT model is evaluated using the TP, TN, FP, and FN evaluation metrics derived from the confusion matrix. These metrics provide an assessment of the model's accuracy, precision, recall, and overall performance. For a more detailed explanation of the proposed model, please refer to the subsequent steps, and for a visual representation, please consult Fig. 1.

Step 1: Divide the Data into Two Parts In the first step, we collect the dataset and pass the dataset through the Negex[1] Python library. This Python library detects the negation in the text with the help of which we divide the data into two parts. One part of the data contains the text which has negation, and the other part contains the text which does not contain negation.

Step 2: Evaluation of Context Scores This stage involves utilizing SentiWordNet (SWN) to calculate the context score of textual comments [23]. The SWN database stores a list of words that can be either positively or negatively charged. We extract the POS-tagged content words from a user review. We analyze SWN for each word's positive and negative polarity values to determine comment context scores. Thus, the context score values are between 0 and 1. We evaluate the normalized positive context score ($PCS^{\otimes}$), negative context score ($NCS^{\otimes}$), and neutral context score ($OCS^{\otimes}$) following Algorithm 1.

Step 3: Evaluation of Emotion Scores We evaluate emotion scores using the *text2emotion*[2] library. Emotions can

[1] *pip install negspacy.*

[2] https://pypi.org/project/text2emotion/

Input: A sequence of words (W) in a review R.
Output: Normalised Positive Emotion Score ($PES^{\otimes}$), Negative Emotion Score ($NES^{\otimes}$), and Neutral Emotion Score ($OES^{\otimes}$).
1: Load a set of emotion words and associated scores.

$$EmotionWords \leftarrow LoadEmotionWords()$$

2: Initialise *emotion score variables* $H=0$, $A=0$, $S=0$ and $S_p=0$.
3: Tokenize each review into words.
4: For all W in R, *if* $W \in EmotionWords$

$$EmotionScores \leftarrow GetEmotionScores(W)$$

5: Update emotion score variables.

$$For\ each\ emotionScore\ variables$$
$$H \leftarrow H + emotion_score(H)$$
$$A \leftarrow A + emotion_score(A)$$
$$S \leftarrow S + emotion_score(S)$$
$$S_p \leftarrow S_p + emotion_score(S_p)$$

6: Total emotion score E of W.

$$E \leftarrow H + A + S + S_p$$

7: Positive emotion score (EP) and negative emotion score (EN).

$$EP \leftarrow H + S_p \parallel EN \leftarrow A + S + S_p \parallel EO \leftarrow \frac{H + A + S + S_p}{2}$$

8: Normalised Positive Emotion Score ($PES^{\otimes}$), Negative Emotion Score ($NES^{\otimes}$), Neutral Emotion Score.

$$(PES^{\otimes}) = \frac{EP}{E} \parallel (NES^{\otimes}) = \frac{EN}{E} \parallel (OES^{\otimes}) = \frac{EO}{E}$$

Algorithm 2 Evaluate the Emotion Score of Reviews

be categorized as happy, angry, surprised, sad, or fear. We ignore the fear emotion in the current study. Expressions of happiness (H), anger (A), sadness (S), and surprise (S) are the four components of Eq. 5. (Sp). Using these four feelings, we estimated the normalized positive emotion score ($PES^{\otimes}$), negative emotion score ($NES^{\otimes}$), and neutral emotion score ($OES^{\otimes}$) following Algorithm 2.

Step 4: Apply the NEGVOT Model The objective of the NEGVOT model is to give the sentiment tag to the text that contains negation and when it does not contain negations. In this initial step, we construct a decision matrix, as illustrated in Table 1. Table 1 consists of alternatives and criteria, where positive, negative, and neutral are the alternatives, and context and emotion are the criteria we calculated in *step 1*. Next, following Algorithm 3, *Norm* (Q_i) and (Q_i^c) are calculated. The working of the NEGVOT model is shown in Fig. 1. The proposed VIKOR technique is described in detail in Algorithm 2, which uses a 2×2 decision matrix as illustrated in Table 1. Then, the proposed VIKOR algorithms are applied for all text types from step 1 to step 7 (having negations or not having negations). If the text contains a negation, proceed from step 1 to step 7 to rank the alternatives, with the top-ranked alternative serving as the text's final tag. If there is no negation in the text, step 9 will ask you to rank the alternatives. The alternative with rank 1 is the final tag for that text. That is, the ranking of alternatives will be done according to the highest value Q_i in case of text having negations, and in case of the text having no negations, rank the alternatives using Q_i^c.

Table 1 The decision matrix consists of alternatives and criteria

Criteria→ Alternative↓	Context	Emotion
Positive	($PCS^{\otimes}$)	($PES^{\otimes}$)
Negative	($NCS^{\otimes}$)	($NES^{\otimes}$)
Neutral	($OCS^{\otimes}$)	($OES^{\otimes}$)

Table 2 Context and emotion score of the reviews

Alternative	Context	Emotion
Positive	0.5714	0.0055
Negative	0.1429	0
Neutral	0.001	0.02

Input: *Positive, negative, score of context score of text* $((PCS^{\otimes}), (NCS^{\otimes}), (OCS^{\otimes}))$, *emotion* $((PES^{\otimes}), (NES^{\otimes}), (OES^{\otimes}))$.
Output: *The sentiment tag to each text will be the highest rank alternative.*
1: *Construct decision matrix* $M = [M_{ij}]_{2\times2}$ *using the input scores.*
2: *Calculate* x_i^* *and* x_j^-.

$$x_i^* = \max[(x_{ij}) \mid j = 1,2] \,\&\, x_j^- = \min[(x_{ij}) \mid j = 1,2]$$

3: *Compute the value* S_j.

$$S_j = \sum_{i=1}^{n} w_i \frac{x_i^* - x_{ij}}{x_i^* - x_i^-} \quad where \quad w1 > w2 \ \ (\text{i.e } w1 = 0.6, w2 = 0.4)$$

4: *Compute the values* R_j

$$R_j = \max[w_i \frac{x_i^* - x_{ij}}{x_i^- - x_{ij}}]$$

5: *Compute Values of* S^* *and* S^-.

$$S^* = \min(S_j) \,\&\, S^- = \max(S_j)$$

6: *Compute Values of* R^* *and* R^-.

$$R^* = \min(R_j) \,\&\, R^- = \max(R_j)$$

7: *Compute the values* Q_i.

$$Q_i = v(\frac{S_j - S^*}{S^- - S^*}) + (1-v)(\frac{R_j - R^*}{R^- - R^*})$$

8: *Calculated weighted normalized valve of Qi, i.e. (Norm (Qi).*

$$Norm(Q_i) = \frac{Q_i}{Max(Q_i)}$$

9: *Calculate* Q_i^c *if negation is not present in the text.*

$$Q_i^c = (1 - Norm(Q_i))$$

10: *Rank the alternatives based on* Q_i^c.
11: *The alternative with the highest rank is the tag to each text with sentiments.*
12: *If the Text contains negation, rank the alternatives according to step 7.*
13: *If the text content does not contain the negation, proceed to step 9.*

Algorithm 3: Sentiment Tagging Using VIKOR Algorithm

NEGVOT Model Numerical Illustration

In this section, we implemented the NEGVOT model on two illustrative examples. Illustrative example 1 tags the sentiment tag to that text which contains no negation as shown in the "Illustrative Examples of Text Containing No Negation" section. Similarly, illustrative example 2 shows the procedure for sentiment tagging of such text, which contains negations.

Illustrative Examples of Text Containing No Negation

We used the two illustrative scenarios to demonstrate how Algorithms 1, 2, and 3 operate. Illustrative example 1 does not contain negations; thus, we apply the above-mentioned procedures to obtain the desired outcome. After completing steps 2 and 3, we assess the context and emotion scores in Table 2. Table 3 contains the decision matrix's normalized values for the steps depicted in Table 1's decision matrix. The values of the various

Table 3 Normalized decision matrix

Alternative	Context	Emotion
Positive	0.970143	1
Negative	0.242536	0
Neutral	0.0087	0

Table 4 Evaluating Max ($Norm(Q_i)$) value and rank the alternatives

Alternative	S_i	R_i	Q_i	$Norm\ (Q_i)$	Q_i^c	Rank
Positive	0.412	0.398	0.011	0.01088	0.98912	1
Negative	0.651	0.589	1.011	1	0	2
Neutral	0.043	0.0123	0.0087	0.0013	0.012	3

parameters are displayed in Table 4, and after that, we rank the alternatives based on their $Max(Norm(Q_i)$ values. According to Table 4, the final tag for the reviews is the alternative with the highest rank.

Example 1: "I am very satisfied with my Twizzlers purchase. I shared these with others and we have all enjoyed them. I will definitely be ordering more."

Illustrative Example 2 on Text Containing Negations

To demonstrate the execution of Algorithm 2 for texts including negation, we used illustrative example 2, which has reviews containing negation. Following steps 2 and 3, the context and emotion scores provided in Table 5 are evaluated. Table 6 contains the normalized values of Table 5's decision matrix, evaluated from Algorithms 1 and 2. Table 7 displays the values of several parameters, followed by a ranking of the alternatives based on their ($Norm(Q_i)$) value. The alternative with rank 1 is the final tag for the reviews, as indicated in Table 7. Figure 2 shows the deduced tag of reviews R_1 and R_2 following all the procedures as explained previously.

Example 2: "I would never recommend having lunch in that hotel."

Result and Discussion

We implemented the NEGVOT model on three datasets. The results were then compared to conventional approaches. Several metrics were analyzed, and their efficacies were inspected. To demonstrate the model's language independence, we applied it to an English- and Hindi-language dataset and analyzed the encouraging results.

Data Collections

The proposed technique was implemented on three text datasets. The first dataset was the Amazon product reviews crawled from online sources. The dataset is SemEval 2013 Twitter corpus, a widely adopted dataset for text sentiment analysis (TSA) tasks. The corpus was originally released as part of the SemEval 2013 competition, specifically Task 2, which focused on sentiment analysis on Twitter. Task 2 involved two subtasks: term-level sentiment classification and message-level polarity classification. The third dataset is the Stanford Sentiment Treebank movie domain dataset. Table 8 shows the data statistics of these datasets. The specific objective of this work was message-level polarity classification, which involves assigning positive, negative, or neutral labels to individual tweets based on their overall sentiment.

Evaluation of the NEGVOT Performance Concerning Distinct Datasets

In this study, the NEGVOT model's resilience was evaluated across diverse domains within an English-language dataset. The dataset consisted of three distinct sets of English-language datasets. The proposed NEGVOT model was applied to diverse datasets, resulting in the assignment of sentiment tags to reviews. Subsequently, a comparative analysis was conducted between the model's outcomes and the annotated dataset. The evaluation involved the computation of multiple metrics associated with confusion matrices. These metrics, such as accuracy, precision, recall, F1 score, and MCC score, are instrumental in quantifying the performance of the model. By systematically comparing the model's predictions with the annotated dataset, we gained insights into the efficacy of the proposed approach in

Table 5 Context and emotion score of the reviews

Alternative	Context	Emotion
Positive	0.10	0.30
Negative	0.00	0.20
Neutral	0.02	0.02

Table 6 Normalized decision matrix

Alternative	Context	Emotion
Positive	1.00	0.83
Negative	0.00	0.55
Neutral	0.02	0.44

Table 7 Evaluating ($Norm(Q_i)$) value and ranking the alternatives

Alternative	S_i	R_i	$Norm(Q_i)$	Rank
Positive	0.40	0.40	0.50	2
Negative	0.58	0.60	1.00	1
Neutral	0.02	0.03	0.20	3

Table 8 Statistical information about diverse datasets

Dataset	Language	Positive	Negative	Neutral
Amazon product reviews[a]	English	1576	1292	1345
SemEval 2013 Twitter corpus[b]	English	1657	1132	1370
Stanford Sentiment Treebank movie domain dataset[c]	English	1512	1350	1189

[a]https://www.kaggle.com/datasets/saurav9786/amazon-product-reviews

[b]https://paperswithcode.com/dataset/semeval-2013-task-2

[c]https://nlp.stanford.edu/sentiment/code.html

accurately discerning sentiment in reviews. The evaluation metrics employed in this study include accuracy, precision, recall, F1 score, and MCC score, as defined in Eqs. 5–9.

These metrics are essential for assessing the performance of classification models and are widely used to measure and compare the effectiveness of different models in sentiment analysis and various classification tasks. Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total number of instances. Precision evaluates the accuracy of positive predictions by determining the proportion of correctly predicted positive instances among all instances predicted as positive. Recall quantifies the model's ability to capture positive instances by calculating the proportion of correctly predicted positive instances out of all actual positive instances. F1 score combines precision and recall, providing a balanced assessment of the model's performance. Matthews correlation coefficient (MCC) offers a measure of the correlation between predicted and actual labels in binary classification tasks. MCC values range from -1 to $+1$, with $+1$ indicating perfect classification, 0 representing random predictions, and -1 indicating complete disagreement between predictions and actual labels. In Eqs. 5–9, TP, TN, FP, and FN are fundamental metrics used to evaluate classification models. True positives (TPs) represent instances correctly identified as positive, while true negatives (TNs) are instances accurately identified as negative. False positives (FPs) occur when the model incorrectly predicts an instance as positive when it should be negative, and false negatives (FNs) occur when the model incorrectly predicts an instance as negative when it should be positive. TP and TN indicate correct predictions, while FP and FN represent prediction errors. These metrics play a critical role in assessing the model's accuracy, precision, recall, and other performance indicators, providing valuable insights into its ability to accurately classify instances. In the next subsection based on these metrics, we compared the performance of the NEGVOT model over diverse datasets.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1_{score} = \frac{(2 * precision * recall)}{(precision + recall)} \tag{4}$$

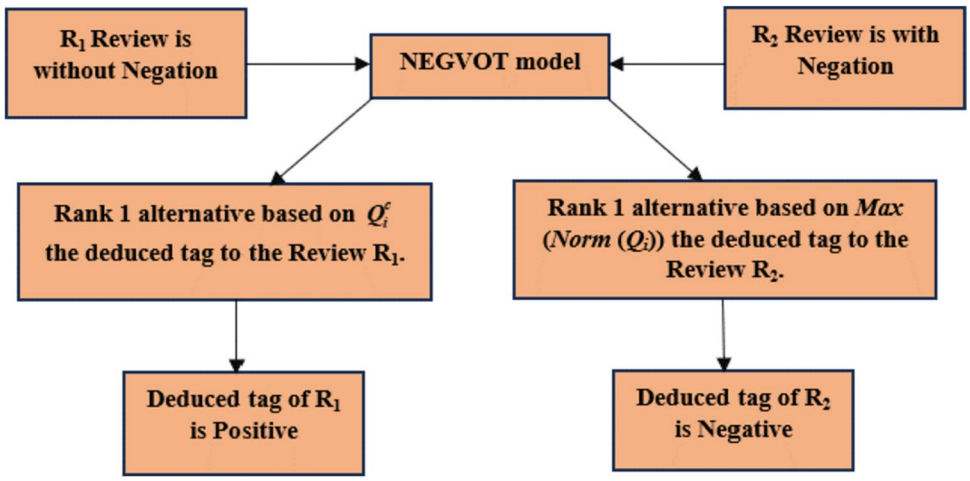


Fig. 2 Deduced tag of reviews using the NEGVOT model

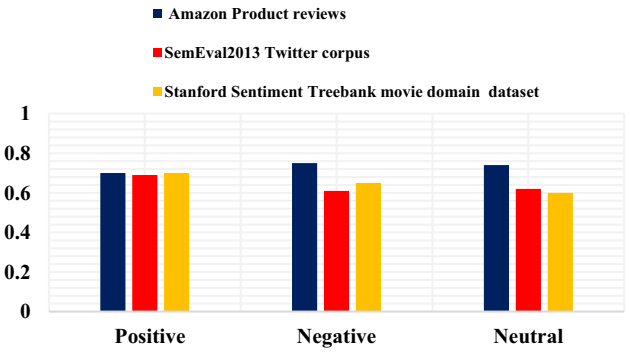


Fig. 3 Comparing the performance of the NEGVOT model for sentiment categorization of tertiary classes across three datasets containing negations

the examined datasets, the Amazon product reviews dataset exhibited the highest frequency of negative feedback provided by customers. Notably, this dataset also contained a substantial proportion of neutral feedback. Conversely, the SemEval 2013 Twitter corpus dataset demonstrated the highest accuracy of 69% in correctly categorizing positive sentiments. However, it struggled comparatively in accurately classifying negative sentiments, achieving a lower accuracy rate of 61%. Regarding the Stanford Sentiment Treebank movie domain dataset, the NEGVOT model displayed commendable performance, attaining the highest accuracy of 70% in correctly identifying positive sentiment instances. It also achieved a respectable accuracy of 65% in effectively categorizing negative sentiment instances. In summary, the

$$Matthews\ correlation\ coefficient(MCC) = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TN + FN) * (FP + TN) * (TP + FN)}} \quad (5)$$

Performance of NEGVOT Model of Tertiary Classification over Three Datasets

In this section, we present a comprehensive evaluation of the NEGVOT model's performance on three distinct review datasets used for sentiment categorization. Our primary aim was to rigorously assess the model's ability to accurately classify sentiments into positive, negative, and neutral categories. The obtained results of the NEGVOT model for three class classifications over three datasets are visually depicted in Fig. 3. The primary objective of the NEGVOT model is to proficiently ascertain the sentiment orientation of written feedback, even in cases where the text contains negations. Through its robust negation handling mechanism, the NEGVOT model demonstrated remarkable performance, resulting in an overall accuracy of approximately 81%. Among NEGVOT model exhibited substantial efficacy across the evaluated datasets, effectively assigning accurate sentiment labels, even in the presence of negations. Its capacity to tackle the complexities associated with negated expressions underscores its practical utility in sentiment analysis tasks.

Performance Comparison of the NEGVOT Model on the Reviews with Negation

In this section, we conduct a comparative analysis of the NEGVOT model against three other models designed for handling negations in text: SentiNeg [24] model, the model by Gupta and Joshi [15], and MiMuSA [25]. We implemented the NEGVOT model on three different datasets, as outlined in Table 8, and present the results in Figs. 4, 5, and 6.

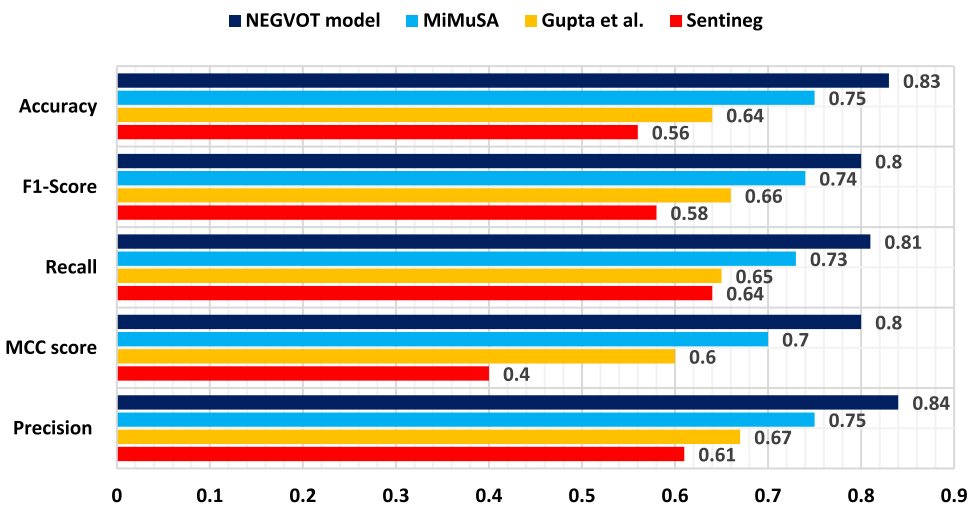


Fig. 4 Comparing the efficacy of the NEGVOT model over other algorithms for Amazon product reviews

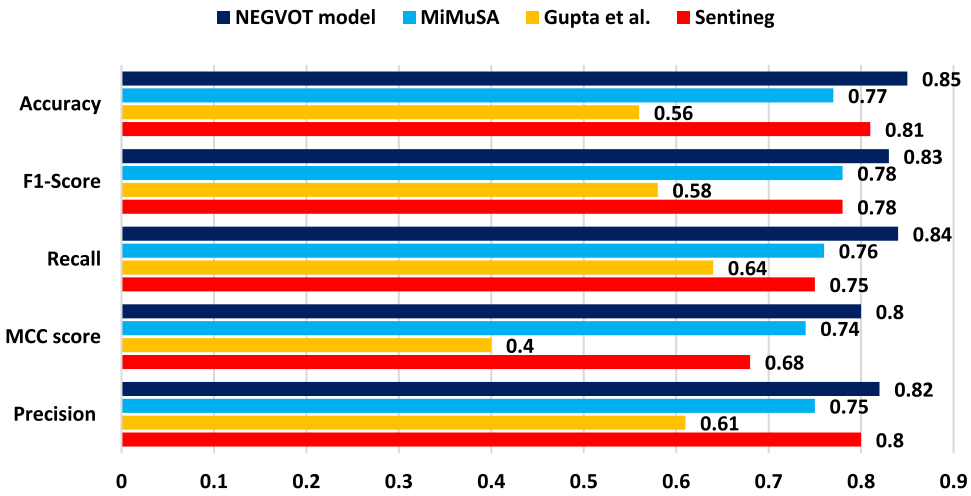


Fig. 5 Comparing the efficacy of the NEGVOT model over other algorithms for the SemEval 2013 Twitter corpus

Gupta et al. employed a feature-based approach for sentiment analysis on Twitter, with an enhanced method to handle negations. On the Amazon product review dataset, their model achieved an accuracy that is 0.19 lower than that of the NEGVOT model. Similarly, on the SemEval 2013 and Stanford review datasets, Gupta et al.'s model achieved accuracies of 0.56 and 0.48, respectively, which are both lower than the NEGVOT model's performance. The overall performance of Gupta et al.'s model is consistently inferior to that of the NEGVOT model, as depicted in Figs. 4, 5, and 6. This indicates that the NEGVOT model is more efficient in handling negations in text. Another model, SentiNeg, was designed to handle negations at the sentence level. However, its performance, as measured by five evaluation metrics, falls short of the NEGVOT model across all three datasets, as shown in Figs. 4, 5, and 6. The highest accuracy achieved by SentiNeg is 0.81 on the SemEval 2013 dataset, which is 0.04 lower than that of the NEGVOT model. Similarly, SentiNeg's highest MCC score on SemEval 2013 is 0.68, which is 0.13 lower than the NEGVOT model's score. These results suggest that the NEGVOT model is more reliable than SentiNeg in handling negations in text. Next in the queue, is MiMuSA, which was introduced to address the negation handling task in text. MiMuSA aims to emulate human language comprehension processes through a modular structure with multiple levels. On the Stanford reviews dataset, MiMuSA achieved the highest accuracy of approximately 0.79, which is 0.03 lower than the NEGVOT model's performance, as illustrated in Fig. 6. On the SemEval 2013 dataset, MiMuSA achieved a precision of 0.75, which is 0.07 lower than the NEGVOT model's precision. Consequently, all evaluation measures for the MiMuSA model are lower than those of the NEGVOT model, as shown in Figs. 4, 5, and 6.

Performance Comparison of the NEGVOT Model on the Reviews Without Negation

In the current study, we employed the NEGVOT model and compared it with other models. Another five models such as GRA (grey relational analysis) [26], SOTM (SAW-MCDM and game model–based sentiment analysis) [27], TOGT-ST

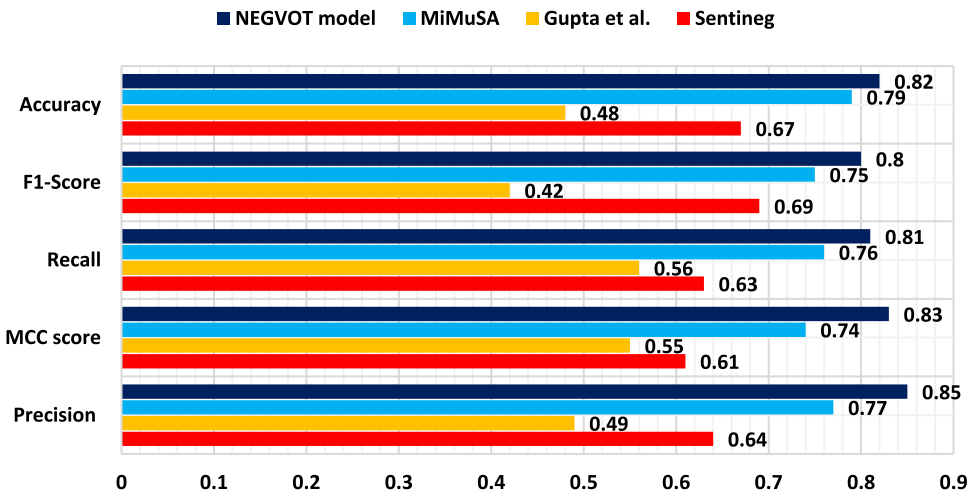


Fig. 6 Comparing the efficacy of the NEGVOT model over other algorithms for the Stanford Sentiment Treebank movie domain

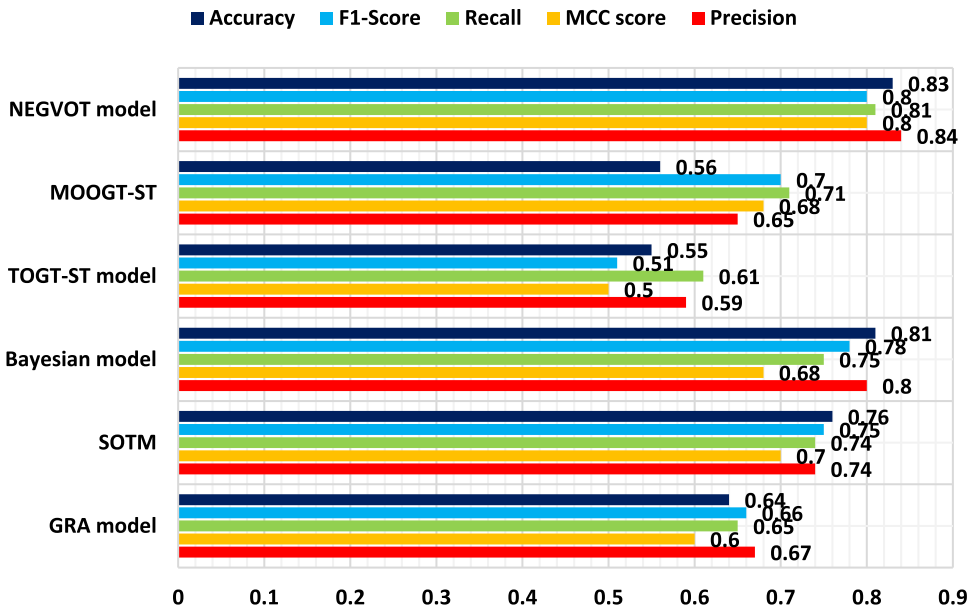


Fig. 7 Comparing the efficacy of the NEGVOT model over other algorithms for Amazon product reviews

model [28], MOOGT-ST [29] model, and Bayesian game [30] mathematical optimization technique–based model. We implemented the NEGVOT model on three datasets as explained in Table 8, and results are illustrated in Figs. 4, 5, and 6.

All these models are unsupervised in nature. The MCDM-based GRA approach was used to assign sentiment labels to textual data. As demonstrated by the findings, the GRA technique obtained an accuracy of approximately 0.64 for Amazon reviews, roughly 0.74 for the SemEval 2013 Twitter corpus, and roughly 0.56 in the Stanford Sentiment Treebank's movie domain. The performance of the NEGVOT model is 0.2 times better than the GRA model as illustrated in Figs. 7, 8, and 9. In contrast, the SOTM model uses a game model in conjunction with the SAW-MCDM technique to classify attitudes in reviews. Applying the SOTM model to the Amazon marketplace dataset demonstrated noteworthy F1 scores of 0.75, recalls of 0.74, MCC scores of 0.7, and precisions of 0.74. It obtained an F1 score of 0.66 on the SemEval 2013 Twitter corpus and an average accuracy of 0.769 in the Stanford Sentiment Treebank's movie domain. Bayesian game model is focused on unsupervised sentiment analysis of product reviews. While obtaining a marginally lower F1 score of 0.51 on the SemEval 2013 Twitter corpus, it demonstrated impressive performance with an F1 score of 0.78 on the Amazon review dataset. Furthermore, the Bayesian game model showed an accuracy of roughly

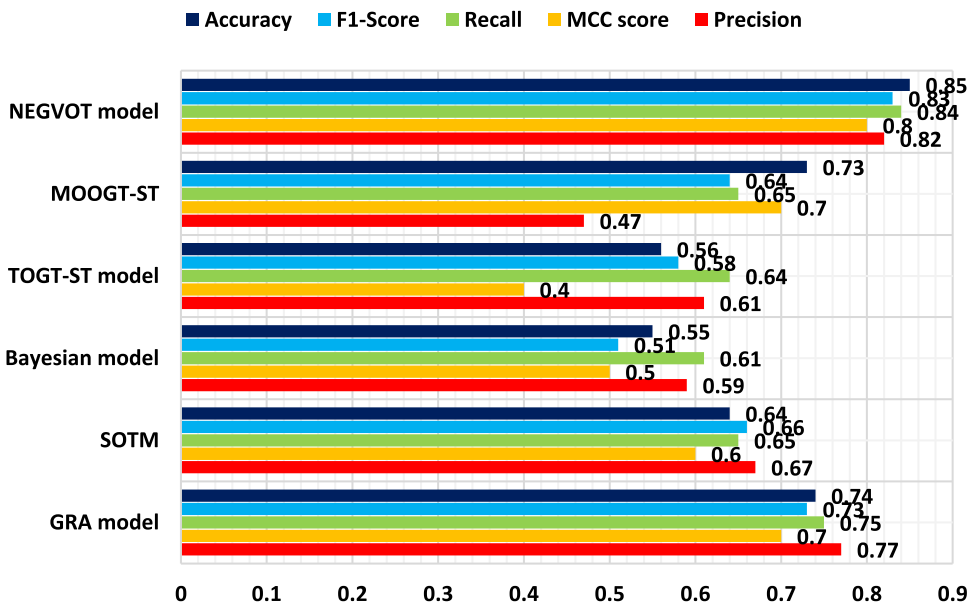


Fig. 8 Comparing the efficacy of the NEGVOT model over other algorithms for the SemEval 2013 Twitter corpus

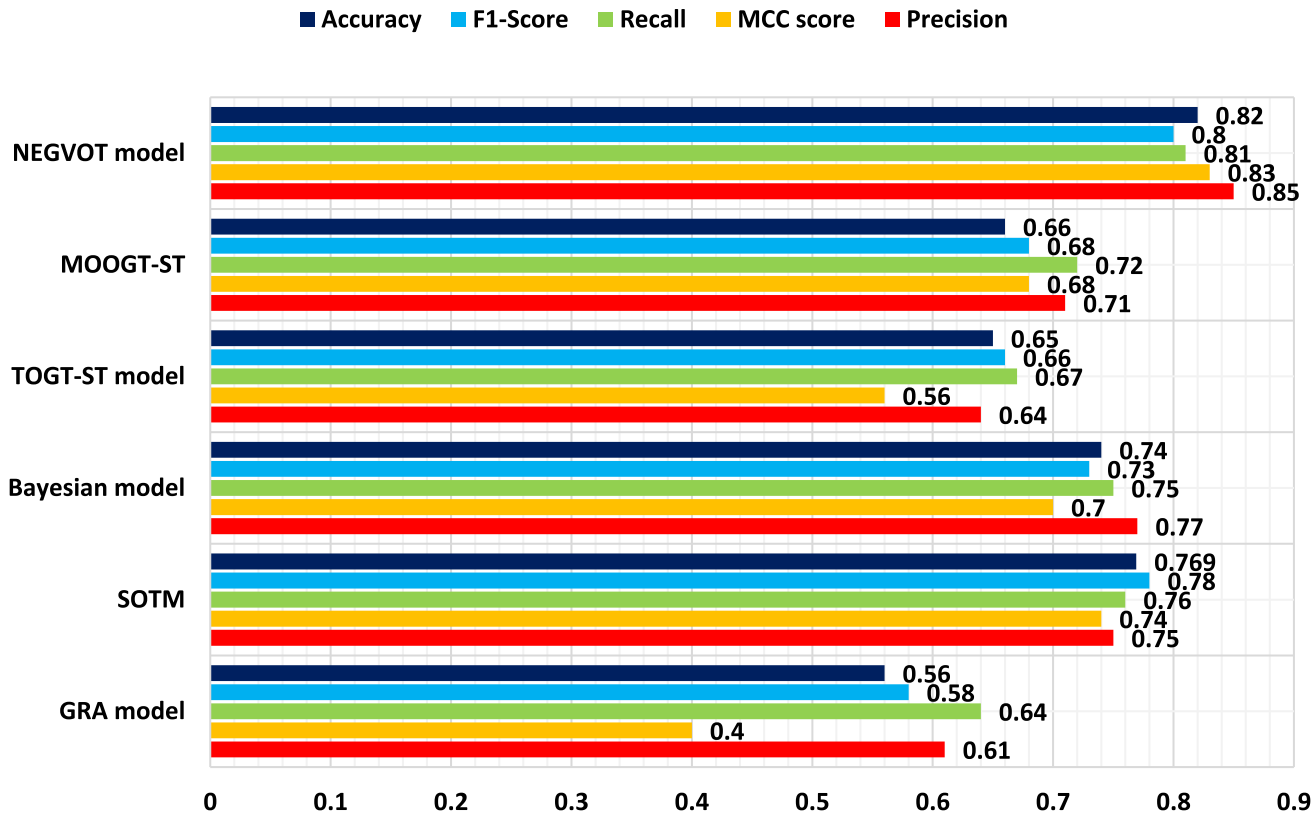


Fig.9 Comparing the efficacy of the NEGVOT model over other algorithms for the Stanford Sentiment Treebank movie domain

0.74 and an F1 score of 0.73 when applied to the movie domain of the Stanford Sentiment Treebank. Similarly, the MOOGT-ST model was tested on all three datasets. It was developed using the MCDM and game theory frameworks. Nonetheless, it continuously showed worse accuracy when compared to the NEGVOT model. Sentiment tagging is the main emphasis of the TOGT-ST model, which integrates the TOPSIS technique with a non-cooperative game model. When applied to the SemEval 2013 Twitter corpus, this model's accuracy increased significantly to around 0.56 from its approximate 0.55 on the Amazon review dataset. The TOGT-ST model performed moderately in the Stanford Sentiment Treebank movie domain dataset, with accuracy, F1 score, recall, MCC score, and precision all circling 0.65.

Notably, the NEGVOT model continuously showed exceptional performance across many datasets, outperforming all other models that were evaluated. It received an astounding 0.83 accuracy, 0.80 F1 score, 0.81 recall, 0.8 MCC score, and 0.84 precision on Amazon product reviews. These findings highlight the remarkable sentiment classification accuracy of the NEGVOT algorithm, especially when applied to Amazon reviews. Similarly, the NEGVOT model demonstrated impressive results with an accuracy of 0.85 on the SemEval 2013 Twitter corpus, demonstrating its efficacy in sentiment categorization for tweets. With an accuracy of 0.82, the NEGVOT model demonstrated its exceptional performance even in the difficult Stanford Sentiment Treebank movie domain dataset, indicating its usefulness and dependability in handling a variety of sentiment classification tasks. Hence, the performance of the NEGVOT model is superior as compared to the existing model as illustrated in Figs. 7, 8, and 9.

Statistical Validation

Model validation ensures that the study's objectives will be accomplished by comparing the model's output to the output of the data-generation process. Utilizing statistical significance, researchers validate their conclusions by testing the validity of their models. Two z-tests were conducted to validate the performance of the suggested model. Using the dataset of meal reviews, we picked two samples with varying sizes. We collected 497 reviews for the first sample (N_1). There were 500 reviews in total, and 471 (X_1) were assigned the correct tags. As a sample, we obtain a $p_1 = 0.956$ for a fraction of reviews with the correct tags. We surveyed 700 users for our second sample (N_2). For 678 (X_2) reviews, the proportion of accurate tags is 0.9685 (p_2). We developed the following set of alternate and null hypotheses to determine if our model is accurate, as shown in Fig. 10.

Fig. 10 A graphical illustration of the key region of the theory stated above

H_o : Both datasets have the same proportion of correctly labeled reviews, $p_1 = p_2$

H_a : Both datasets have a different proportion of correctly labeled reviews, $p_1 \neq p_2$

As $|z| = 1.1359 \leq z_c = 1.96$, it may be inferred that the null hypothesis is not rejected. The null hypothesis is not rejected because the p value is less than 0.05, which is $p = 0.256$. The null hypothesis H_o is not rejected; hence, the conclusion is drawn. At the $\alpha = 0.05$ significance level, there is insufficient evidence to conclude that the population proportions p_1 and p_2 differ. The 95% confidence interval for $p_1 - p_2$ is $-0.0346 < p_1 - p_2 < 0.0096$.

Discussion

In this section, we examine a few aspects of this study, including macro- and micro-evaluation across three datasets, time and spatial complexity, error analysis, and the significance and limitations of the NEGVOT model.

Macro- and micro-evaluation

When dealing with diverse datasets, we examined overall performance with macro- and micro-averages. Figure 9 demonstrates the general performance on three different datasets. We present the average accuracy, F-score, and recall across all n datasets, both large and small, from Eqs. 6 to 11. The macro-F1 score, recall, and precision are 0.788, 0.75, and 0.68. Similarly the micro F1 score, recall, and Precision are 0.77, 0.69, and 0.74 over three datasets as illustrated in Fig. 11.

$$Macro - averaged\ Precision = \frac{\sum_{i=1}^{n} Precision_i}{n} \tag{6}$$

$$Macro - averaged\ Recall = \frac{\sum_{i=1}^{n} Recall}{n} \tag{7}$$

$$Macro - F1\ score = 2.\frac{Macro - Precision_i \times Macro - Recall}{Macro - Precision_i + Macro - Recall} \tag{8}$$

$$Micro - averaged\ precision = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \tag{9}$$

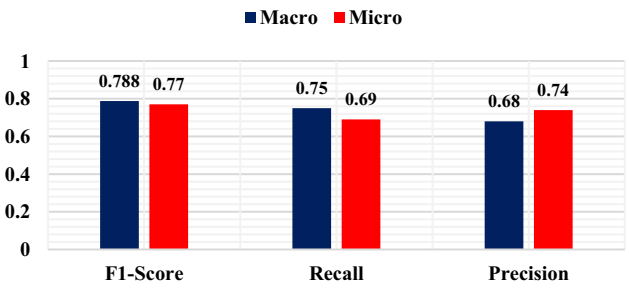


Fig. 11 Measures of macro- and micro-performance based on a sampling of three datasets

Table 9 Time and space complexity of the VIKOR method

Efficiency	Best case	Average case	Worst case
Time Complexity	Ω(n)	Ω(2n)	Ω(2n)
Space Complexity	Ω(n)	Ω(n)	Ω(n)

$$Micro-averaged\ recall = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)} \tag{10}$$

$$Micro-averaged\ F1\ score = 2 \cdot \frac{Micro-Precision_i \times Micro-Recall}{Micro-Precision_i + Micro-Recall} \tag{11}$$

Computational Complexity

The number of iterations an algorithm needs to finish a task for a given input length measures its temporal complexity, which may be used to categorize algorithms into broad categories. This analysis focuses on time complexity, a subset of computational complexity that quantifies the work required to implement an algorithm. The time complexity of an algorithm is the sum of the times it takes to run all of its statements. As a result, its success highly depends on the amount of data being handled. It helps quantify and analyze the algorithm's efficacy. Out of all the MCDM methods, the VIKOR method has the lowest complexity [31]. The complexity of an algorithm is proportional to the amount of input space it requires. The input and additional spaces are combined to form the total space complexity. The amount of memory needed to store the algorithm's inputs and outputs before it can be run and returned. Table 9 shows that the time and space complexity of the VIKOR method is polynomial time.

Significance

Mathematical optimization can be considered a powerful tool for sentiment analysis, as it can help identify the most significant words and phrases in a piece of text and assign weights to them based on their importance for sentiment analysis. This allows for more accurate and efficient sentiment analysis and can help to identify patterns and trends in large datasets. One approach to sentiment analysis using mathematical optimization is to use a technique called linear programming. This involves defining an objective function, a mathematical expression that represents the sentiment of a piece of text, and then optimizing this function subject to constraints. The constraints can include factors such as the length of the text, the presence of certain words or phrases, and the overall sentiment score. Overall, sentiment analysis using mathematical optimization is a powerful tool for understanding the emotional tone of a piece of text and can be applied to a wide range of applications, including social media monitoring, market research, and customer feedback analysis. Combining mathematical optimization techniques with natural language processing allows highly accurate and efficient sentiment analysis that can help businesses and organizations make more informed decisions and improve their products and services. The offered study lays the path for using mathematical optimizations in various NLP tasks.

Challenges

(i) The NEGVOT model consists of VIKOR MCDM techniques. VIKOR being an MCDM technique consists of criteria and alternatives. Here, NEGVOT model consists of two criteria that are context and emotion. The weight for context criteria is $w_1 = 0.6$, and for emotion, $w_2 = 0.4$. The essential condition for the accurate working of the NEGVOT model is that $w_1 > w_2$. The proposed NEGVOT model fails when $w_1 = w_2 = 0.5$ and $w_1 > w_2$.

(ii) The second drawback of the NEGVOT model is that it fails to give the correct tag in some cases, as shown in Table 10. Table 10 contains some examples where the NEGVOT model fails to identify the correct sentiment tag of the text. In Ex 1, one part of the text is positive and the other part is negative, but the overall sentence is positive but the NEGVOT model gives a sentiment tag negative. Ex 2 is the sarcasm and emojis; NEGVOT model does not identify such sentences. Ex 3 contains the negation + bilingual words in the sentences, which is overall positive, but the sentiment tag is negative. Ex 4 contains hashtags in the text, which NEGVOT models did not identify.

Table 10 Examples where the NEGVOT model failed

Examples	Text	Actual	Sentiment tag from NEGVOT
Ex 1	The smartphone is very fast and it does not lag anywhere between the high-end games	Positive	Negative
Ex 2	After using this oil my hair became like Dwayne Johnson 😭😒	Negative	Positive
Ex 3	I have never seen such a behetareen movie in my entire life	Positive	Negative
Ex 4	#i_love_my_country	Positive	NULL

Conclusion

The article proposes a methodology for classifying the sentiments of English-language texts, both with and without negations. The model was evaluated using three different datasets. To use words, the NEGVOT model makes use of the SWN. The sentiment scores of textual feedback are calculated using SWN, and the emotion scores are extracted using the Python Negex tools. The VIKOR MCDM approach was employed in the NEGVOT model. The NEGVOT model creates a sentiment label for each review by employing the MCDM approach. An observed 91% accuracy suggests that the method is reliable. Lexicon-based and mathematical optimization technique-based frameworks compare the results achieved by the suggested methodology. Further improvements to the predicted performance of the proposed model can be achieved by fixing the problems highlighted in the discussion section. Furthermore, we will use a WordNet-based strategy to increase the precision of the suggested method for extending sentiment lexicons. The suggested method has demonstrated encouraging results, and it may lead the way for other new applications including the analysis of sentiment in product reviews and the use of social media for effective policy, government transparency, public input into decision-making, gender equality, and the suppression of unrest and crime. In future research, we incorporate the mathematical optimization technique for generating multimodal that work on a multilingual dataset that helps in different levels of sentiment analysis.

Appendix

The results of Figs. 4, 5, 6, 7, 8, and 9 are numbered in Table 11 and Table 12.

Table 11 Comparison on the text having negation

	Approach	Accuracy	F1-score	Recall	MCC score	Precision
Amazon Reviews	**Sentineg**	0.56	0.58	0.64	0.4	0.61
	MiMUSA	0.75	0.74	0.73	0.7	0.75
	Gupta et al.	0.64	0.66	0.65	0.6	0.67
	NEGVOT model	**0.83**	**0.8**	**0.81**	**0.8**	**0.84**
SemEval2023	**Sentineg**	0.81	0.78	0.75	0.68	0.8
	MiMUSA	0.77	0.78	0.76	0.74	0.75
	Gupta et al.	0.56	0.58	0.64	0.4	0.61
	NEGVOT model	**0.85**	**0.83**	**0.84**	**0.8**	**0.82**
Stanford sentiment Treebank Movie review	**Sentineg**	0.67	0.69	0.63	0.61	0.64
	MiMUSA	0.79	0.75	0.76	0.74	0.77
	Gupta et al.	0.48	0.42	0.56	0.55	0.49
	NEGVOT model	**0.82**	**0.8**	**0.81**	**0.83**	**0.85**

Table 12 Comparison on the text without negation

	Approaches	Accuracy	F1-score	Recall	MCC score	Precision
Amazon reviews	**MOOGT-ST model**	0.56	0.7	0.71	0.68	0.65
	TOGT-ST model	0.55	0.51	0.61	0.5	0.59
	Bayesian model	0.81	0.78	0.75	0.68	0.8
	SOTM model	0.76	0.75	0.74	0.7	0.74
	GRA model	0.64	0.66	0.65	0.6	0.67
	NEGVOT model	**0.83**	**0.8**	**0.81**	**0.8**	**0.84**
SemEval2023	**MOOGT-ST model**	0.73	0.64	0.65	0.7	0.47
	TOGT-ST model	0.56	0.58	0.64	0.4	0.61
	Bayesian model	0.55	0.51	0.61	0.5	0.59
	SOTM model	0.64	0.66	0.65	0.6	0.67
	GRA model	0.74	0.73	0.75	0.7	0.77
	NEGVOT model	**0.85**	**0.83**	**0.84**	**0.8**	**0.82**
Stanford sentiment Tree movie	**MOOGT-ST model**	0.66	0.68	0.72	0.68	0.71
	TOGT-ST model	0.65	0.66	0.67	0.56	0.64 T
	Bayesian model	0.74	0.73	0.75	0.7	0.77
	SOTM model	0.769	0.78	0.76	0.74	0.75
	GRA model	0.56	0.58	0.64	0.4	0.61
	NEGVOT model	**0.82**	**0.8**	**0.81**	**0.83**	**0.85**

Author Contribution All authors contributed equally to this study.

Data Availability https://www.kaggle.com/datasets/saurav9786/amazon-product-reviews, https://paperswithcode.com/dataset/semeval-2013-task-2, https://nlp.stanford.edu/sentiment/code.html.

Code Availability The code generated during the current study are available from the corresponding author on reasonable request.

Declarations

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors declare no competing interests.

Human and Animal Ethics No humans or animals were harmed in any way.

References

1. Pandey S, Sagnika S, Mishra BSP. A technique to handle negation in sentiment analysis on movie reviews. Proc 2018 IEEE Int Conf Commun Signal Process ICCSP. 2018;737–43. https://doi.org/10.1109/ICCSP.2018.8524421.
2. Hupont I, Cerezo E, Ballano S, Baldassarri S. On the origin of the methodology for the scalable fusion of affective channels in a continuous emotional space and the 'emotional kinematics' filtering technique - a correction. Inf Fusion. 2021;67:1–2.
3. Valle-Cruz D, Fernandez-Cortez V, López-Chau A, Sandoval-Almazán R. Does Twitter affect stock market decisions? Financial sentiment analysis during pandemics: a comparative study of the H1N1 and the COVID-19 periods. Cogn Comput. 2022;14(1):372–87. https://doi.org/10.1007/S12559-021-09819-8/TABLES/17.
4. Popescu AM, Etzioni O. Extracting product features and opinions from reviews. Nat Lang Process Text Min. 2007;9–28. https://doi.org/10.1007/978-1-84628-754-1_2/COVER.
5. Na JC, Khoo C, Wu PHJ. Use of negation phrases in automatic sentiment classification of product reviews. Libr Collect Acquis Tech Serv. 2005;29(2):180–91. https://doi.org/10.1016/J.LCATS.2005.04.007.
6. Kumar V. Spatiotemporal sentiment variation analysis of geotagged COVID-19 tweets from India using a hybrid deep learning model. Sci Rep. 2022;12(1). https://doi.org/10.1038/s41598-022-05974-6.
7. Sigari S, Gandomi AH. Analyzing the past, improving the future: a multiscale opinion tracking model for optimizing business performance. Human Soc Sci Commun. 2022;9(1). https://doi.org/10.1057/s41599-022-01325-y.
8. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Comput Linguist. 2011;37(2):267–307. https://doi.org/10.1162/COLI_A_00049.
9. Limsopatham N, Macdonald C, McCreadie R, Ounis I. Exploiting term dependence while handling negation in medical search. SIGIR'12 - Proc Int ACM SIGIR Conf Res Dev Inf Retr. 2012;1065–66. https://doi.org/10.1145/2348283.2348471.
10. Xia R, Xu F, Zong C, Li Q, Qi Y, Li T. Dual sentiment analysis: considering two sides of one review. IEEE Trans Knowl Data Eng. 2015;27(8):2120–33. https://doi.org/10.1109/TKDE.2015.2407371.
11. Xia R, Xu F, Yu J, Qi Y, Cambria E. Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. Inf Process Manage. 2016;52(1):36–45. https://doi.org/10.1016/J.IPM.2015.04.003.
12. Ghag KV, Shah K. Negation handling for sentiment classification. Proc 2nd Int Conf Comput Commun Control Autom ICCUBEA. 2016. https://doi.org/10.1109/ICCUBEA.2016.7860016.
13. Amalia R, Bijaksana MA, Darmantoro D. Negation handling in sentiment classification using rule-based adapted from Indonesian language syntactic for Indonesian text in Twitter. J Phys Conf Seri. 2018;971(1). https://doi.org/10.1088/1742-6596/971/1/012039.
14. Ljajić A, Marovac U. Improving sentiment analysis for twitter data by handling negation rules in the Serbian language. Comput Sci Inf Syst. 2019;16(1):289–311. https://doi.org/10.2298/CSIS180122013L.
15. Gupta I, Joshi N. Feature-Based Twitter Sentiment Analysis with Improved Negation Handling. IEEE Trans Comput Soc Syst. 2021;8(4):917–27. https://doi.org/10.1109/TCSS.2021.3069413.
16. Mao R, Li X. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. 35th AAAI Conf Artif Intell AAAI. 2021;15:13534–542. https://doi.org/10.1609/aaai.v35i15.17596.
17. Singh PK, Paul S. Deep learning approach for negation handling in sentiment analysis. IEEE Access. 2021;9:102579–92. https://doi.org/10.1109/ACCESS.2021.3095412.
18. Albatayneh NA, Ghauth KI, Chua FF. Discriminate2Rec: negation-based dynamic discriminative interest-based preference learning for semantics-aware content-based recommendation. Expert Syst Appl. 2022;199. https://doi.org/10.1016/j.eswa.2022.116988.
19. Lal U, Kamath P. Effective negation handling approach for sentiment classification using synsets in the WordNet lexical database. 2022 1st Int Conf Electr Electron Inf Commun Technol ICEEICT. 2022. https://doi.org/10.1109/ICEEICT53079.2022.9768641.
20. Mao R, Liu Q, He K, Li W, Cambria E. The Biases of pre-trained language models: an empirical study on prompt-based sentiment analysis and emotion detection. IEEE Trans Affect Comput. 2022. https://doi.org/10.1109/TAFFC.2022.3204972.
21. Zhang H, et al. Leveraging statistical information in fine-grained financial sentiment analysis. World Wide Web. 2022;25(2):513–31. https://doi.org/10.1007/S11280-021-00993-1/TABLES/5.
22. Ghadikolaei AS, Esbouei SK, Antuchevičienė J. Applying fuzzy MCDM for financial performance evaluation of Iranian companies. Technol Econ Dev Econ. 2014;20(2):274–91. https://doi.org/10.3846/20294913.2014.913274.
23. Esuli A, Sebastiani F. SENTIWORDNET: A publicly available lexical resource for opinion mining. Proc 5th Int Conf Lang Resour Eval LREC. 2006;417–22 [Online]. Available: http://nmis.isti.cnr.it/sebastiani/Publications/2007TR02.pdf. Accessed Sep 23, 2021.
24. Savanur SR, Sumathi R. SentiNeg: algorithm to process negations at sentence level in sentiment analysis. Int J Softw Innov. 2023;11(1). https://doi.org/10.4018/IJSI.315741.
25. Wang Z, Hu Z, Ho SB, Cambria E, Tan AH. MiMuSA—mimicking human language understanding for fine-grained multi-class sentiment analysis. Neural Comput Appl. 2023. https://doi.org/10.1007/s00521-023-08576-z.
26. Punetha N, Jain G. Sentiment Analysis of Stock Prices and News Headlines Using the MCDM Framework. 2023;1–4. https://doi.org/10.1109/aist55798.2022.10065221.
27. Punetha N, Jain G. Aspect and orientation-based sentiment analysis of customer feedback using mathematical optimization models. Knowl Inf Syst. 2023;1–30. https://doi.org/10.1007/s10115-023-01848-z.
28. Punetha N, Jain G. Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews. Appl Intell. 2023;1–22. https://doi.org/10.1007/s10489-023-04471-1.
29. Punetha N, Jain G. Unsupervised sentiment analysis of Hindi reviews using MCDM and game model optimization

techniques. Sādhanā. 2023;48(4):1–13. https://doi.org/10.1007/S12046-023-02255-4.
30. Punetha N, Jain G. Bayesian game model based unsupervised sentiment analysis of product reviews. Expert Syst Appl. 2023;214:119128. https://doi.org/10.1016/j.eswa.2022.119128.
31. Ghaleb AM, Kaid H, Alsamhan A, Mian SH, Hidri L. Assessment and comparison of various MCDM approaches in the selection of manufacturing process. Adv Mater Sci Eng. 2020;2020. https://doi.org/10.1155/2020/4039253.

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/374861277

Organic Farming in India and its Way Forward

Article *in* Defence Life Science Journal · May 2023

DOI: 10.14429/dlsj.8.18975

CITATIONS
0

READS
181

4 authors, including:

Monika Sharma
Delhi Technological University
2 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Rajeev Kumar Mishra
Delhi Technological University
57 PUBLICATIONS **512** CITATIONS
SEE PROFILE

All content following this page was uploaded by Monika Sharma on 03 November 2023.

The user has requested enhancement of the downloaded file.

Defence Life Science Journal, Vol. 8, No. 3, July 2023, pp. 239-247, DOI : 10.14429/dlsj.8.18975
© 2023, DESIDOC

Organic Farming in India and its Way Forward

N. Krithika, Rishabh Jain, Monika Sharma, and Rajeev Kumar Mishra*

Department of Environmental Engineering, Delhi Technological University, Delhi–110 042, India
**Email: rajeevkumarmishra@dtu.ac.in*

ABSTRACT

India is home to 30% of the world's organic growers and is likely to expand in the coming years. The solution to the issues of sustainability, global warming, land degradation and food security is Organic Farming, which is seen as a sign of dynamic change for the agricultural industry. Organic Farming discards the use of synthetic fertilizers and promotes sustainable agricultural practices. Organic Farming holds immense potential to revive the degrading state of the agricultural sector in the world by offering environmental benefits, quality products and conserving non-renewable resources. It is a promising alternative to conventional farming and is expanding quickly. Organic Farming is gaining worldwide attention with 2.30 million hectares of land being used for the purpose. It helps to reduce greenhouse gas (GHG) emissions and improves soil fertility, boosting productivity and crop health. Organic Farming can also be used for land reclamation purposes. The aim of the present study is to examine the development of Organic Farming in India and globally, as well as identify any potential barriers to its implementation.

Keywords: Agriculture; Sustainability; Organic farming; Pesticides; Land degradation

1. INTRODUCTION

India derives 20 percent of its GDP from agriculture, with 66.6 percent of India's population being dependent on it for their livelihood. Over the last 30 years, cropping intensity has increased from 118 percent to 135 percent[1]. Due to the expansion of agriculture, arable land is rapidly decreasing. The most significant barrier is the excessive consumption of chemical fertilizers, causing soil quality to degrade over time and thus be harmful in the long term. Another important aspect is land degradation due to anthropogenic activities such as mining which disrupts the soil and nearby land. Reclamation of such areas is highly recommended, especially in India, a major player in the mining sector.

To remedy this alarming situation, organic fertilizers are highly recommended. However, the availability and adoption of Organic Farming in India are low. These nonchemical options are critical for the agricultural transition to a sustainable farming practice but are expensive due to their low cost compared to chemical fertilizers[2]. Organic Farming is an eco-friendly technique that originated due to the adverse effects of chemical fertilizers. Organic Farming holds immense potential to revive the degrading state of the agricultural sector in the world by offering environmental benefits, quality products and conserving non-renewable resources. It not only helps to sustain the soil and ecosystem but also improves the health of producers and consumers[3].

Received : 20 March 2023, Revised : 18 April 2023
Accepted : 21 April 2023, Online published : 12 October 2023

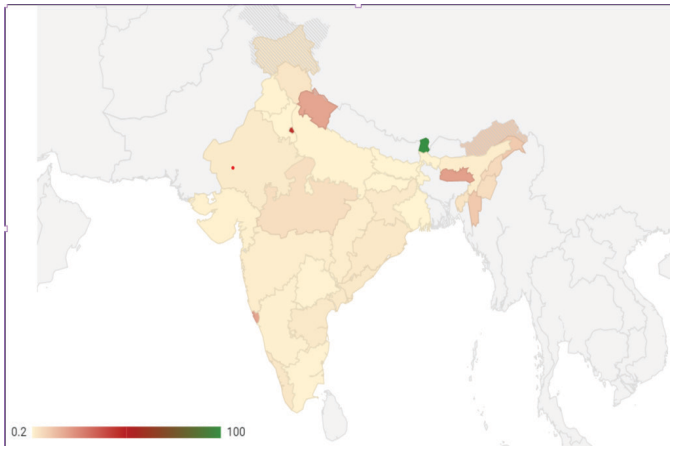


Figure 1. State-wise organic coverage in India (net sown area percentage) Data Source: apeda.gov.in

Organic Farming is an amalgamation of innovation and science to promote a sustainable relationship between plants, wildlife, and humans[4]. They help promote a balanced supply of nutrients, keeping the plants healthy, enhancing soil biological activity and nutrient mobilization in the soil. Due to Organic Farming practices, there has been an increase in the soil water retention capacity and a decrease in the acidity and salinity of soil. Organic Farming also helps to supply food for microorganisms and earthworms beneficial for soil fertility. Organic fertilizers have a low nutrient content and are required in large amounts to work effectively. This might lead to nutrient deficiency due to slow nutrient release to meet the crop

requirements. The cost of pursuing Organic Farming is very high compared to chemical and conventional farming resulting in its slower growth and adoption amongst farmers. Figure 1 depicts the current progress of the Indian states under Organic Farming. With 35% of its poor population, Sikkim was long regarded as one of India's poorest states and became the first organic state in the world[5]. Going entirely organic was not an easy transition for Sikkim. On the high, terraced slopes of the Himalayas, there were countless small farms. The 66,000 farmers in the state received agroecological farming instruction and a robust policy framework from the state government[6]. Many harvests failed in the first few years, which reduced agricultural output. The rapid removal of synthetic fertilizers from the soil necessitated many years for the earth to regain fertility. Ladakh, Meghalaya, Uttarakhand, and Goa follow this.

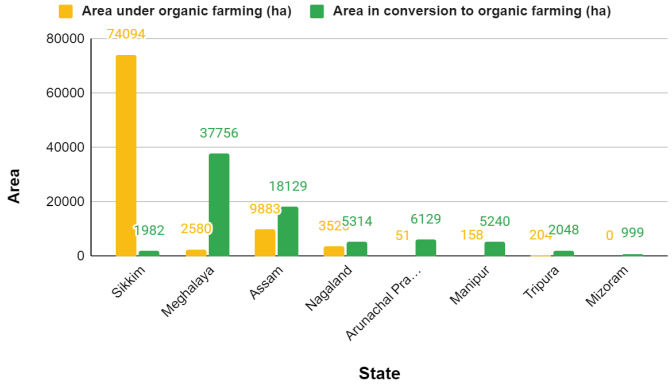


Figure 2. Area under organic cultivation and area in conversion to organic farming of North Eastern States (from 2017-18) (Source: apeda.gov.in).

Farming practices in North East India have traditionally been organic, using little chemical fertilizers (Fig. 2). Islands and tribal territories are now getting developed to maintain their biological history. India owns 26% of certified land under cultivation and 74% of the forest area, making it ranked 9th in terms of area under Organic Farming. With 585200 producers and 0.7% organic share, Madhya Pradesh and Uttar Pradesh are the greatest regions covered under Organic Farming, followed by Sikkim, which is recognized as a fully organic state as it completely disregards the usage of chemical fertilizers [7]. India has one of the largest organic wild harvest collections. More than 1.34 million metric tonnes of highly attested organic products, comprising all types of food items, which are consumable and non-consumable, were produced in India in the years 2015 and 2016[8].

Table 1. Organic products exported from India

Product Type	Organic Produce
Cereals	Wheat, Rice, Maize or Corn
Pulses	Red gram, Black gram
Fruits	Banana, Orange, Mango, Pineapple, Passionfruit
Oil and Seeds	Soyabean, Sunflower, Mustard, Groundnut, Cotton Seed
Vegetables	Brinjal, Garlic, Potato, Tomato, Onions
Herbs and Spices	Chilli, Peppermint, Cardamom, Turmeric, Black Pepper, Amla, Ginger
Others	Jaggery, Sugar, Tea, Coffee

Source: Chandrashekar 2010

Furthermore, the Agri-export Policy 2018 asserts the country is a potentially significant player in the world organic market. According to government estimates, India's top organic exports include flax seeds, sesame, soybean, tea, medicinal plants, rice, and pulses (Table 1), expanding India's organic exports by about 50% in 2018–19, reaching Rs 5,151 million[9].

This study aims to understand the history of Organic Farming in India and the world through various scientific literature and draw inferences about Organic Farming's potential growth in the country. Organic Farming is slowly gaining pace in India and is becoming widespread. However, Organic agricultural methods are unknown to the local farmers, who favor traditional agriculture, which yields goods faster by utilizing poisonous chemical pesticides and fertilizers, affecting human health and the environment. Awareness-raising initiatives should be planned to increase the practice of Organic Farming so that farmers may learn about Organic Farming and its benefits. The ultimate goal is to evaluate the factors which may facilitate the adoption of Organic Farming in the country and provide recommendations to expedite the adoption of organic agriculture in India[10].

Based on scientific literature and reports on Organic Farming practices at the Global and Indian levels, a thorough literature review was carried out. Most studies seek to minimize extraneous variations such as soil type, geographic location and living standards with different degrees of success. We further read and researched case studies to understand the real-life impact of Organic Farming on countries. China and Sri Lanka have been chosen to show contrasting viewpoints on the advantages and disadvantages of Organic Farming and to emphasize formulating an accurate action plan for India. The study concludes by identifying potential challenges of expanding Organic Farming in India and providing a long-term and short-term implementation plan to expand Organic Farming culture in India with strengthening governmental policies and support.

2. INTERNATIONAL STATUS OF ORGANIC FARMING

Lockeretz, *et al*. (1987) compared the economic performance of 14 organic crop/livestock farms in the Midwest to that of 14 conventional farms[11]. The farms

were linked based on the physical characteristics and different types of agricultural operations. Organic farms' market value of the crops produced per unit area was 11% lower. But, because the cost of production was also lower, both systems had equivalent net revenue per unit area. The net returns on organic farms were often more. According to both research, organic farms had reduced production costs.

The yield of the organic movement is a frequent topic of inquiry. Can Organic Farming feed everyone on the planet? Even high-input, high-yield systems are now failing to feed the globe due to problems with food distribution, social organization, and grave concerns about poverty, racism, and gender[12]. Less food will be available if the land is converted from inorganic to Organic Farming because of yield losses during the conversion phase. Such organically grown food is only purchased by the wealthy. As a result, less food is accessible to the poor. The price of the food they can buy goes up. Issues with equity result from this. Despite all of this, Organic Farming is effective. Some research showed sustained yields (albeit somewhat lower in many years) without agrochemicals in temperate climatic conditions based on tests running for 25 years in Switzerland. On the other hand, many agricultural experts think that large amounts of farm yard manure (FYM) and other biomass material, which will be required to make up for the fertilizers, are inaccessible without chemical fertilizers. Moreover, they think that some crops need the help of agrochemicals, particularly fertilizers to achieve greater output yield-wise and that adopting Organic Farming results in food insecurity for the nation[13].

Darnhofer (2010) noted that consumer tastes shifted towards chemical-free organic foods [14]. Organic Farming has spread to 187 nations, with 3.1 million farmers managing 72.3 million hectares of land. Australia (35.69 mha) has the highest amount of organic land, followed by Argentina (3.63 mha) and Spain (2.35 mha)[15]. All areas have seen growth in organic agricultural land and retail sales. Organic Farming is a promising sector with immense potential to grow and revolutionize the global agricultural industry.

Due to its numerous advantages, Organic Farming has become globally popularized as a propitious alternative to conventional agriculture. Since 2000, there has been a prominent rise in the overall Organic Farming production area, especially in industrialized countries[16]. Oceania accounts for approximately half of the world's farmed land with an estimated 35.9 million hectares (mha) of Organic Farming, followed by Europe (14.6 mha), South America (8 mha), Asia (6.1 mha), North America (3.2 mha), and Africa (3.2 mha).

Figure 3 depicts the Top 10 countries for Organic Farming in Asia. China has an Organic Farming land area of 1,90,000 hectares and is the fourth largest consumer of organic goods. Under arable and permanent crops, oil seeds, coconuts, cotton, coffee, temperate fruits, and tea make up the majority of the organic output

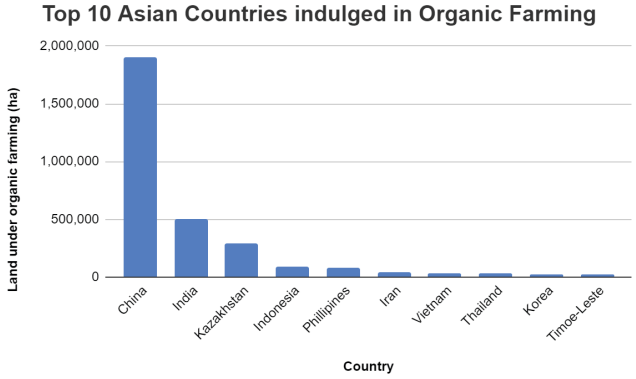


Figure 3. Top countries for organic farming in asia [Source: USDA] (Source: agricoop.nic.in)

in Asia. In 2015, 5.5 million hectares of wild organic agricultural land was used for growing fungi produce, medicinal plants, and oil[17]. In 2015, organic food sales in Asia, Australia, and other areas totaled 7.2 billion US dollars, boosting Asia's export market. The key nations with the highest sales of domestic organic products are South Korea, Taiwan, Japan, Malaysia, and Singapore[18].

3. STATUS OF ORGANIC FARMING IN INDIA

According to Tambe 2020, Organic Farming is a labor-intensive procedure that is highly restricted to farmers with abundant resources and the export market. It relies on external support systems for the price, market data, and product certification. As a result, he has concluded that Organic Farming likewise has a narrow area of application and societal value. Instead, he has advocated ecological farming; the major goals are maintaining high output, lowering production costs, and increasing self-sufficiency[19]. Both resource-poor and resource-rich farmers may benefit from it. The method is straightforward, it targets local markets, and its scope of coverage and social significance are decently high.

According to the Ministry of Agriculture and Farmers' Welfare, in 2020-21, 0.98 million farmers were brought under Organic Farming. Under the schemes, about 0.94 million ha area in India is under Organic Farming. The Government of India has been consistently promoting the country's chemical-free farming practices through dedicated schemes since 2015. Schemes such as PKVY and MOVCDNER provide end-to-end support to farmers in establishing, certifying and marketing the Organic Farming practice on their land[20]. However, Organic Farming production has been volatile, as seen in Table 2. The amount spent on schemes and programmes to promote Organic Farming is dwarfed by the annual subsidy provided on chemical fertilizers[21].

Table 2. Growth of organic farming in India over the years

Year	Organic Farming Area (Hectares)	Organic Production (MT)*	No. of Farmers

2015-16	19281	6321661	19355
2016-17	96291.16	8760811	173846
2017-18	6455	17132676	84618
2018-19	124990	989255	166571
2019-20	222369	2047536	365253
2020-21	7568	3399520	12074

Source: NCOF report; *MT- Metric Tonnes

There has been a rapid decrease in organic fertilizer production in India from 338720 million tonne in 2017-18 to 3879 million tonne in 2020-21. The possible reasons could be the organic fertilizer industry is hampered by low and uncertain demand, which prevents it from using its capacity optimally and discourages additional investment. Duplicity in securing licenses and official permission for manufacturing, selling, and quality testing of organic fertilizers hinder the manufacturing, selling, and quality testing of organic fertilizers. Using organic fertilizers and on-farm inputs can take time and resources compared to using chemical fertilizers[22]. A minimal financial allocation has been set out in public research institutes for organic fertilizer research and development.

Drylands that are semiarid and arid, on various occasions, have inadequate organic matter and water-holding ability[23]. Soil fertility is declining and certain serious pests are developing a resistance to synthetic pesticides in many areas where heavy input agricultural systems are used[24]. Many of these are signs of inefficient land use, which contributes to desertification; adopting Organic Farming techniques suited for drylands can assist in improving these situations.

4. ORGANIC VS CONVENTIONAL FARMING

Organic agriculture varies fundamentally from conventional agriculture. Thus, the ongoing use of Organic Farming practices appears to change the agro-functioning ecosystem. Organic farms typically support 30% more biodiversity than traditional farms. Even while conventional farming consistently adheres to some ecological principles, Organic Farming tends to increase soil fertility in a way that conventional farming cannot[25]. One approach to sustainable agriculture is Organic Farming. Many of these techniques utilized crop rotation, intercropping, mulching, double digging, and crop integration[26].

Figure 4 explains the existing drawbacks of conventional farming, emphasizing why there must be a significant shift to Organic Farming from conventional farming. Health consciousness and public readiness to pay for expensive organic goods are the main variables affecting consumer demand for organic food. Organic products have a rich, health-conscious customer base driven by a liberal price premium and environmental concerns [27].

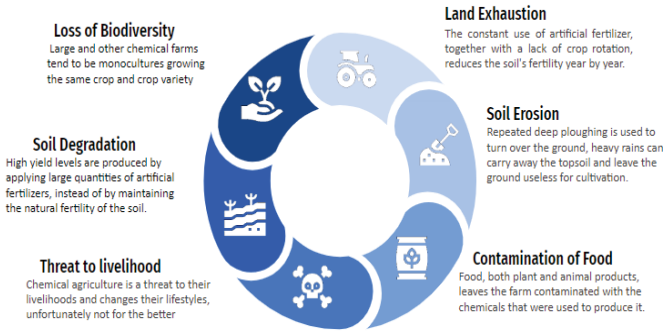


Figure 4. Bottlenecks of conventional farming

The literature research has demonstrated differences in viewpoints on Organic Farming, particularly among specialists. Although there is much disagreement on the profitability and yield growth of Organic Farming, there is broad agreement regarding its potential to safeguard human health and preserve the environment. Studies testify that Organic Farming is environmentally friendly. Still, there have been fewer efforts to promote Organic Farming in the form of subsidies, training facilities, etc., compared to conventional farming. This drawback can be addressed with the right incentives and policies. India, particularly in the dryland areas of the nation, may benefit from the varied soil and climatic conditions.

5. POLICIES FOR ORGANIC FARMING IN INDIA

Some initiatives the Ministry of Agriculture and Farmers Welfare took to promote Organic Farming in the country and equip farmers with the proper knowledge and skills are mentioned in Table 3[28].

In order to develop a plan of action for Organic Farming implementation in India, it is essential to study the growth and failures of other countries and learn from their methods. The prime focus of this section is to scrutinize the past experiences of India's neighbouring countries - Sri Lanka and China's approaches to Organic Farming, as well as the key actions that have been made to ensure the success or failure of this type of farming.

6. THE SRI LANKA CASE

In 2021, the President of Sri Lanka, Gotabaya Rajapaksa, decided to stop the import of chemical fertilizers and pesticides. Due to this, Organic Farming was picked up by all Sri Lankan farmers, and then the devastating effects ensued[36]. According to a source, Sri Lanka's president Gotabaya Rajapaksa planned to save $400 million yearly on chemical fertilizers and pesticides. Along with this, the president believed that using chemicals in agriculture affected the health of humans and the environment. Once independent in rice production, the country spent $450 million on its imports. The ban reduced production by 20% in six months[37]. Due to this reason, Sri Lanka saw a decrease in its output

Table 3. Policies to promote organic farming in India

Paramparagat Krishi Vikas Yojana (PKVY)	The PKVY started in 2015 to encourage cluster-based Organic Farming through a participatory guarantee system for India. The Bhartiya Prakritik Krishi Paddhati, based on principles of natural farming, was included in the PKVY as a sub-scheme in 2020–21. The National Mission on Clean Ganga (NMCG) has also been made part of PKVY[29].
Mission Organic Value Chain Development for North Eastern Region (MOVCDNER)	With a vision to establish certified organic production in a sustainable value chain, the scheme has helped to connect farmers with customers since 2015. They aim to create a sustainable value chain from the input of seeds to establishing facilities for aggregation, processing and marketing of organic products via brand-building initiatives[30].
One District-One Product (ODOP)	It was initiated in 2018 and the program uses the One District One Product (ODOP) strategy to leverage scale in input procurement, shared services, and product marketing. In districts, there could be more than one cluster of ODOP products[31].
National Policy for Farmers, 2007	This strategy supports farmers' adoption of Organic Farming and highlights its significance. It supports using organic inputs and traditional knowledge for sustainable agriculture and offers incentives for Organic Farming[32].
National Program for Organic Production (NPOP)	It is an Indian certification scheme for organic goods. The program specifies requirements for organic cultivation, certification, and body accreditation. In India, NPOP certification is a requirement for all organic[33].
Sikkim Organic Mission	The first entirely organic state in India is Sikkim. The Sikkim Organic Mission was established in 2010 to make all of the state's agricultural land organic. Farmers that embrace organic agricultural methods receive incentives and subsidies from the state government[34].
Karnataka State Organic Farming Policy	The Karnataka State Organic Farming Policy was introduced in 2017 to elevate Organic Farming practices in the state. The policy aims to promote Organic Farming practices and increase the area under organic cultivation in the state. It encourages farmers to adopt Organic Farming practices through awareness campaigns, capacity building, and financial incentives[35].

of crops by 18%. The Sri Lankan government had to spend a lot of money in the form of subsidies to help the farmers cover up for this loss in productivity.

7. THE CHINA CASE

China is rated 3rd internationally in certified organic products and 4th globally in sales. In China, organic food production and sustainable agricultural methods are increasing, based on a government study from 2019. Between 2005 and 2018, the total area under cultivation for certified organic agriculture expanded more than five times to 3.1 million hectares[38]. China exports agri-food products worth US$65 billion annually. With various forms of official assistance, farmers are abandoning using chemicals in their agriculture to promote and advance individuals' health, environmental preservation, and economic motivations. Chinese customers are eager to eat food without chemicals, mainly for health reasons. Organic and "green" food demand is rising quickly, particularly among the middle and upper classes[39].

8. INFERENCE FROM THE CASES

Chemical fertilizers and pesticides considerably affect the health of humans and the environment. This was the primary reason for the ban, but it is also important to realize that these chemicals allowed the farmers to grow more crops on a smaller piece of land. This quality of chemicals that enable the farmer to produce higher yields is a huge plus in upcoming nations like Sri Lanka. In 2000, 83% of Sri Lankan population was categorized as nourished. However, in 2019, the population of Sri Lankans who were nourished was just 7%. At the same time, China has made significant research investments and has institutions supporting the organic industry that offers training and information sharing. Twenty-one million farmers were taught soil, water, and fertilizer management. This initiative increased maize, wheat, and rice yields by 11% while reducing nitrogen fertilizer consumption by 15% to 18%.

The Chinese experience has been innovative in that it has established a local standard, the Green Food standard, and has spread, tested, and refined it away from the international spotlight. It has also differentiated certification into Grade A and Grade AA. This tactic has aided in China's quick adoption of organic foods. While the Chinese gradually acclimated the farmers towards Organic Farming through flexible and lucrative policies, the sudden policy shift in Sri Lanka and the lack of a concrete supply chain system led to the downfall of the economy and the environment in Sri Lanka.

9. CHALLENGES IN ORGANIC FARMING

India can learn from other countries' successes and mistakes in expediting the Organic Farming development process. To implement an effective action plan, it is necessary to identify potential challenges which may be a roadblock to Organic Farming expansion in India.

9.1 Shortage of Biomass

Many erudite scholars and experienced farmers are skeptical about the nutrient value that organic fertilizers could impart. Researchers believe this organic matter is not enough to be accessible to all.

9.2 Disparity of Supply and Demand

Fruits and vegetables cannot be grown or moved to any area like non-perishable grains can, yet non-perishable grains can be cultivated anywhere. It ought to be produced close to the demand, and there ought to be willing businesses and farmers there. However, the demand mostly originates from urban areas where no farms can grow organic fruits. The solution to this smart transfer is through experienced routes.

9.3 Time

More interaction between a farmer and his crops is required in Organic Farming, such as weed control, early intervention, and crop monitoring. With conventional techniques and chemical fertilizers, the time taken to produce more crops is less.

9.4 High MRP

Organic Farming is cost sensitive due to the extreme care needed in its practice, as a result their cost of production is also high. This leads to customer segmentation, mostly the rich and environmentally conscious people would buy organic products which hinders its expansion over the country.

9.5 Lack of Special Infrastructure

Handling organic produce is a very necessary step in expanding and flourishing the organic farming industry in the country. Due to lack of infrastructure and facilities there is massive spoilage which is a major challenge.

10. APPLICATION OF ORGANIC FARMING IN LAND RECLAMATION

India accounts for 16% of the global population, although its land only covers 2% of its total area. Naturally, the soil is sometimes under more stress than it can handle. As a result, India's productive resources, particularly its agricultural land, are constantly undergoing varying degrees of deterioration and are quickly becoming a wasteland. In India, there are currently 68.35 million hectares of wastelands on the planet. About 50% of these areas are non-forest lands of this type, which, with the proper care, may be restored to fertility[40]. The chemical and physical functions of the soil, such as sorption capacity, nutrient mobilization, and long-term fixation, are all improved by organic additions. Organic matter improves soil quality and serves as a store for nutrients, preventing nutrient loss through leaching and erosion.

Green manures act as mulch for the soil, protecting it from moisture loss and wind/water erosion and increasing the soil's organic matter. Nitrogen-fixing bacteria of their root nodules aid in absorbing nitrogen from the atmosphere. Land Remediation Techniques cover methods like composting and farming on land. Inland farming, polluted soils are transported to farming locations where they are repeatedly turned over and tilled to provide for aeration. The government or local authority concerned must provide financing or approve large-scale, expensive soil remediation projects[41].

11. SOCIO-ECONOMIC IMPACT THROUGH ORGANIC FARMING

Because Organic Farming needs more effort, each farm creates more employment that generates cash. Their occupational health is enhanced since farmers and laborers on organic fields are less likely to be exposed to agriculture-related chemicals[42].

Organic produce is healthier and more durable due to low nitrate and high levels of antioxidants which assist in extending their shelf life[43]. The industry is experiencing a boom in the economy due to increased profits generated by Organic Farming. The idea of industry-driven "One district - One product" is now being encouraged, as is the creation of new clusters close to larger towns where the demand for organics will be considerably greater. By eliminating synthetic fibers and carbon sequestration, Organic Farming can also help lower emissions and the atmospheric carbon dioxide level[44].

The second aspect impacting farmer inclination towards Organic Farming is perceived benefit. An individual is more open to adopting new behavior the larger its benefits. Farmers are more likely to use Organic Farming if they believe that cultivating organic crops would minimize crop disease and produce better harvest quality yielding more profits.

12. THE ANNUAL BUDGET OF INDIA 2023

The government would help 10 million farmers switch to natural (or chemical-balanced) farming during the following three years, according to Budget 2023. The budget initiated a PM PRANAM program for the restoration, awareness, nutrition and improvement of Mother Earth. The increase in soil health in India is one of the main advantages of natural farming. In traditional agriculture, the misuse of synthetic fertilizers and pesticides can degrade the soil, lowering its fertility and capacity to sustain strong plant development[45].

The livelihoods of Indian farmers may potentially be improved through natural farming. Natural food is frequently more expensive than traditionally produced, giving farmers better incomes and more stable living

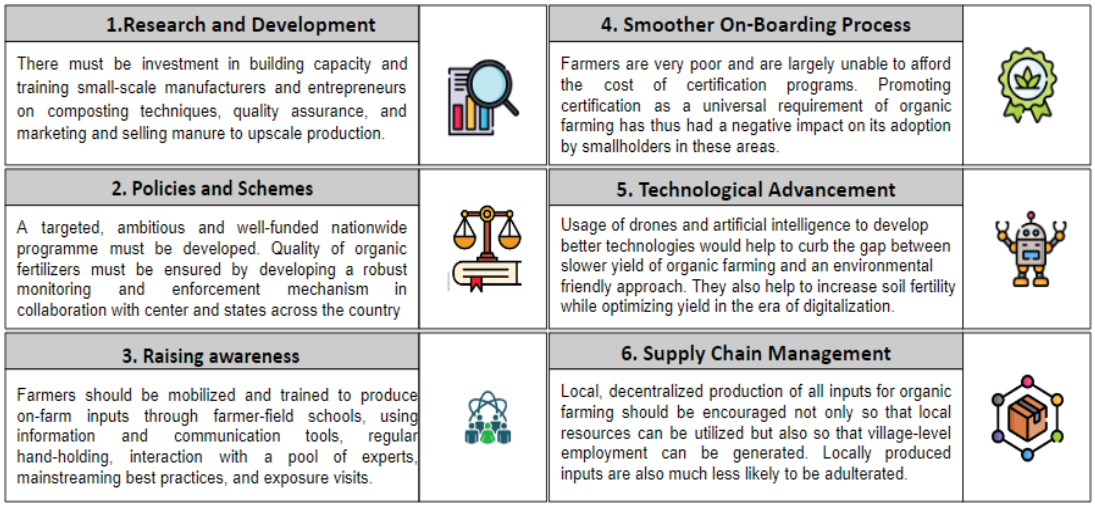


Figure 5. Strategies to promote organic farming in India.

conditions. Natural farming can also lower the danger of crop failures since organic farmers use a variety of crops, which lowers the chance of crop failures brought on by pests and diseases[46].

13. RECOMMENDATIONS

13.1 Strategies to promote Organic Farming in India

After analyzing the scenario of Organic Farming in Sri Lanka and China, we feel it would be in India's best interest to adopt the following practices and develop a comprehensive action with both long-term and short-term strategies explained in Figure 5 and 6.

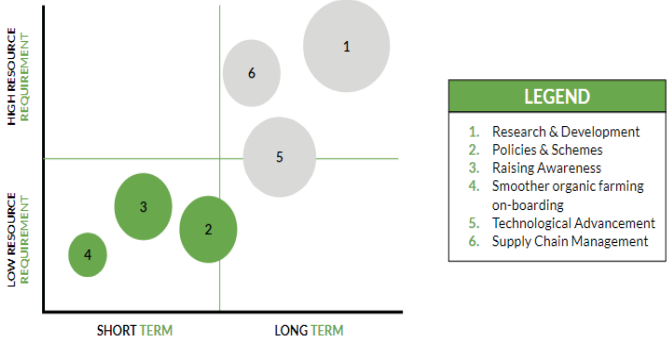


Figure 6. Long-term and Short-term map to implement strategies.

This chart bifurcates the above-mentioned strategies into long-term and short-term implementation goals. The size of these bubbles is directly proportional to the expected impact, while the color represents the priority of the steps mentioned above. Green bubbles are recommended for immediate implementation, while grey bubbles depend on government budgetary constraints.

14. CONCLUSION

Organic Farming is an amalgamation of innovation and science to promote a sustainable relationship between plants, wildlife, and humans. They are beneficial to the environment and increase the quality of life by producing healthier products. Organic Farming also helps to supply food for microorganisms and earthworms beneficial for soil fertility. The sector is pacing up with time and has spread to 187 nations, with 3.1 million farmers managing 72.3 million hectares of land. Despite being a consumer-centric and market-controlled sector, the organic industry has expanded to approximately 30% annually over the past ten years. The government is taking relevant measures to promote Organic Farming in India through multiple schemes and policies. However, organic agricultural methods are unknown to the local farmers, who favor traditional farming, which yields goods faster using poisonous chemical pesticides and fertilizers. Awareness-raising initiatives should be planned so that farmers can learn about Organic Farming and its benefits. It can be used in land rehabilitation practices across the country and help to reclaim barren land by improving soil quality and serving as a store for nutrients, preventing nutrient loss through leaching and erosion.

However, there is a need to create the right technology and infrastructure to distribute the farmers. Farmers must receive training in value-added technologies. India must learn from the cases of Sri Lanka and China and focus on prioritizing Organic Farming by addressing the existing problems and difficulties. The nation needs a robust organic policy creation.

ACKNOWLEDGEMENTS

The authors express their gratitude to everyone who contributed to the project at various stages and shared their research and findings. The authors thank the Advance Air and Acoustics Research laboratory members, Department of Environmental Engineering, Delhi Technological University, for supporting the entire study.

REFERENCES

1. FAO Publications Catalogue 2022, *Food and Agriculture Organization of the United Nations*, Rome, Italy, Oct. 2022, 136 p. doi: 10.4060/CC2323EN.
2. Kumar, Vineet. CSE report flags poor state of

organic fertilisers, biofertilisers sector in India, India, 2022. https://www.downtoearth.org.in/blog/agriculture/cse-report-flags-poor-state-of-organic-fertilisers-biofertilisers-sector-in-india-82647 (Accessed Feb. 28, 2023).

3. A. K. Barik & Sarkar, N.C. Organic Farming in India: Present Status, Challenges and Technological Break Through. *In* Proceedings of Third International Conference on Bio-resources and Stress Management, 2017, 84-93.
4. Babu, C. & Karunakaran, N. Status, benefits and future prospects of Organic Farming in India: A review. *J. of Management Res. and Anal.*, 2021, **8**(3), 103–111. doi: 10.18231/J.JMRA.2021.022.
5. Chadha, D. & Srivastava, S. K. Growth Performance of Organic Agriculture in India. *Current J. of Appl. Sc. and Tech.*, 2020, **39**(33), 86–94**.** doi: 10.9734/CJAST/2020/V39I3331022.
6. Agriculturepost.com. 5 Govt schemes, promoting Organic Farming in India . India, 2023, https://agriculturepost.com/farm-inputs/5-govt-schemes-promoting-organic-farming-in-india/ (Accessed Feb. 28, 2023).
7. Department of Agriculture & Horticulture, Government of Assam, India . Mission Organic Value Chain Development in Assam (MOVCD). https://agri-horti.assam.gov.in/schemes/mission-organic-value-chain-development-in-assam-movcd (Accessed Feb. 28, 2023).
8. Meena, S.S. & Bhusanar, S.B. Organic Farming in India Present Status and Challenges. *In* Handbook of Emerging Trends in Agribusiness and General Management , edited by Satyaveer Singh Meena, Vikash and Hanuman Prasad, Himanshu Publication Delhi and Udaipur, 2022, 207p. https://www.researchgate.net/publication/364322060_Organic_Farming_in_India_Present_Status_and_Challenges (Accessed Feb. 28, 2023).
9. Manifest Learning Academy. Status of Organic Farming in India – Journals of India. https://journalsofindia.com/status-of-organic-farming-in-india/ (Accessed Feb. 28, 2023).
10. Roychowdhury, R. ; Banerjee, U. ; Sofkova, S. ; & Tah, J. Organic Farming for crop improvement and sustainable agriculture in the era of climate change. *Online J. of Biological Scs.*,2013, **13**(2), 50–65, doi: 10.3844/OJBSCI.2013.50.65.
11. Lockeretz, W. ; Freedgood, J. ; & Coon, K. Farmers' views of the prospects for agriculture in a metropolitan area. *Agricultural Systems*, 1987, **23**(1), 43–61. doi: 10.1016/0308-521X(87)90072-2.
12 Letourneau, D. K & Goldstein, B. Pest damage and arthropod community structure in organic vs. conventional tomato production in California. *J. of Applied Ecology*, 2001, **38**(3), 557–570, Jun. 2001, doi: 10.1046/J.1365-2664.2001.00611.X.
13. Mäder, P. ; Fließbach, A. ; Dubois, D.; Gunst, l. ; Fried, P. ; & Niggli, U. Soil fertility and biodiversity in Organic Farming. *Science,* 2002, **296**(5573),1694–1697. doi: 10.1126/SCIENCE.1071148.
14. Darnhofer, I. Strategies of family farms to strengthen their resilience. *Environmental Policy and Governance*, 2010, **20**(4), 212–222. doi: 10.1002/EET.547.
15. Debnath, R.M. Opportunities and challenges of organic agri-food organisations: a case study from india. *Indian Journal of Public Administration*, 2021, **67**(4), 608–619. doi: 10.1177/00195561211042978.
16. Brar, B.S., Singh, J., Singh, G. and Gurpreet-Kaur. 2015. Effects of long-term application of inorganic and organic fertilizers on soil organic carbon and physical properties in maize–wheat rotation. *Agronomy,* 2015*,* **5**, 220- 238. doi: 10.1377/001355656211072978.
17. M. M and G. M. K., "New Agriculture Technology in Modern Farming," *International Journal of Management Research and Social Science*, **8**(3), doi: 10.30726/IJMRSS/V8.I3.2021.83016.
18. Garg A., "Recent Trends in Agriculture: Vertical Farming and Organic Farming," *Advances in Plants & Agriculture Research*, 2014, **1**(4), doi: 10.15406/APAR.2014.01.00023.
19. Satish S. Tambe, "Our Heritage Recent Trends In Organic Farming In India: A Review Article." https://www.researchgate.net/publication/350134533_ (Accessed on Mar. 01, 2023).
20. Singh P., "Management of the Pandemic: Agriculture, Food Management and Resilience During Covid-19 in India," *Indian Journal of Public Administration*, 2019, **67**(3), 324–336, doi: 10.1177/00195561211045094.
21. Yadav C. and Pandey S., "Status of the use of Organic Fertilizers in India: A Review," *Agricultural Reviews*, 2020, **7**, doi: 10.18805/AG.R-1961.
22. Aryal J.P., Sapkota T.B., Krupnik T. J., Rahut D. B., Jat M. L., and Stirling C. M., "Factors affecting farmers' use of organic and inorganic fertilizers in South Asia," *Environmental Science and Pollution Research*, 2021, **28**(37), 51480–51496, doi: 10.1007/S11356-021-13975-7/TABLES/10.
23. Sharma A.K., "ALN No. 58: Sharma: The potential for Organic Farming in the drylands of India **58**, Winter 2005 Soil management for drylands the potential for Organic Farming in the drylands of India", http://cals.arizona.edu/OALS/ALN/aln58/sharma.html (Accessed Feb. 28, 2023)
24. Mahadeo S Deshmukh, Nitin A Babar, "Present Status and Prospects of Organic Farming in India.", *Journal of Management Research and Analysis* https://www.researchgate.net/publication/282243181_Present_Status_and_Prospects_of_Organic_Farming_in_India (Accessed on Mar. 01, 2023).
25. Imran, Amanullah, Hussain I., "Agricultural soil reclamation and restoration of soil organic matter and nutrients via application of organic, inorganic and bio fertilization (Mini review)," *IOP Conference Series: Earth and Environmental Science*, **788**(1) , 012165, doi: 10.1088/1755-1315/788/1/012165.
26. Aryal J.P., Sapkota T., Rahut D.B., and Jat M.L.,

"Agricultural sustainability under emerging climatic variability: the role of climate-smart agriculture and relevant policies in India," *International Journal of Innovation and Sustainable Development*, **14**(2), 219–245, 2020, doi: 10.1504/ijisd.2020.106243.
27. C. Liu, H. Cutforth, Q. Chai, and Y. Gan, "Farming tactics to reduce the carbon footprint of crop cultivation in semiarid areas. A review," *Agronomy for Sustainable Development*, **36**(4), doi: 10.1007/s13593-016-0404-8.
28. "MOVCD-NER." https://movcd.dac.gov.in/ (Accessed on Feb. 28, 2023).
29. "Paramparagat Krishi Vikas Yojana (PKVY) Manual for District-Level Functionaries 2017". (Accessed Feb. 28, 2023)
30. "Mission Organic Value Chain Development for North Eastern Regions (MOVCDNER) INM Division, Ministry of Agriculture and Farmers' Welfare Mission Organic Value Chain Development for North Eastern Region a Sub-Mission under National Mission for Sustainable Agriculture (NMSA) Operational Guidelines," 2018.
31. "Official Website of One District One Product Uttar Pradesh" https://odopup.in/en/page/district-wise-products (Accessed on Feb. 28, 2023).
32. "National Policy for Farmers 2007 by Department of Agriculture and Farmers Welfare| National Portal of India." https://www.india.gov.in/national-policy-farmers-2007-department-agriculture-and-cooperation (accessed Feb. 28, 2023).
33. "5 Govt schemes, promoting Organic Farming in India - Agriculture Post." https://agriculturepost.com/farm-inputs/5-govt-schemes-promoting-organic-farming-in-india/ (Accessed on Feb. 28, 2023).
34. "Sikkim: India's First Organic State; RTSV - In Depth programme analysis." https://byjus.com/free-ias-prep/sikkim-indias-first-organic-state-rstv-in-depth/ (Accessed on Feb. 28, 2023).
35. "Organic Farming in Karnataka, How to Start | Agri Farming." https://www.agrifarming.in/organic-farming-in-karnataka-how-to-start (accessed Feb. 28, 2023).
36. "Sri Lanka's Organic Farming Experiment Went Catastrophically Wrong." https://foreignpolicy.com/2022/03/05/sri-lanka-organic-farming-crisis/ (Accessed on Feb. 28, 2023).
37. Narayanan D.S., "Organic Farming in India: relevance, problems and constraints,"*Sustainable Agriculture,* **5**(4), 2005.
38. "Why China is emerging as a leader in sustainable and organic agriculture." https://theconversation.com/why-china-is-emerging-as-a-leader-in-sustainable-and-organic-agriculture-132407 (accessed Feb. 28, 2023).
39. Scott S., Si Z., Schumilas T., and Chen A., "Organic food and farming in china: Top-down and bottom-up ecological initiatives," *Organic Food and Farming in China: Top-down and Bottom-up Ecological Initiatives*, pp. 1–218, 2018, doi: 10.4324/9780203701706.
40. S. M and Basavaraju S.M., "Growth and present status of Organic Farming in india and global level," *Paripex Indian Journal of Research*, pp. 56–57, 2022, doi: 10.36106/PARIPEX/0807327.
41. Singh R.K., Singha M., Singh S. K., Pal D., Tripathi N., and Singh R. S. , "Land use/land cover change detection analysis using remote sensing and gis of dhanbad district, india," *Eurasian Journal of Forest Science*, 2018, doi: 10.31195/EJEJFS.428381.
42. Doran J.W. and Zeiss M., "Soil health and sustainability: Managing the biotic component of soil quality," *Applied Soil Ecology*, **15**(1), 3–11, doi: 10.1016/S0929-1393(00)00067-6.
43. Urra J., Alkorta I., and Garbisu C., "Potential Benefits and Risks for Soil Health Derived from the Use of Organic Amendments in Agriculture," *Agronomy 2019,* **9**(9), 542, doi: 10.3390/AGRONOMY9090542.
44. "Reclamation of Problem Soils (RPS)-A Sub Scheme of RKVY Government of India Ministry of Agriculture & Farmers Welfare Department of Agriculture, Cooperation& Farmers Welfare (Natural Resource Management Division)". https://pib.gov.in/PressReleseDetail.aspx?PRID=1741966 (Accessed Feb. 28, 2023)
45. McArthur J.W. and McCord G.C., "Fertilizing growth: Agricultural inputs and their effects in economic development," *Journal of Development Economics*, 2017, **127**, 133–152, doi: 10.1016/j.jdeveco.2017.02.007.
46. Kalpana P., SaiBramari G., and Anitha L., "Formulation of potential vegetable waste compost in association with microorganisms and Spirulina platensis," *Asian Journal of Plant Science & Research*, 2011,**1**(1)**,** doi: 10.1107/s14573-026-0404-6.

CONTRIBUTORS

Ms N.Krithika is a final year undergraduate student pursuing B.Tech. in Environmental Engineering from Delhi Technological University, Delhi, India. In this study, she has contributed to data interpretation, literature analysis of the work and text writing of the manuscript

Mr Rishabh Jain is a final year undergraduate student pursuing B.Tech. in Environmental Engineering from Delhi Technological University, Delhi, India. In this study, he has contributed to data interpretation, literature analysis of the work and text writing of the manuscript.

Ms Monika Sharma holds a Master of Science degree in Environmental Science from Dr. Bhimrao Ambedkar Univesity, Agra, India. She is currently pursuing her PhD at Delhi Technological Univesity, Delhi, India. In this study, she has performed a critical review and edited the manuscript.

Dr Rajeev Kumar Mishra is an Assistant Professor in the Department of Environmental Engineering at Delhi Technological University, Delhi, India. He holds a PhD from the Indian Institute of Technology, Roorkee. He has contributed to the guidance and framing of the research study for this work. He has also performed a critical review and edited the manuscript.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/375870704

Parametrical Analysis and Thermal Sensitivity Study of 8-bit Vedic Multiplier

Conference Paper · July 2023

DOI: 10.1109/ICCCNT56998.2023.10306486

CITATIONS
0

READ
1

2 authors:

Hemanshi Chugh
Delhi Technological University
8 PUBLICATIONS **1** CITATION
SEE PROFILE

Sonal Singh
Delhi Technological University
21 PUBLICATIONS **857** CITATIONS
SEE PROFILE

All content following this page was uploaded by Hemanshi Chugh on 27 November 2023.

The user has requested enhancement of the downloaded file.

Parametrical Analysis and Thermal Sensitivity Study of 8-bit Vedic Multiplier

Hemanshi Chugh
Electronics and Communications Department
Delhi Technological University
Delhi, India
hemanshi.chugh@gmail.com

Sonal Singh
Electronics and Communications Department
Delhi Technological University
Delhi, India
sonalsingh@dtu.ac.in

***Abstract*—The Vedic multiplier, as inspired by classical Indian mathematical principles, offers a unique strategy to improve computational efficiency. This manuscript focuses on the implementation of an 8-bit multiplier employing the Urdhva Tiryakbhyam algorithm in a 90 nm node technology file through Cadence Virtuoso. The suggested Vedic multiplier is thoroughly compared with the traditional multiplier in terms of performance parameters, such as delay, area, power delay product (PDP), and energy-delay product (EDP). The proposed implementation also indicates a substantial decline in transistor count, EDP and PDP of about 25.9%, 80.3% and 62.8%, respectively. Furthermore, the study examines the thermal sensitivity of the multipliers by analyzing the energy consumption in response to temperature variations. The energy consumption graph displays fluctuations with increasing temperature for the traditional 8-bit multiplier, while a stable and linear trend is observed for the proposed 8-bit multiplier. These findings highlight the potential of the Vedic multiplier in optimizing the performance and energy efficiency of digital systems. The energy versus temperature graph emphasizes the superiority of the Vedic multiplier across various operating temperatures. The proposed design exhibits reduced energy consumption compared to the traditional approach, making it an attractive solution for low-power applications. Overall, the study validates the efficiency and effectiveness of the Vedic multiplier, further establishing its potential to revolutionize digital system designs.**

***Index Terms*—Vedic Multiplier, Parametric Analysis, 8-bit multiplier, Urdhva-Tiryakbhyam, Thermal Sensitivity**

I. Introduction

THE multiplication is a fundamental operation in Digital Signal Processing (DSP) systems and plays a crucial role in a wide range of applications, including audio and video processing, communications, image processing, and scientific computing [1]–[4]. Efficient and high-speed multiplication is essential to meet the demanding computational requirements of these systems. As DSP applications continue to evolve and demand increasingly complex algorithms and higher processing speeds, the need for fast multipliers becomes even more critical. Traditional multiplication algorithms, such as the traditional binary multiplier, often struggle to meet these requirements due to their inherent complexities and longer computation times. Therefore, there is a growing demand for innovative multiplication techniques that can deliver improved performance and energy efficiency. Vedic multiplication, inspired by ancient Indian mathematics [5], has emerged as a promising alternative to traditional multiplication techniques in recent years. Among the various Vedic multiplication algorithms [6], the Urdhva Tiryakbhyam (UT) technique has gained significant attention for its ability to perform high-speed multiplication operations efficiently. This manuscript serves as an extension to the paper presented [7]. The proposed architecture in this study utilizes a novel 5-bit special adder, in conjunction with full adders and half adders, to implement a 4x4 multiplier based on the UT technique of Vedic mathematics. The results demonstrated significant improvements in the proposed 4x4 Vedic Multiplier(VM) compared to the traditional CMOS VM. Specifically, the circuit exhibited a 48.6% reduction in delay 52.2% reduction in power consumption, and a 75% decrease in the power-delay product (PDP). Additionally, this paper explores the energy efficiency of the traditional and proposed 4-bit multipliers under varying temperature conditions. Furthermore, an 8-bit multiplier is proposed by utilizing the modified 4-bit multiplier using Cadence Virtuoso in a 90 nm technology file, and a comparison is made between the traditional and proposed multipliers. By harnessing the power of ancient mathematical principles, the VM offers the potential to revolutionize multiplication operations in DSP systems [2]. The UT technique, with its parallelized and streamlined approach, addresses the limitations of traditional multipliers and enables faster computation with reduced energy consumption. The structure of the paper is outlined as follows: Section II provides a detailed explanation of the UT sutra, accompanied by an illustrative example. In Section III, the architectural designs, including block diagrams and schematic diagrams, of both the traditional and proposed designs are presented. Additionally, this section includes the verification of these designs through simulated waveforms. Section IV offers a comprehensive analysis of the 2-bit, 4-bit, and 8-bit multipliers, presenting the results and drawing inferences based on the findings. Finally, Section V summarizes the manuscript.

II. Urdhva Tiryakbhyam: A Vedic Sutra

The UT technique, also known as the vertical and crosswise algorithm [8], breaks down the multiplication process into a series of more uncomplicated steps that can be executed in parallel. This approach eliminates the need for time-consuming

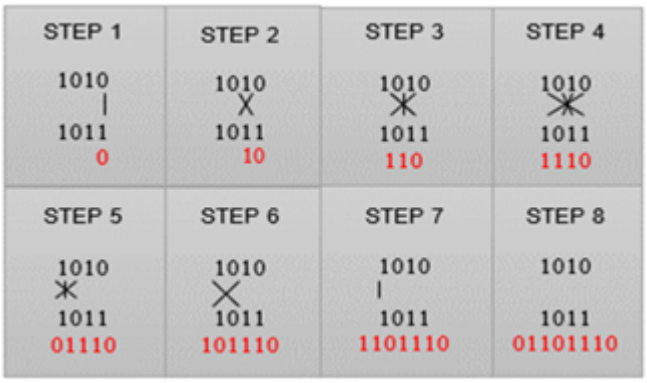


Fig. 1. 4-bit Vedic multiplication illustrated with the UT sutra

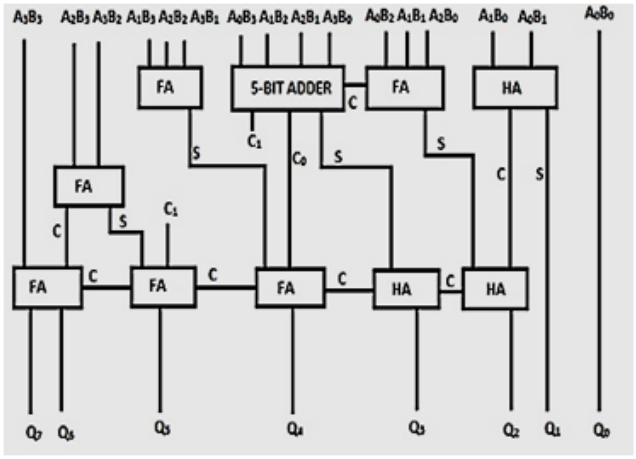


Fig. 2. Architecture of the proposed 4-bit VM using a special 5-bit adder

shifting and iterative addition, resulting in faster computation times. By leveraging the principles of Vedic mathematics, UT offers superior performance and energy efficiency compared to traditional multiplication algorithms [9]. Fig. 1 the step-by-step representation of the algorithm of the UT technique for 4-bit operands. In this illustration, we consider binary 4-bit operands: 1010 (decimal 10) and 1011 (decimal 11). The multiplication process involves the generation of partial products using the vertical and crosswise methods, progressing through 7 steps. The final product is obtained in step 8 and results in 01101110 (decimal 110). [7]. Similarly, the basics of this sutra can be applied to higher-order binary multiplication. The results of this sutra provide valuable insights into the benefits of Vedic multiplication techniques and their applicability in modern digital design. The findings contribute to the ongoing efforts in developing efficient and high-performance multipliers for DSP systems [3], [10], opening up new possibilities for improved computational capabilities in various domains.

III. Architecture Design & Verification

The proposed 4-bit Vedic multiplier architecture is depicted in Fig. 2 and the schematic representation is displayed in Fig. 3 This 4-bit VM is composed of 16 AND gates, 6 FAs, 3

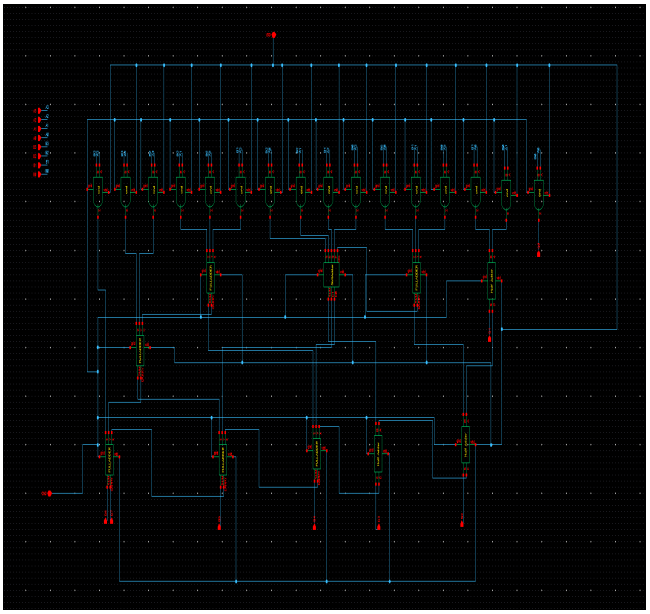


Fig. 3. Schematic of the proposed 4-bit VM using a special 5-bit adder

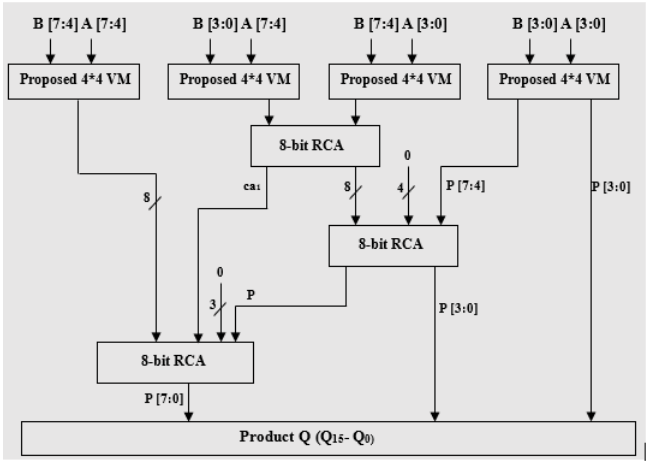


Fig. 4. Architecture of the proposed 8*8 multiplier

HAs, and a 5-bit special adder—a circuit that accepts 5-bit binary data and provides a 3-bit output. These components are intelligently combined to create an optimized VM that enhances performance while minimizing power consumption [7]. To implement an 8-bit VM, four modules of the proposed 4-bit Vedic multiplier are utilized, along with three 8-bit Ripple Carry Adders (RCAs) to perform the necessary additions and generate the final 16-bit output (Fig. 4). The schematic design of the 8-bit VM (refer to Figure 5) is implemented using the Cadence Virtuoso Design platform v6.1.5. The CMOS implementation of the VM leverages the principles of the UT algorithm, enabling faster computation for low-power and high-speed multiplication operations. The multiplier takes 8-bit inputs A and B and produces a 16-bit output. Inputs A and B represent the numbers to be multiplied, while the output represents the result of their multiplication. The schematic

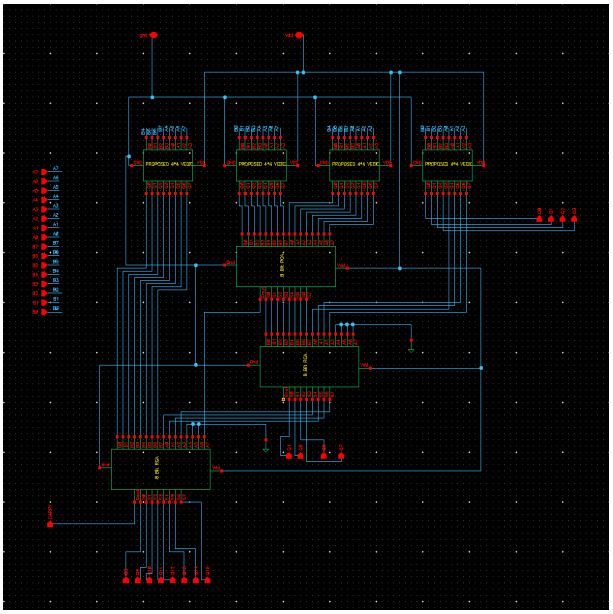


Fig. 5. Schematic of proposed Vedic 8-bit Multiplier

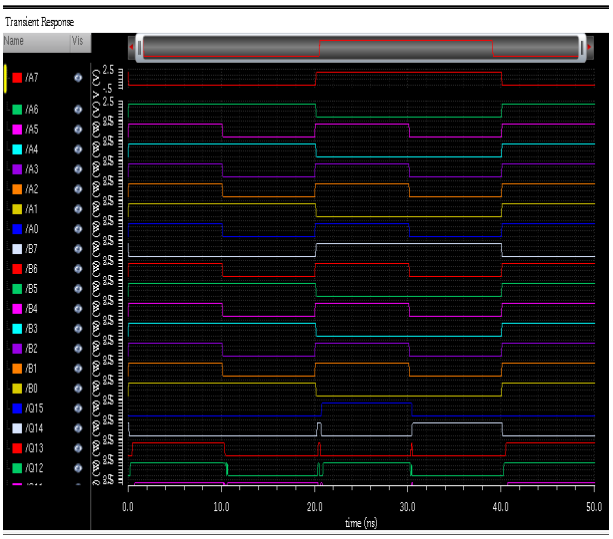


Fig. 6. Input Waveform of proposed Vedic 8-bit Multiplier

designs of the 8-bit VM have been verified and simulated in the Analog Design Environment (ADEL) using the Spectre simulation platform. The simulations were conducted using 90nm technology and a power supply of 2V. The simulated waveforms of the novel 4-bit Vedic multiplier are depicted in Fig. 6 and 7. The inputs to the multiplier, $A_7 - A_0$ and $B_7 - B_0$, are shown in the waveform (Fig. 6). The output of the multiplier, $Q_{15} - Q_0$ (as shown in Fig. 7), represents the 16-bit product. To ensure the functionality of the proposed multiplier across various scenarios, verification for all possible input combinations has been conducted.

In addition to the waveform simulations, the multiplication results for different input combinations have been tabulated in Table I. Through the verification and simulation process, the

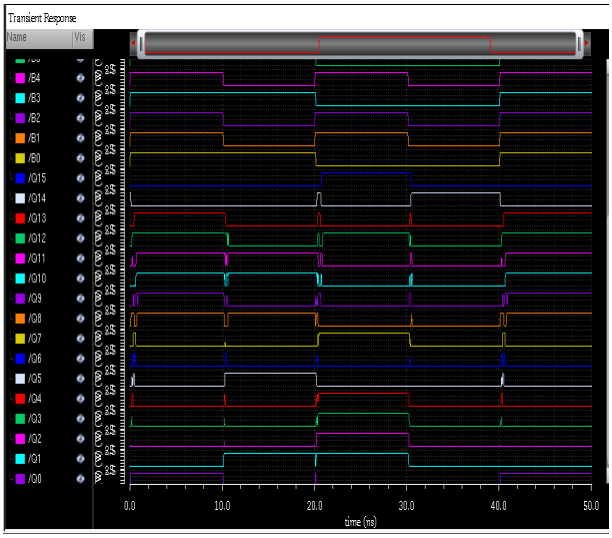


Fig. 7. Output Waveform of proposed Vedic 8-bit Multiplier

proposed 8-bit Vedic multiplier has proven its effectiveness and suitability for real-world applications. The combination of schematic verification, waveform simulations, and multiplication result analysis comprehensively evaluates the multiplier's performance and functionality.

TABLE I
MULTIPLICATION RESULTS OF PROPOSED 8-BIT VEDIC MULTIPLIER

Input A		Input B		Output Q	
Decimal	Binary	Decimal	Binary	Decimal	Binary
127	1111111	127	1111111	16129	0011111100000001
82	1010010	41	101001	3362	0000110100100010
173	10101101	214	11010110	37022	1001000010011110
128	10000000	128	10000000	16384	0100000000000000

IV. PARAMETRICAL ANALYSIS

The PDP is an indicator of the energy efficiency of the design, and sometimes designs with a low PDP tend to perform slowly. To evaluate the performance of the proposed design, another performance metric, EDP, is calculated. The PDP and EDP can be easily calculated using equations 1 and 2.

$$PDP = Power * Delay \tag{1}$$

$$EDP = PDP * Delay \tag{2}$$

A. Comprehensive Performance Evaluation of Multipliers

Table II provides a comprehensive parametric analysis of a 2-bit Vedic module. This module holds significant importance as it acts as a fundamental building block for higher-order architectures. By examining various parameters, the table offers valuable insights into the performance and characteristics of the module.

Additionally, Table III presents a comparative analysis between a traditional 4-bit multiplier and the proposed Vedic multiplier. The table offers quantitative data and metrics that facilitate a comprehensive understanding of the performance

TABLE II
PARAMETRICAL ANALYSIS OF 2-BIT VM

Parameters	2-bit Vedic Multiplier
Power	11.65 μW
Delay	0.11 ns
Energy	0.252 pJ
PDP	1.28 fJ
EDP	0.028 e^{-21} Js

disparities between traditional and Vedic multipliers.

TABLE III
COMPARATIVE PARAMETRICAL ANALYSIS OF 4-BIT VM

Parameters	4-bit Vedic Multipliers	
	traditional	**Proposed**
No. of accepted trans steps	1568	1329
Total transient time	24.32 sec	13.95 sec
Peak resident memory	57.5 MB	50.5 MB
Transistor Count	762	504
Delay	0.34 ns	0.17 ns
Power	143.6 μW	68.5 μW
Energy (0-90 ns)	13.10 pJ	9.30 pJ
PDP	49.06 fJ	12.01 fJ
EDP	0.016 e^{-21} Js	0.002 e^{-21} Js

The following inferences were drawn from Table III:

1) The delay of the 4-bit proposed VM was less in comparison to other designs.
2) In contrast to the traditional design, the proposed 4-bit VM exhibits minimum total power, PDP and EDP of 68.5 μW, 12.01 fJ, and 0.002 e^{-21} Js, respectively.
3) The energy consumption of the proposed circuit for the 0-90ns period is 29% less than that of the traditional multiplier.
4) As compared to the traditional design, the proposed architecture uses 504 transistors.

Furthermore, Table IV extends this comparative analysis to 8-bit multiplier designs when the traditional and proposed multiplier schematics are simulated for a transition time of 50 ns. The detailed examination and exploration of various parameters and performance indicators are provided in the table for a comprehensive overview of the benefits and advantages of the proposed VM over its traditional counterpart. The following inferences were drawn from Table IV:

1) The delay of the proposed 8-bit VM was lower compared to the traditional designs.
2) In contrast to the traditional design, the proposed 8-bit VM exhibits minimum total power, PDP, and EDP of 0.58 mW, 378.6 fJ, and 0.25 e^{-21} Js, respectively.
3) Compared with the traditional design, 1739 transistors are used in the proposed architecture.
4) The simulations of the total settling time of the proposed VM are 30.6% lower.
5) The peak value of resident memory decreases by 18.5% in the proposed 8-bit VM.

TABLE IV
COMPARATIVE PARAMETRICAL ANALYSIS OF 8-BIT VM

Parameters	8-bit Vedic Multipliers	
	traditional	**Proposed**
No. of accepted trans steps	1799	1739
Total transient time	39.521 sec	27.41 sec
Peak resident memory	135 MB	110 MB
Transistor Count	3984	2952
Delay	1.25 ns	0.65 ns
Power	0.82 mW	0.58 mW
Energy (0-40 ns)	33.49 pJ	27.38 pJ
PDP	1019.1 fJ	378.6 fJ
EDP	1.27 e^{-21} Js	0.25 e^{-21} Js

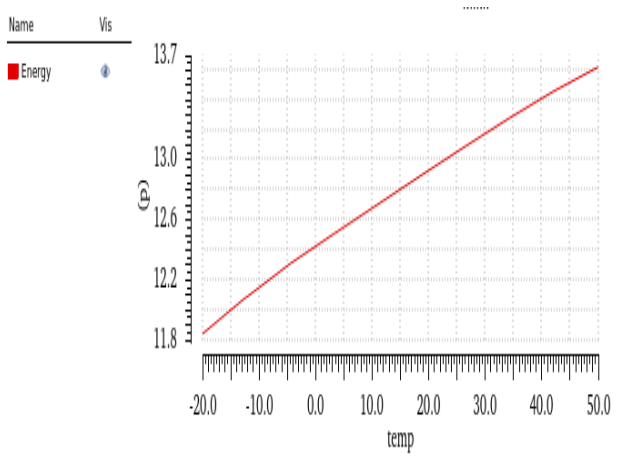


Fig. 8. Temperature-dependent Energy Consumption of Traditional 4-bit VM

6) The energy consumption of the proposed circuit for the 0-40ns period is 18% less than that of the traditional multiplier.

B. Impact of Temperature Variation on Multiplier's Energy Consumption

This section aims to analyze the thermal sensitivity of the multipliers. Fig. 8 and Fig. 9 show the impact on energy consumption in terms of temperature variations for 4-bit implementations, while Fig. 10 and Fig. 11 show for the 8-bit multiplier implementations. The effect of temperature fluctuations (ranging from -20 to 50° Celsius) on energy usage are depicted graphically. This analysis is crucial for understanding the thermal behavior of the circuit and its impact on overall energy efficiency. Furthermore, the energy graph described in the figures offers a visual representation of the observed patterns in energy usage. It provides a precise illustration of how energy usage changes linearly with temperature variations. For a more detailed breakdown of the energy data, Table V and VI accompany the figures, providing a comprehensive summary of the energy analysis conducted in the manuscript.

The presented Tables V and VI clearly illustrate the energy efficient characteristics of the proposed 4-bit and 8-bit VMs across varying temperature levels. The comparative analysis

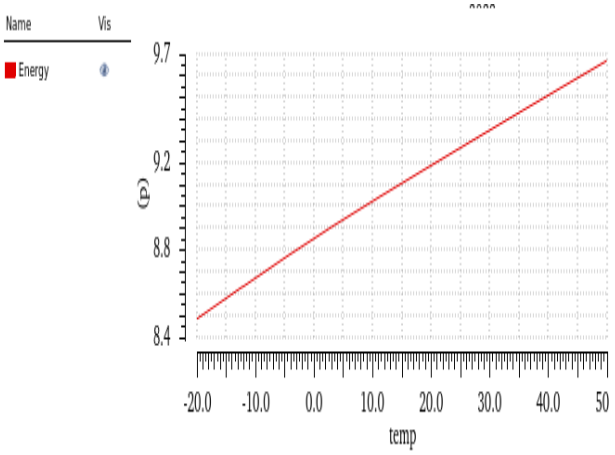


Fig. 9. Temperature-dependent Energy Consumption of Proposed 4-bit VM

TABLE V
IMPACT OF TEMPERATURE VARIATION ON THE ENERGY CONSUMPTION OF 4-BIT VM

Temperature (°C)	Energy (pJ)	
	Traditional 4-bit VM	Proposed 4-bit VM
-20	11.83	8.48
-10	12.14	8.67
0	12.44	8.85
10	12.67	9.02
20	12.93	9.18
30	13.17	9.35
40	13.41	9.51
50	13.61	9.66
Average	12.78	9.09

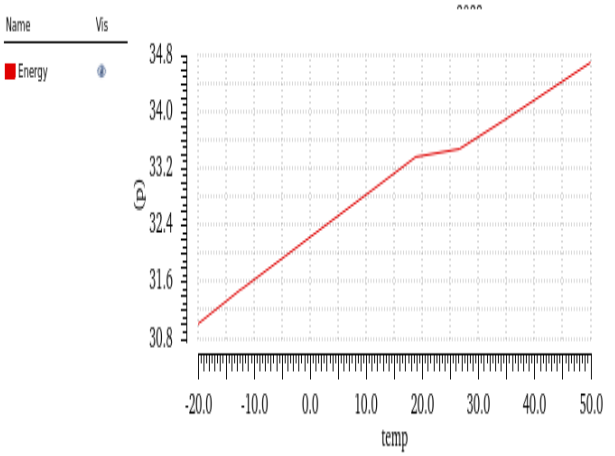


Fig. 10. Temperature-dependent Energy Consumption of Traditional 8-bit VM

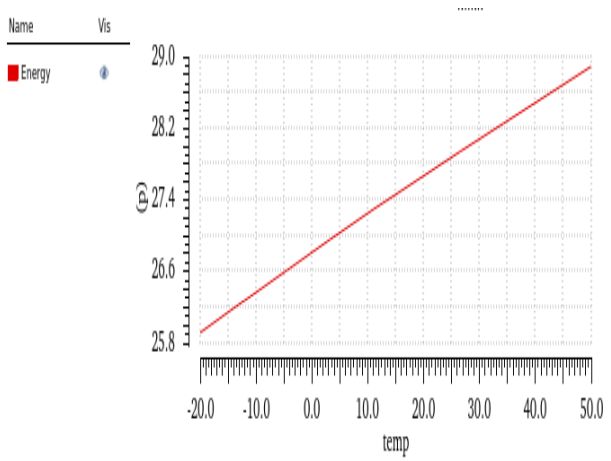


Fig. 11. Temperature-dependent Energy Consumption of Proposed 8-bit VM

TABLE VI
IMPACT OF TEMPERATURE VARIATION ON THE ENERGY CONSUMPTION OF 8-BIT VM

Temperature (°C)	Energy (pJ)	
	Traditional 8-bit VM	Proposed 8-bit VM
-20	31.01	25.91
-10	31.63	26.36
0	32.25	26.85
10	32.83	27.24
20	33.39	27.68
30	33.66	28.07
40	34.17	28.48
50	34.69	28.88
Average	32.95	27.43

shows a substantial decrease in energy consumption compared to the traditional design. Specifically, the average energy consumption of the proposed 4-bit VM is 28.8% lower than that of the traditional design, while the 8-bit VM architecture achieves a reduction of 16.7%. Moreover, the graph depicting the energy consumption of the traditional 8-bit VM shows fluctuations with increasing temperature, while the graph for the proposed 8-bit VM shows a linear and stable trend. This remarkable reduction in energy consumption combined with the stability of the proposed designs, highlights the efficiency and effectiveness of the proposed architecture.

V. CONCLUSION

In conclusion, this manuscript presents an extension to a previous paper [7], highlighting a novel 4-bit Vedic multiplier architecture and its subsequent application in an 8-bit multiplier design. The proposed architecture incorporates a 5-bit special adder, FAs, and HAs, intelligently combined to optimize performance and minimize power consumption. The implementation of the 8-bit VM design using the proposed 4-bit multiplier modules, along with Ripple Carry Adders (RCAs), further highlights the scalability and applicability of the proposed architecture. The architecture exhibits significant improvements in delay, EDP, and PDP compared to traditional

multipliers as presented in Tables III and IV. The analysis of temperature-dependent energy efficiency adds further value to the research. The results indicate the potential of the proposed VM to contribute to energy savings and improved performance in various applications. By leveraging the principles of Vedic mathematics and incorporating intelligent design techniques, the proposed multiplier offers a promising solution for reducing energy consumption and enhancing overall circuit efficiency. These findings contribute to the field of circuit design and offer potential advancements in low-power, high-speed multiplication operations.

REFERENCES

[1] P. Paz and M. Garrido, "Efficient implementation of complex multipliers on fpgas using dsp slices," *Journal of Signal Processing Systems*, pp. 1–8, 2023.

[2] N. Gaur, P. Kumar, and A. Mehra, "Design and analysis of high performance and low power fft for dsp datapath using vedic multipliers," *International Journal of Electronics Letters*, vol. 10, no. 2, pp. 188–199, 2022.

[3] V. J. Ramya, A. Selvarani, L. A. Kumar, and J. Navarajan, "Implementation of high speed fft using reversible logic gates for wireless dsp applications," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 6, pp. 8071–8082, 2021.

[4] B. Venkateshwarlu, S. K. Reddy, and G. S. Kumar, "A novel programmable 16 bit alu using vedic multiplier and kogge-stone adder," *International Journal of Management, IT and Engineering*, vol. 9, no. 1, pp. 386–390, 2019.

[5] S. B. K. Tirtha and V. S. Agrawala, *Vedic mathematics*, vol. 10. Motilal Banarsidass Publ., 1992.

[6] V. Kaushik and H. Saini, "A review on comparative performance analysis of different digital multipliers," *Advances in Computational Sciences and Technology*, vol. 10, no. 5, pp. 1257–1272, 2017.

[7] H. Chugh and S. Singh, "Design and implementation of a high-performance 4-bit vedic multiplier using a novel 5-bit adder in 90nm technology," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pp. 1–6, IEEE, 2022.

[8] K. Ramya and S. K. S. Manvi, "Design of a 4-bit vedic multiplier using transistors," in *Proc. 9th International Conference on Electrical Electronics and Computer Science (EECS 2014)*, pp. 39–44, 2014.

[9] C. R. Patel, V. Urankar, B. Vivek, and V. K. Bharadwaj, "Vedic multiplier in 45nm technology," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 21–26, IEEE, 2020.

[10] R. Yadav, Y. Kumar, K. K. Verma, D. Joshi, H. Yadav, and A. Dev, "Review on fir filters using different adders and multipliers based on vedic mathematics," in *2021 International Conference on Simulation, Automation & Smart Manufacturing (SASM)*, pp. 1–5, IEEE, 2021.

View publication stats

General Relativity and Gravitation (2023) 55:133
https://doi.org/10.1007/s10714-023-03181-w

RESEARCH

Parametrization of the deceleration parameter in a flat FLRW universe: constraints and comparative analysis with the ΛCDM paradigm

Himanshu Chaudhary[1,2,3] · Saadia Mumtaz[4] · Amine Bouali[5] · Ujjal Debnath[6] · G. Mustafa[7,8]

Received: 19 July 2023 / Accepted: 31 October 2023 / Published online: 17 November 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

The constraint of the deceleration parameter associated with dark energy stands as one of the most captivating subjects in the present cosmological framework. This study centers on the parametric reconstruction of the deceleration parameter in a flat Friedmann–Robertson–Walker (FLRW) Universe that encompasses radiation, dark energy, and pressure-less dark matter. In this context, we thoroughly investigate a highly motivated parametrization of $q(z)$, which offers an evolutionary scenario from deceleration to the acceleration phase of the Universe. The crucial task of estimating the parametrization of the Hubble parameter is accomplished through its incorporation into the Friedmann equation. The free parameters are subsequently constrained utilizing a comprehensive set of observational data, including $H(z)$, type Ia supernovae (SNIa), Baryon Acoustic Oscillation (BAO), Gamma Ray Burst (GRB), and Quasar (Q) measurements. Implementing the Markov Chain Monte Carlo (MCMC) technique and the $H(z)$ + BAO + SNIa + GRB + Q dataset, we derive the best-fit values for the model parameters. Consequently, we provide a graphical analysis of the cosmographic parameters such as deceleration, jerk, and snap parameters by applying these optimized model parameter values. Finally, we compare our results with those of the standard ΛCDM paradigm to evaluate the viability of our proposed models.

Keywords Parametrization · FLRW Universe · Friedmann equation

Contents

1 Introduction 2
2 Basic equations of the FLRW model 4
2.1 The model 6

✉ G. Mustafa
gmustafa3828@gmail.com

Extended author information available on the last page of the article

3 Data analysis . . . 6
3.1 Methodology . . . 7
3.2 Data description . . . 7
3.2.1 $H(z)$ dataset . . . 7
3.2.2 Type Ia supernovae (SNIa) . . . 8
3.2.3 Baryon acoustic oscillations (BAO) . . . 10
4 Observational and theoretical comparisons with $H(z)$ measurements and distance modulus $\mu(z)$ . . . 11
4.1 Contrasting with hubble measurements . . . 11
4.2 Contrasting with type Ia supernovae (SNIa) . . . 12
4.3 Relative differences between model and ΛCDM . . . 13
5 Cosmographic analysis . . . 14
5.1 The deceleration parameter . . . 14
5.2 The Jerk parameter . . . 15
6 Statefinder diagnostic . . . 17
7 $Om(z)$ diagnostic . . . 18
8 Information criteria . . . 19
9 Results . . . 20
10 Discussions and conclusions . . . 22
References . . . 23

1 Introduction

It is evident through numerous astronomical studies that our Universe is presently undergoing an expansion epoch. This remarkable phenomenon opens up a new and unprecedented window for pioneering research in the domain of modern cosmology [1–3]. Delving into the underlying physical mechanisms responsible for this cosmic expansion has initiated a captivating journey for the scientific community. To this end, various cosmological models have been proposed to clarify this puzzle, primarily through two distinct approaches. The initial approach involves modifying the gravitational sector of the field equations [4, 5], while the latter focuses on unconventional matter sources, notably dark energy (DE) models, among others. Our emphasis is placed on the second approach, assuming that DE serves as the primary driving force behind the accelerated cosmic expansion [6–8]. The violation of the strong energy conditions by DE leads to anti-gravitational effects, initiating the transition from an early deceleration phase to late-time acceleration [9]. Various DE models have been put forward so far to interpret the fundamental principles driving the ongoing accelerated cosmic expansion [10–14]. Amidst this array of models, the ΛCDM model stands as the widely accepted DE paradigm. Nonetheless, it is associated with certain ambiguities, including the enigmatic *coincidence*, *fine-tuning*, and *age* dilemma [13, 15, 16]. To overcome these challenges, the quest for alternative models assumes significant importance, aiming to provide a more comprehensive portrayal of the current cosmic phases. Various DE models, such as phantom [17–21], quintessence [22] tachyon, k-essence, and scalar field models have played pivotal roles in elucidating the origin and ultimate fate of DE. However, the enigma of cosmic accelerated expansion continues to puzzle the frontiers of modern physics.

The selection of an appropriate approach clarifying the phenomenon of cosmic accelerated expansion has remained a captivating subject for the scientific community. At the phenomenological level, the parametrization of the equation of state (EoS)

parameter has attracted considerable attention in investigating the accelerating Universe. This approach entails assuming a specific evolutionary mechanism rather than adopting a particular DE model and scrutinizing the intrinsic nature of the exotic constituents driving this acceleration. Being a model-independent approach, it involves estimating model parameters using empirical datasets. However, certain limitations arise from these parametric assumptions including (i) the tendency for divergence issues in many parametrizations and (ii) potential oversight of subtle aspects related to DE due to the implicit assumptions underlying the parametric framework. Extensive literature exists offering insights into cosmic acceleration through viable EoS parametrization (see Refs. [23–25]). Drawing inspiration from this, the parametrization of the deceleration parameter q has also received significant attention, as it plays a pivotal role in characterizing the evolutionary mechanism. In any cosmological model, a transition from deceleration to the acceleration phase unfolds, summarizing the entire cosmic evolutionary process. The sign of q distinguishes whether the cosmic expansion is accelerating ($q < 0$) or decelerating ($q > 0$).

In recent times, numerous theoretical frameworks have emerged to comprehensively investigate the evolutionary history of the Universe by employing the parametrization of $q(z)$ as a function depending on the scale factor ($a(t)$), time (t), or redshift (z) [26–51]. By adopting the parametrization of the deceleration parameter, finite outcomes can be achieved without necessitating the selection of a specific gravitational theory. However, similar to the drawbacks associated with the parametrization of EOS, these approaches may also come with certain drawbacks. Due to the lack of conclusive arguments regarding the evolutionary history of the Universe, the adoption of a parametric formalism may be regarded as a preliminary step towards comprehending the expanding nature of the cosmos.

In this manuscript, our key goal is to explore the characteristics of these two kinematic parameters concerning the parametrization of $q(z)$. In a recent study, Manon et al. [52] delved into a particular form of the deceleration parameter and identified the optimal ranges using a χ^2 minimization approach, utilizing available observational data. They also examined the evolutionary scenario of the jerk parameter within the parametrized model. Gadbali et al. [38], on the other hand, explored the parametrization of the deceleration parameter in the context of $f(Q)$ gravity and employed Bayesian analysis to constrain the model parameters with observational data. Inspired by their approaches, we aim to explore the Barboza-Alcaniz-Zhu-Silva (BAZS) type parametrization of the deceleration parameter, striving to identify the most viable model when compared to the ΛCDM model [53–56]. This work serves as an extension of our previous research efforts. Notably, a recent publication has presented a three-parameter model of the deceleration parameter [39] for which we adopted a well-grounded BAZS-type parametrization for the deceleration parameter, which encompassed three unknown parameters.

The primary focus of this study revolves around the precise resolution of the model parameters through diverse observational datasets. Specifically, we employ the $H(z)$ dataset, comprising 57 measurements, and the SNIa dataset, which encompasses 1048 measurements, 17 BAO measurements, 162 measurements of GRBs, and 4 observations of compact radio quasars to derive the optimal values of the model parameters. To accomplish this, we employ the Markov Chain Monte Carlo analysis, which effectively

aligns the model parameters with the most recent observational data. Our analysis encompasses the utilization of the χ^2-minimization scheme, incorporating the $H(z)$ + SNIa + BAO + GRB dataset, as well as Quasar's dataset. Through this comprehensive analysis, we can establish the parameter constraints of q_0, q_1, and β within the confidence levels of 1σ and 2σ. The format of the paper is as follows: Sect. 2 provides an in-depth discussion of the fundamental equations governing the FLRW Universe. In Sect. 2.1, we adopt a parametrization approach for the deceleration parameter of the model and determine the corresponding Hubble solution for each parametrization. Section 3 is dedicated to constraining the model parameters using various datasets while simultaneously deriving the best-fitted values of the model parameters through the MCMC technique consistent with observational data. In Sect. 5, we present graphical analyses of cosmographic parameters, such as deceleration and jerk, for the considered model. Furthermore, Sects. 6 and 7 provide a detailed analysis of the Statefinder and $Om(z)$ Diagnostic parameters. In Sect. 8, informative criteria for the model have been presented. Lastly, Sect. 10 encompasses our concluding remarks.

2 Basic equations of the FLRW model

In this section, we consider a homogeneous, isotropic FLRW Universe with a spatially flat geometry. The line element is given by

$$ds^2 = -dt^2 + a^2(t)\left[dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right], \tag{1}$$

where $a(t)$ represents the scale factor. The energy-momentum tensor for the cosmic fluid is described by

$$T_{\mu\nu} = (\rho + p)u_\mu u_\nu + p g_{\mu\nu}, \tag{2}$$

where ρ and p denote the energy density and pressure density of the fluid, respectively. The 4-velocity $u^\mu = \frac{dx^\mu}{ds}$ satisfies the relation $u^\mu u_\mu = -1$. The Friedmann equations in the Einstein's gravity can be written as

$$H^2 = \frac{8\pi G}{3}\rho, \tag{3}$$

and

$$\dot{H} = -4\pi G(\rho + p), \tag{4}$$

where $H = \frac{\dot{a}}{a}$ represents the Hubble parameter whereas the dot symbolizes derivative with respect to cosmic time t. Considering that the Universe is composed of radiation, dark matter (DM), and DE, the total energy density ρ and total pressure p can be expressed as $\rho = \rho_r + \rho_m + \rho_d$ and $p = p_r + p_m + p_d$, respectively. Assuming

separate conservation equations for radiation, DM, and DE, we have

$$\dot{\rho}_r + 3H(\rho_r + p_r) = 0, \tag{5}$$
$$\dot{\rho}_m + 3H(\rho_m + p_m) = 0, \tag{6}$$

and

$$\dot{\rho}_d + 3H(\rho_d + p_d) = 0. \tag{7}$$

For radiation, we have $p_r = \frac{1}{3}\rho_r$, leading to $\rho_r = \rho_{r0}a^{-4}$ from Eq. (5). Since DM is pressureless, i.e., $p_m = 0$, Eq. (6) yields $\rho_m = \rho_{m0}a^{-3}$.

Now let us introduce the deceleration parameter as follows

$$q = -1 - \frac{\dot{H}}{H^2}. \tag{8}$$

Similarly, we define the deceleration parameter for DE as $q_d = -1 - \frac{\dot{H}_d}{H_d^2}$, where H_d represents the Hubble expansion rate for the DE component. Applying Eqs. (3) and (4), we can express

$$H_d^2 = \frac{8\pi G}{3}\,\rho_d, \tag{9}$$

and

$$\dot{H}_d = -4\pi G(\rho_d + p_d). \tag{10}$$

Using the field Eqs. (9), (10), and the energy-conservation Eq. (7), we can find the energy density of the fluid as

$$\rho_d = \rho_{d0}\, e^{\int \frac{2(1+q_d)}{1+z}dz}, \tag{11}$$

where ρ_{d0} represents the present value of the density parameter and z is the redshift parameter given by $1 + z = \frac{1}{a}$ (with the present scale factor $a_0 = 1$). We define the dimensionless parameters $\Omega_{r0} = \frac{8\pi G\rho_{r0}}{3H_0^2}$, $\Omega_{m0} = \frac{8\pi G\rho_{m0}}{3H_0^2}$, and $\Omega_{d0} = \frac{8\pi G\rho_{d0}}{3H_0^2}$, where H_0 is the present value of the Hubble parameter. Using Eq. (3), the Hubble parameter can be expressed as

$$H^2(z) = H_0^2\left[\Omega_{r0}(1+z)^4 + \Omega_{m0}(1+z)^3 + \Omega_{d0}\, e^{\int \frac{2(1+q_d)}{1+z}dz}\right], \tag{12}$$

where $\Omega_{d0} = 1 - \Omega_{r0} - \Omega_{m0}$.

Consequently, this section establishes the basic equations of the spatially flat FLRW model. Considering a Universe composed of radiation, DM, and DM, we derive the Friedmann equations, conservation equations, and the deceleration parameter. The

energy densities and pressures for each component are calculated, with specific attention given to the DE component. The resulting equations provide a foundation for investigating the dynamics and evolution of the Universe in subsequent sections of the manuscript.

2.1 The model

In this study, we adopt the BAZS type model parameterization [57] for the deceleration parameter [58, 59] given by

$$q_d(z) = q_0 + \frac{q_1}{(1+z)^\beta}, \tag{13}$$

where q_0 and q_1 are constants, and β is considered to be a positive constant. The corresponding energy density parameter (11) can be expressed as

$$\rho_d = \rho_{d0}\,(1+z)^{2(1+q_0)}\; e^{-\frac{2q_1}{\beta(1+z)^\beta}}. \tag{14}$$

Using Eq. (12), we obtain the Hubble parameter as

$$H^2(z) = H_0^2\Big[\Omega_{r0}(1+z)^4 + \Omega_{m0}(1+z)^3 + (1-\Omega_{r0}-\Omega_{m0})\,(1+z)^{2(1+q_0)}\; e^{-\frac{2q_1}{\beta(1+z)^\beta}}\Big]. \tag{15}$$

The chosen BAZS parametrization enables the exploration of the deceleration parameter's pattern in terms of the model parameters q_0, q_1, and β. It provides a flexible and versatile approach to characterize the acceleration of the Universe, capturing both early deceleration and late-time acceleration. Using this parametrization, we can explore the evolution of the Hubble parameter and study the dynamics of the Universe in different cosmic epochs.

3 Data analysis

To constrain the model parameters, we intend to perform a comprehensive analysis using multiple observational datasets, including $H(z)$ measurements, SNIa, GRB, quasars, and BAO measurements. This aims to find the best-fit values for the proposed model by employing the MCMC technique [60]. The total χ^2 function, denoted by χ^2_{tot}, incorporates contributions from different datasets, namely $\chi^2_{H(z)}$ for $H(z)$, χ^2_{SNIa} for SNIa, χ^2_{BAO} for BAO, χ^2_{GRB} for GRB, and χ^2_Q for quasars. This function quantifies the agreement between the model predictions and the corresponding observational data. By minimizing the total χ^2 function through the MCMC analysis, the model parameters can be effectively constrained. The MCMC algorithm explores the parameter space, generating a Markov chain of parameter values that converge to the regions of

the highest probability. The best-fit values are obtained based on the parameter combinations that minimize the total χ^2 value, indicating the optimal agreement between the model and the observational datasets. This comprehensive approach, combining multiple datasets and utilizing MCMC, allows a robust and thorough characterization of the proposed model. It takes a wide range of cosmological observations into account, ensuring a comprehensive analysis and providing more reliable constraints on the model parameters.

3.1 Methodology

In the realm of cosmology, MCMC methods have emerged as invaluable tools for exploring and analyzing intricate parameter spaces, especially in the context of Bayesian inference [61, 62]. The MCMC techniques provide a powerful means to estimate model parameters and ascertain their uncertainties through efficient sampling from probability distributions [63]. By implementing algorithms like the Metropolis-Hastings and Gibbs sampling, MCMC generates a Markov chain of parameter values that converge to the desired target distribution [64, 65]. This target distribution typically represents the posterior distribution, combining prior knowledge with the observational data [66]. Through iterative proposals and acceptance or rejection steps based on the likelihood of the data and prior information, MCMC systematically explores the parameter space, sampling regions proportionate to their relative probability [61]. This enables reliable estimation of best-fit parameters and associated uncertainties [66]. MCMC plays a significant role in constraining the parameters of theoretical models, including those concerning the cosmic microwave background, large-scale structure, or DE [62, 67]. It allows researchers to quantitatively evaluate the compatibility of various models with observational data [68]. MCMC facilitates the examination of parameter degeneracies and correlations, aiding in the identification of interdependencies among cosmological parameters [61]. Such insights are critical for understanding the underlying cosmological model and the physical processes [68]. MCMC also offers a flexible framework in incorporating prior knowledge, enabling the integration of theoretical expectations and previous observational constraints into the analysis [66].

3.2 Data description

3.2.1 $H(z)$ dataset

In this analysis, $H(z)$ data is utilized to constrain the parameters of the model. In particular, Hubble data may be determined by calculating the BAO in the radial direction of galaxy clusters [69] or by the differential age technique, which additionally yields the redshift dependence of the Hubble parameter as

$$H(z) = -\frac{1}{1+z}\frac{dz}{dt}, \tag{16}$$

where dz/dt is determined appropriately by using two moving galaxies in a proportionate manner. To compute the model's parameters, the analysis incorporates 57 Hubble measurements, spanning the redshift range from $0.07 \leq z \leq 2.42$. The chi-square function is used to correlate the theoretical predictions of the model against observations.

$$\chi^2_{H(z)}(\phi) = \sum_{j=1}^{57} \frac{\left[H_{th}\left(z_j, \phi\right) - H_{obs}\left(z_j\right)\right]^2}{\sigma^2_{H(z_j)}}, \tag{17}$$

where $H_{th}\left(z_j, \phi\right)$, $H_{obs}(z_j)$ and $\sigma_{H(z_j)}$ denote the theoretical Hubble function values at redshift z_j with model parameters ϕ, observed value of Hubble function at z_j, and standard error at the redshift z_j, respectively. The numerical values of the Hubble function with appropriate redshifts are shown in Table 1.

3.2.2 Type Ia supernovae (SNIa)

Our primary dataset consists of two essential components. The initial component comprises the Pantheon compilation, which encompasses 1048 measurements of Type Ia Supernovae (SNIa). These measurements provide detailed information on relative luminosity distances. These measurements span the extensive redshift range of $0.01 < z < 2.3$ [88]. This publicly accessible SNIa catalog offers magnitudes that have been meticulously corrected for systematic effects. These corrections account for various factors, such as the stretch of the light curve, the color at maximum brightness, and the stellar mass of the host galaxy. These corrections are essential to ensure the accuracy of the observed data. However, it's important to note that the apparent magnitude of each SNIa must be calibrated using an arbitrary fiducial absolute magnitude, denoted as M. In our analyses, we treat M as a nuisance parameter, which allows us to account for uncertainties in the calibration process. This calibration is achieved by leveraging the theoretical values of the distance moduli, providing a means to link observed apparent magnitudes to their corresponding absolute magnitudes.

$$\mu\left(z_j, \phi\right) = 5\log_{10}\left[D_L\left(z_j, \phi\right)\right] + M \tag{18}$$

at redshift z_j by calculating the corresponding luminosity distance.

$$D_L\left(z_j, \Theta\right) = c\left(1 + z_j\right)\int_0^{z_j} \frac{\mathrm{d}z'}{H\left(z', \phi\right)} \tag{19}$$

In the above equation, the variable c represents the speed of light, while the nuisance parameter M encodes the Hubble constant. This parameter needs to be marginalized in our analyses to account for its uncertainty. The associated expression for χ^2_{SN} is outlined in [89].

$$\chi^2_{\mathrm{SN}}(\phi) = \left(\Delta\mu\left(z_i, \phi\right)\right)^T C^{-1}_{\mathrm{SN}} \Delta\mu\left(z_j, \phi\right) + \ln\left(\frac{S}{2\pi}\right) - \frac{k^2(\phi)}{S} \tag{20}$$

Table 1 In this table, we present a compilation of 57 data points utilized in this manuscript

z	$H(z)$	σ_H	References	z	$H(z)$	σ_H	References
0.070	69	19.6	[70]	1.363	160	33.6	[71]
0.120	68.6	26.2	[70]	1.965	186.5	50.4	[71]
0.480	97	62	[70]	0.24	79.69	2.99	[72]
0.880	90	40	[70]	0.34	83.8	3.66	[72]
0.170	83	8	[73]	0.43	86.45	3.97	[72]
0.270	77	14	[73]	0.30	81.7	6.22	[74]
0.400	95	17	[73]	0.31	78.18	4.74	[75]
0.900	117	23	[73]	0.36	79.94	3.38	[75]
0.90	69	12	[73]	0.40	82.04	2.03	[75]
1.300	168	17	[73]	0.44	84.81	1.83	[75]
1.430	177	18	[73]	0.48	87.79	2.03	[75]
1.530	140	14	[73]	0.52	94.35	2.64	[75]
1.750	202	40	[73]	0.56	93.34	2.3	[75]
0.1791	75	4	[76]	0.59	98.48	3.18	[75]
0.1993	75	5	[76]	0.64	98.82	2.98	[75]
0.3519	83	14	[76]	0.35	82.7	9.1	[77]
0.593	104	13	[76]	0.38	81.5	1.9	[78]
0.6797	92	8	[76]	0.51	90.4	1.9	[78]
0.7812	105	12	[76]	0.61	97.3	2.1	[78]
0.8754	125	17	[76]	0.44	82.6	7.8	[79]
1.037	154	20	[76]	0.60	87.9	6.1	[79]
0.200	72.9	29.6	[80]	0.73	97.3	7.0	[79]
0.280	88.8	36.6	[80]	0.57	87.6	7.8	[81]
0.3802	83	13.5	[82]	0.57	96.8	3.4	[83]
0.4004	77	10.2	[82]	2.30	224	8.6	[84]
0.4247	87.1	11.2	[82]	2.33	224	8	[85]
0.4497	92.8	12.9	[82]	2.34	222	8.5	[84]
0.4783	80	99	[82]	2.36	226	9.3	[86]
0.470	89	34	[87]				

Among these, 31 data points are derived through the differential age technique, while the remaining 26 data points are obtained using Baryon Acoustic Oscillations (BAO) and various other methods. These data points span the redshift range of $0.07 \leq z \leq 2.42$ and include their corresponding errors denoted as σ_H

In this context, $C_{\rm SN}$ represents the overall covariance matrix. We calculate the sum of all the components within the inverse of C_{SN} as denoted by S". The variable k is determined as follows

$$k(\phi) = \left(\Delta\mu\left(z_j, \phi\right)\right)^T \cdot C_{\rm SN}^{-1} \tag{21}$$

with $\Delta\mu\left(z_j, \phi\right) = \mu\left(z_j, \phi\right) - \mu_{\rm obs}\left(z_j\right)$.

In our analysis, we include a sample of 162 GRBs as described in [90]. GRBs are remarkable cosmic events known for their extreme luminosity, making them visible even from the furthest reaches of the observable Universe. Their intense brightness enables us to study objects at unprecedented distances and, consequently, they are invaluable probes for cosmological investigations. We also incorporate 24 binned quasar distance modulus data from [91]. Quasars, often referred to as quasi-stellar radio sources, are exceptionally bright and energetic nuclei of distant galaxies, hosting supermassive black holes. They are observable in a range of redshifts from $0.46 \leq z \leq 2.76$. Quasars are crucial in our MCMC analysis as they provide unique insights into the large-scale structure of the Universe and can help us to understand the distribution and evolution of matter over cosmic time. The quasar data, obtained from a Very Long Baseline Interferometry (VLBI) all-sky survey, offers high-resolution observations of these powerful cosmic beacons, enhancing the precision of our cosmological investigations. Both GRBs and quasars play pivotal roles in our MCMC analysis by serving as powerful tools for studying the cosmos at extreme distances and shedding light on the underlying structure and evolution of the Universe.

3.2.3 Baryon acoustic oscillations (BAO)

Baryon acoustic oscillations are intriguing features in the large-scale structure of the Universe. These patterns originate from the sound waves in the early Universe's baryon-photon fluid and are imprinted during recombination. BAO's are detected through surveys, often expressed as $D_A(z)/r_d$, $D_V(z)/r_d$, and $H(z) \cdot r_d$, where r_d representing the comoving sound horizon at recombination ($z_d \approx 1059$). This standard ruler helps in measuring the cosmic expansion independently, by combining BAO data with low-redshift observations, offering insights beyond the cosmic microwave background.

The dataset utilized in our analysis is a subset of a more extensive catalog comprising 333 measurements of Baryon Acoustic Oscillations (BAO). The primary sources contributing to this catalog is [92–103]. Given the high degree of correlation within this extensive dataset, we have opted to work with a subsample of BAO measurements extracted from the larger catalog, as detailed in [104]. To address potential correlations in the BAO dataset, mock datasets from N-body simulations with known parameters are used to estimate covariance matrices. Since exact covariance details are unavailable, we apply a method from [105] introducing non-diagonal elements to account for the correlations. This method randomly adds positive correlations between data points, covering over 66.6% of the dataset. Non-diagonal elements are set to $C_{ij} = 0.5\sigma_i\sigma_j$, where σ_i and σ_j are the 1 σ errors for data points i and j," respectively, hence, a representative subset to minimize errors. Transverse BAO experiments produce measurements of $D_H(z)/r_d = c/H(z)r_d$ with a comoving angular diameter distance [106, 107].

$$D_M = \frac{c}{H_0} S_k \left(\int_0^z \frac{dz'}{E(z')} \right), \tag{22}$$

where

$$S_k(x) = \begin{cases} \frac{1}{\sqrt{\Omega_k}} \sinh\left(\sqrt{\Omega_k} x\right) & \text{if} \quad \Omega_k > 0 \\ x & \text{if} \quad \Omega_k = 0 \\ \frac{1}{\sqrt{-\Omega_k}} \sin\left(\sqrt{-\Omega_k} x\right) & \text{if} \quad \Omega_k < 0. \end{cases} \tag{23}$$

We also consider the angular diameter distance $D_A = D_M/(1+z)$ and $D_V(z)/r_d$. This corresponds to the combination of the BAO peaked coordinates and the sound horizon r_d at the drag epoch. Furthermore, we could immediately derive "line-of-sight" (or "radial") observations from the Hubble parameter (Fig. 1).

$$D_V(z) \equiv \left[z D_H(z) D_M^2(z) \right]^{1/3}. \tag{24}$$

4 Observational and theoretical comparisons with $H(z)$ measurements and distance modulus $\mu(z)$

After finding the values of free parameters in our cosmological model, it is essential to compare the model's predictions with observational data and the well-established ΛCDM paradigm. This step allows us to assess the model's viability and its ability to explain the observed phenomena in the Universe. By contrasting the model predictions with observational data, we can evaluate the agreement and identify any discrepancies or deviations from the standard ΛCDM framework. This analysis plays a crucial role in validating or refining the proposed cosmological model and advancing our understanding of the underlying physical processes driving the evolution of the Universe.

4.1 Contrasting with hubble measurements

We assess the compatibility of our model with observational data by comparing it to the Hubble 57 datasets, represented by pink dots accompanied by error bars. We also include the well-established ΛCDM model as a reference, indicated by a black dotted line. The comparative analysis is presented in Fig. 2. This figure provides important insights into the model's performance in describing the observed Hubble data. Our analysis reveals that our proposed cosmological model closely matches the behavior of the standard ΛCDM model for redshifts $z < 1$. However, as we extend our observations to redshifts in the range $z > 1$, subtle deviations emerge between our model and the standard ΛCDM paradigm. These deviations diminish rapidly beyond a critical redshift of approximately $z \approx 2.5$. This observation indicates that the proposed cosmological model exhibits distinctive characteristics at higher redshifts, setting it apart from the conventional ΛCDM model. These deviations have significant cosmological implications. They suggest that our model may better account for the behavior of the universe at early cosmic times, where redshifts are higher. However,

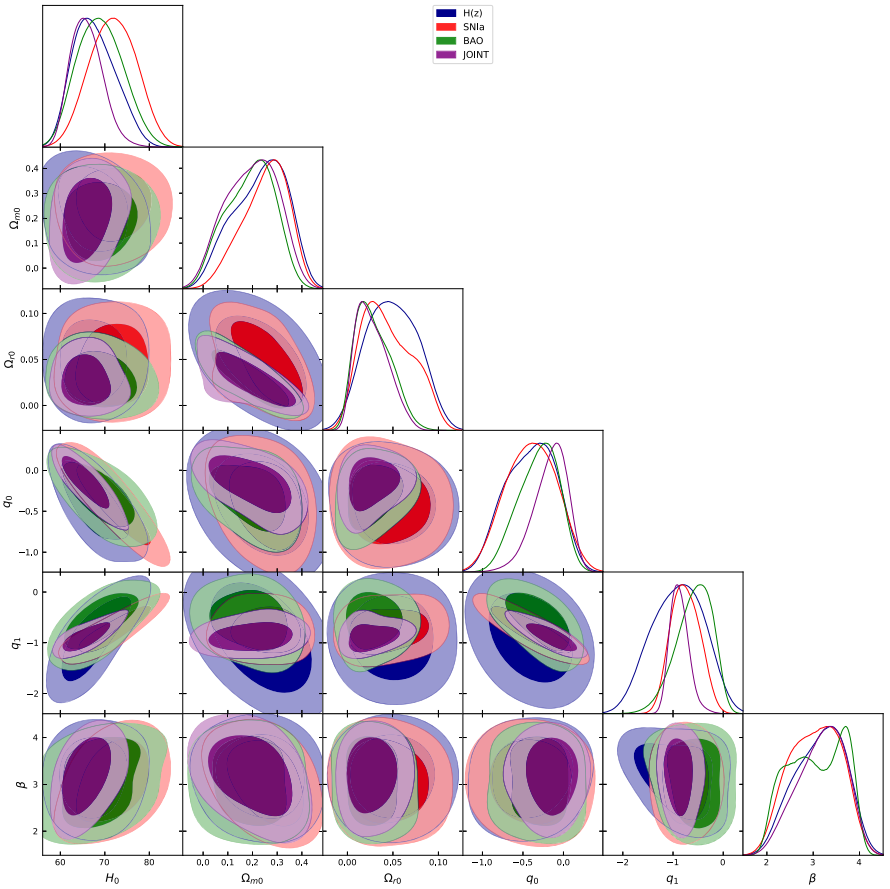


Fig. 1 The figure depicts the posterior distribution for various observational data measurements using the BAZS type Model, illustrating the 1σ and 2σ confidence levels. Specifically, "H(z)" corresponds to the Hubble dataset sourced from Table 1," "SNIa" refers to the type Ia supernovae dataset, and "BAO" represents the baryon acoustic oscillations dataset obtained from [104]. The "JOINT" dataset encompasses a combination of the "H(z)" dataset, type Ia supernovae, gamma-ray bursts, quasars, and BAO measurements

it is crucial to note that these deviations gradually decrease, indicating a convergence between the two models beyond $z \approx 2.5$.

4.2 Contrasting with type Ia supernovae (SNIa)

We have compared the distance modulus function $\mu(z)$ of our model with the measurements obtained from SNIa observations. The SNIa dataset consists of 1048 observational points (denoted by pink dots in the figure) accompanied by error bars to indicate the measurement uncertainties. For comparison, we also include the ΛCDM model, represented by a black dotted line. The distance modulus function $\mu(z)$ is a crucial quantity in cosmology as it describes the apparent brightness of distant objects,

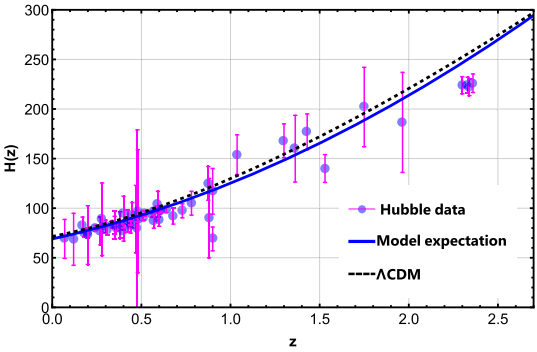


Fig. 2 Theoretical curve of Hubble function $H(z)$ of the model shown in blue line and ΛCDM model shown in black dotted line with $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, against $H(z)$ measurements shown in blue dots with their corresponding error bars in Magenta line

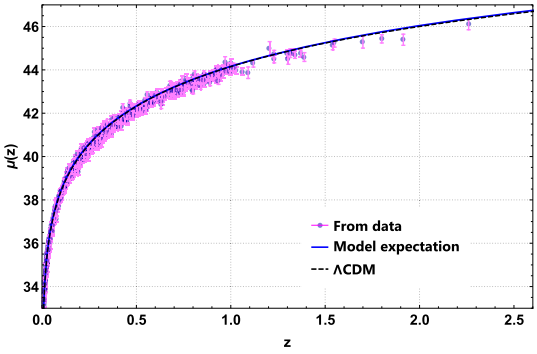


Fig. 3 The theoretical curves of distance modulus $\mu(z)$ of the model shown in blue line and ΛCDM model shown in black dotted line with $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, against $H(z)$ measurements are shown in blue dots with their corresponding error bars in Magenta line

such as supernovae, and allows us to infer their distances. By comparing our model's predictions with the observed SNIa data, we can assess the model's ability to reproduce the observed behavior of the Universe. Figure 3 visually demonstrates the agreement between our model and the SNIa measurements. The fact that our model closely aligns with the dataset suggests that it provides a good fit for the observed supernova data. This agreement supports the viability of our model in describing the cosmological expansion and provides evidence for its ability to explain the behavior of the Universe as captured by the SNIa observations.

4.3 Relative differences between model and ΛCDM

Now we plot Fig. 4 that highlights the relative distinctions between our model and the ΛCDM model. This plot reveals that both models exhibit comparable behaviors for redshifts $z < 1$. However, as the redshift increases, slight discrepancies emerge between the two models. Notably, these differences diminish rapidly beyond a critical redshift of approximately $z \approx 2.5$. The resemblance between our model and the ΛCDM model for $z < 1$ suggests that both models capture the observed cosmic dynamics adequately within this range. The observed deviations for higher redshifts indicate unique features or alternative mechanisms in our model. Nevertheless, it is essential to note that these deviations gradually diminish, implying a convergence between the two models at higher redshifts.

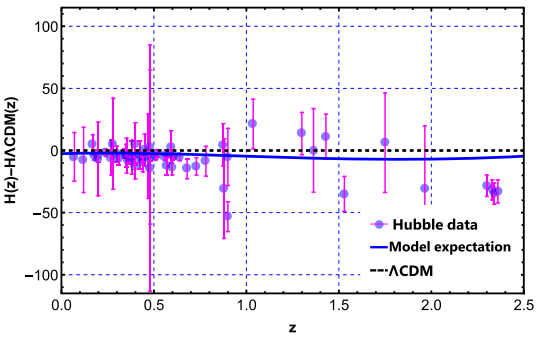


Fig. 4 The variations of the difference between the Model Hubble function shown in the blue line and the ΛCDM Hubble function shown in the black dotted line as a function of the redshift z with $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, against $H(z)$ measurements are shown in blue dots with their corresponding error bars shown in the Magenta line

5 Cosmographic analysis

Cosmographic parameters refer to a set of cosmological quantities that characterize the expansion history of the Universe. These parameters provide valuable insights into the nature of DE and the dynamics of cosmic expansion. The key cosmographic parameters include the Hubble parameter and its derivatives, such as the deceleration parameter ($q(z)$) and the jerk parameter ($j(z)$). These parameters are derived from observational data and theoretical models, allowing us to probe the fundamental properties of the Universe. By studying the behavior of cosmographic parameters at different redshifts, we can constrain and compare various cosmological models, shedding light on the underlying physics driving the evolution of our Universe.

5.1 The deceleration parameter

A dimensionless parameter that governs the cosmological assessment of the expansion rate, is known as the deceleration parameter. This parameter can be defined in terms of the scale factor whose diminishing leads to the uniform expansion of the Universe. The deceleration parameter is a measure of the rate at which the expansion of the Universe is slowing down. If the deceleration parameter is positive, it means that the expansion of the Universe is slowing down due to the gravitational attraction of matter. If the deceleration parameter is negative, it means that the expansion of the Universe is accelerating, which is thought to be due to the influence of DE. Mathematically, it is given by

$$q = -\frac{a\ddot{a}}{\dot{a}^2}. \tag{25}$$

It should be noted that the value of the Hubble parameter either increases or decreases with time according to the sign of the deceleration parameter. The behavior of DP is shown in Fig. 5. Different ranges of q_0 have been predicted so far in certain cosmological scenarios. Ideally, the value of q_0 must be evaluated through observational analysis. For example, the correspondence between apparent brightness and redshift for a class of similar supernovae in distant galaxies relies on the q. Though such estimations seem unfavorable to comprehend, however, the new results entirely support

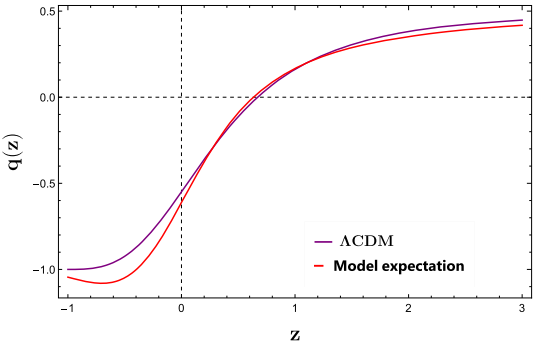


Fig. 5 The figure illustrates the evolution of the deceleration parameter with respect to the redshift (z). It features the ΛCDM model depicted in the purple line, characterized by $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, as well as the BAZS type Model represented by the red line, utilizing the best-fit values obtained from the Joint analysis

the accelerating Universe models. We must choose the value of q_0 very precisely. It turns out to be very model-dependent, and the case for an accelerating Universe is not as strong as usually assumed.

5.2 The Jerk parameter

A constructive and worthwhile diagnostic of DE is required to interpret different cosmological scenarios comprising DE. Sahni et al. [108] gave a new diagnostic of DE using higher derivatives of the scale factor. The dimensionless jerk parameter (also known as jolt refers to the generalization of usual cosmological parameters $a(t)$ and q. Alternative terms include pulse, impulse, bounce, surge, shock, and super-acceleration. The jerk parameter shows how the cosmic acceleration is changing over time. The jerk parameter is defined as the third derivative of the scale factor with respect to time. It is a dimensionless quantity that can be used to describe the deviation of cosmic expansion from a linear function. Mathematically, one could write it as

$$j = \frac{1}{a}\frac{d^3a}{d\tau^3}\left[\frac{1}{a}\frac{da}{d\tau}\right]^{-3} = q(2q+1) + (1+z)\frac{dq}{dz}. \tag{26}$$

This parameter arises from the fourth term in a Taylor series expansion of the scale factor about t_0 given by

$$\begin{aligned}\frac{a(t)}{a_0} &= 1 + H_0(t-t_0) - \frac{1}{2}q_0H_0^2(t-t_0)^2 + \frac{1}{6}j_0H_0^3(t-t_0)^3 \\ &\quad + O\left[(t-t_0)^4\right],\end{aligned} \tag{27}$$

where subscript 0 corresponds to the current values. The diversity of DE proposals tends to create hurdles in finding the most suitable candidate for the physical significance of cosmic dynamics. In this plethora of proposals, a specific value of the respective diagnostic parameter j helps to encompass an association of the DE proposals with the standard Universe models [109, 110]. Transitions between different eras of accelerated expansion can be demonstrated more significantly through the models

Table 2 At a 95% confidence level (CL), we present the constraints on cosmological parameters for both the standard ΛCDM and BAZS models

MCMC Results						
Model	Parameters	Priors	H(z)	SNIa	BAO	JOINT
ΛCDM Model	H_0	[50,100]	$68.601446^{+2.391380}_{-4.796978}$	$74.148786^{+1.214393}_{-2.232916}$	$69.089209^{+4.396874}_{-6.252595}$	$69.854848^{+1.259100}_{-2.386935}$
	Ω_{m0}	[0.,1.]	$0.312846^{+0.035127}_{-0.067481}$	$0.285342^{+0.008887}_{-0.015831}$	$0.256796^{+0.025772}_{-0.069423}$	$0.268654^{+0.012822}_{-0.028134}$
BAZS type Model	H_0	[50,100]	$68.592971^{+5.126186}_{-7.148173}$	$74.873983^{+5.464484}_{-9.640377}$	$69.956978^{+5.616279}_{-8.345873}$	$67.773921^{+3.716017}_{-5.461747}$
	Ω_{m0}	[0.,1.]	$0.232805^{+0.126574}_{-0.213658}$	$0.253057^{+0.102011}_{-0.194818}$	$0.191094^{+0.116464}_{-0.179170}$	$0.197848^{+0.121585}_{-0.185651}$
	Ω_{r0}	[0,0.1]	$0.049321^{+0.030402}_{-0.046350}$	$0.043840^{+0.030156}_{-0.040556}$	$0.028279^{+0.020736}_{-0.027171}$	$0.026929^{+0.018005}_{-0.025413}$
	q_0	[-1.,1.]	$-0.397751^{+0.555549}_{-0.360012}$	$-0.395164^{+0.550275}_{-0.345706}$	$-0.303608^{+0.506796}_{-0.288555}$	$-0.156998^{+0.499696}_{-0.208188}$
	q_1	[-1.,0.]	$-0.899420^{+0.945385}_{-0.594614}$	$-0.754379^{+0.515717}_{-0.282478}$	$-0.573074^{+0.741877}_{-0.360921}$	$-0.875359^{+0.298836}_{-0.184219}$
	β	[2.,4.]	$3.156262^{+1.012974}_{-0.640448}$	$3.052575^{+0.947410}_{-0.648137}$	$3.071269^{+1.010929}_{-0.710796}$	$3.190088^{+1.054048}_{-0.558220}$

These constraints are derived from the Hubble dataset, type Ia supernovae (SNIa), and Baryon Acoustic Oscillations (BAO) datasets. Additionally, the 'JOINT' dataset combines the 'H(z)' dataset with data from type Ia supernovae, gamma-ray bursts, quasars, and BAO measurements

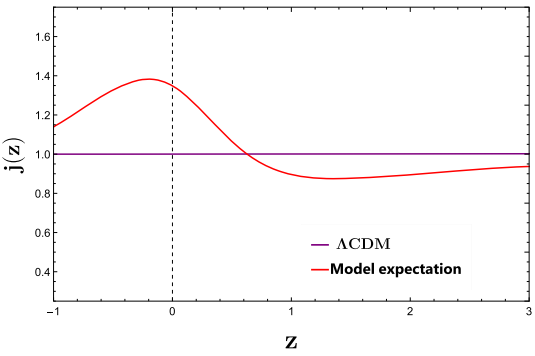


Fig. 6 The figure illustrates the evolution of the jerk parameter with respect to the redshift (z). It features the ΛCDM model depicted in the purple line, characterized by $\Omega_{m0} = 0.3$ and $\Omega_\Lambda = 0.7$, as well as the BAZS type Model represented by the red line, utilizing the best-fit values obtained from the Joint analysis

involving a cosmic jerk. Blandford et al. [111] discussed how the jerk parameterization reports an alternative formalism to classify cosmological models to concordance ΛCDM model. The value $j = 1$ yields the flat ΛCDM model. The generalization of the jerk mechanism thus allows fixing the departure from the ΛCDM value in a conclusive way.

6 Statefinder diagnostic

statefinder diagnostics [108, 112–115, 115, 116] is a formalism generally used to differentiate various DE models and compare their behavior using the higher-order derivatives of the scale factor. The cosmological statefinder diagnostic pair $\{r, s\}$ allows one to explore the cosmic characteristics of DE in a model-independent manner which can be calculated using the relations

$$r = \frac{\dddot{a}}{aH^3}, \qquad s = \frac{r - 1}{3\left(q - \frac{1}{2}\right)}. \tag{28}$$

Here the parameter s comes up as a linear combination of r and q. This diagnostic pair is dimensionless and geometrical as it is computed in terms of the cosmic scale factor. Different trajectories in the $\{r, s\}$ and $\{q, r\}$ planes are displayed to illustrate the time evolution of different DE models with the help of statefinder diagnostics pair. In these circumstances, some specific pairs commonly correspond to standard models of DE like $\{r, s\} = \{1, 0\}$ represents ΛCDM model and $\{r, s\} = \{1, 1\}$ indicates standard cold dark matter model (SCDM) in FLRW background. Also, $(-\infty, \infty)$ yields static Einstein Universe. In the $r - s$ plane, $s > 0$ and $s < 0$ define a quintessence-like model and phantom-like model of the DE, respectively. Moreover, the evolution from phantom to quintessence can be observed by deviation from $r, s = 1, 0$. On the other hand, $\{q, r\} = \{-1, 1\}$ corresponds to the ΛCDM model while $\{q, r\} = \{0.5, 1\}$ shows SCDM model. It is worth noting that in a $r - s$ plane, if the trajectories of the DE model move away from these standard values, then the respective model comprises distinct characteristics from the standard cosmic models.

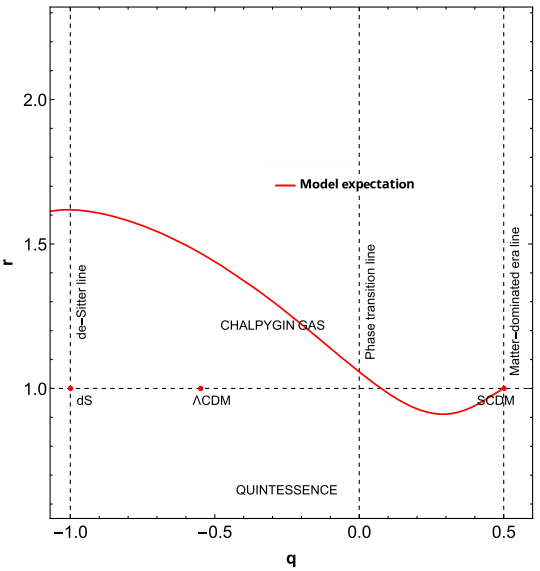


Fig. 7 The figure illustrates the behavior of $\{q, r\}$ profile of the BAZS type Model represented by the red line, utilizing the best-fit values obtained from the Joint analysis

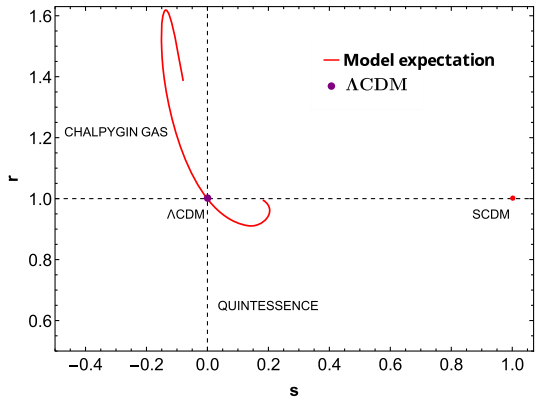


Fig. 8 The figure illustrates the behavior of $\{s, r\}$ profile of the BAZS type Model represented by the red line, utilizing the best-fit values obtained from the Joint analysis

7 *Om*(*z*) diagnostic

The $Om(z)$ diagnostic is a geometrical analysis for which the Hubble parameter is used to provide a null test for the ΛCDM model [117–120]. Likewise the statefinder diagnostic, Om diagnostic also effectively distinguishes various DE models from ΛCDM through the slope variation of $Om(z)$. A positive slope of the diagnostic parameter gives a quintessence model whereas its negative behavior depicts a phantom model. Also, a constant slope with respect to redshift specifies the nature of DE coinciding with that of the cosmological constant. In the case of a flat Universe, $Om(z)$ is defined

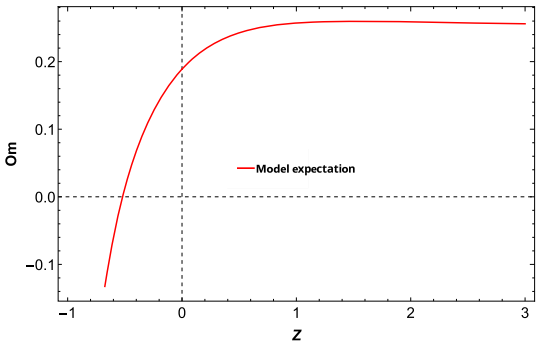


Fig. 9 The figure illustrates the behavior of $Om(z)$ profile of the BAZS type Model represented by the red line, utilizing the best-fit values obtained from the Joint analysis

by

$$Om(z) = \frac{\left(\frac{H(z)}{H_0}\right)^2 - 1}{(1+z)^3 - 1}. \tag{29}$$

In the Om diagnostic, only the first-order temporal derivative is involved, making it a simpler diagnostic than the statefinder diagnosis.

8 Information criteria

Model selection criteria, e.g., the corrected Akaike information criterion (AIC_c) [121–123], play a crucial role in evaluating the sustainability of the models. The AIC_c takes the maximum likelihood function, the number of data, and the number of free parameters of models into account. For larger datasets, the AIC_c reduces to AIC. The AIC is used to figure out the goodness-of-fit of different models. The model having the lowest AIC is the most preferred by observations. The difference, $\Delta AIC = AIC_{preferred} - AIC_{model}$, is more meaningful than the AIC itself. A range of $(0, 2)$ for ΔAIC refers to the models that are strongly supported, values in the range of $(4, 7)$ are less supported, while values beyond 10 refer to models without any support. Therefore, the use of AIC is a very useful tool for classifying a set of models based on their ability to set data. Additionally, we incorporate the Bayesian Information Criterion (BIC), as introduced by [121–123]. The BIC is defined as: $BIC \equiv -2\ln \mathcal{L}_{\max} + p_{\rm tot} \ln(N_{\rm tot})$ .. Similar to the concept of ΔAIC, when $0 < |\Delta BIC| \leq 2$, the model with the higher BIC value is marginally less favored by the data. However, when $2 < |\Delta BIC| \leq 6$ (or $|\Delta BIC| > 6$), the model with the higher BIC value is considered significantly (or highly) less favored by the data. This criterion provides a useful tool for model selection and comparison in our analysis.

Table 3 Summary of $\chi^{2\ min}_{\rm tot}$, $\chi^2_{\rm red}$, AIC_c, ΔAIC_c, BIC, ΔBIC for ΛCDM and BAZS type model

Model	$\chi^{2\ min}_{\rm tot}$	$\chi^2_{\rm red}$	$\mathcal{K}_{\rm f}$	AIC_c	ΔAIC	BIC	ΔBIC
ΛCDM model	1635.56	0.9756	3	1641.56	0	1657.09	0
Model	1630.85	0.9745	6	1642.85	1.29	1673.91	16.8188

9 Results

a. **Confidence levels** In Fig. 1, we present the 1D and 2D posterior distributions at confidence levels of 68.3% (1σ) and 95.4% (2σ). These distributions are derived from the constraints imposed on the BAZS Type Model, incorporating data from various sources, including $H(z)$, Type Ia supernovae (SNIa), Baryon Acoustic Oscillation (BAO), Gamma Ray Burst (GRB), and Quasar observations.

b. **Cosmological and model parameter** Our analysis yields optimal values for several cosmological parameters, namely, Ω_{mo}, Ω_{ro}, q_0, q_1, and β, as well as the present value of the Hubble function, denoted as H_0, which are summarized in Table 2. In the context of the ΛCDM model, the parameter Ω_{mo} exhibits a range of values, spanning from 0.2568 to 0.3128. It's noteworthy that these values can vary across different datasets, illustrating the dataset-dependent constraints. On the other hand, for the BAZS type Model, the parameter Ω_{mo} presents values within the range of 0.1911–0.2531. These values, too, display variations contingent on the specific datasets employed. Within the BAZS type model, the parameter Ω_{ro}, signifying radiation density, is found to fall in the range of 0.0269–0.0493. However, it is important to note that these values do exhibit some degree of variability based on the dataset used. Regarding the BAZS type model, both q_0 and q_1 serve as representations of deceleration parameters. These parameters show relatively wide ranges, and their values are influenced by the choice of datasets. Different datasets yield distinct constraints on the current cosmic acceleration. The unique parameter in the BAZS type model, denoted as β, plays a pivotal role in shaping cosmic dynamics. Its values span from 3.05 to 3.19 across various datasets, with moderate levels of uncertainty. Our analysis also addresses the crucial parameter, the Hubble constant (H_0). In the ΛCDM model, H_0 displays a range of values, approximately between 68.60 and 74.85. This pattern is consistent across different datasets. Similarly, within the BAZS type model, H_0 varies in the range of approximately 67.77–74.87. Although the uncertainties in the BAZS model are somewhat larger, these values remain consistent across datasets. This persistent discrepancy or disagreement in the measurements of the H_0 value among both models underscores the presence of tension.

c. **Decceleration parameter** The comparative behavior of the deceleration parameter between our model and the standard ΛCDM has been shown in Fig. 5. We can observe that at high redshifts, there is minimal variation in the numerical values of q between both models, resembling the behavior of ΛCDM. However, at low redshifts, where the ΛCDM model transitions into a de Sitter phase with $q = -1$, our model exhibits a slightly higher acceleration with $q(-1) \approx -1.14$. This

difference suggests that our model extends the acceleration phase beyond what is predicted by the ΛCDM model. Furthermore, the phase transition redshift, denoted as z_{tr}, corresponds to the epoch where $q = 0$. Remarkably, the numerical values of z_{tr} for both models are very close to each other, indicating a comparable transition from deceleration to acceleration. These findings shed light on the dynamics of the expansion of the Universe. Since ΛCDM is widely accepted as the standard model, our model suggests a slight deviation at low redshifts, allowing a more prolonged acceleration period. The similarities and differences between the two models provide valuable insights into the behavior of the deceleration parameter and the underlying physics governing the evolution of the Universe.

d. **Jerk parameter** In Fig. 6, we compare the behavior of the jerk parameter j in two different models (ΛCDM model and our model). The ΛCDM model, which is widely accepted in cosmology, shows a constant value of $j = 1$ across all epochs and the phase transition redshift. This indicates that the ΛCDM model maintains a steady acceleration throughout the entire cosmic evolution. On the other hand, our model exhibits slight deviations in the jerk parameter at different epochs and the phase transition redshift. At extremely high redshifts, when the Universe was in its early stages ($z \to \infty$), the value of j is slightly lower, around 0.97. This suggests a slightly decelerating expansion during that early period. However, as the redshift approaches zero, indicating the present time ($z \to 0$), the value of j increases to 1.32, indicating a transition to a phase of constant acceleration. Additionally, at the phase transition redshift, the jerk parameter takes on a value of 1.17, signifying a temporal change in the accelerating behavior within our model. The ΛCDM model demonstrates a constant acceleration, in line with the concept of a cosmological constant driving the expansion. In contrast, our model exhibits slight variations in the jerk parameter, indicating deviations from a constant acceleration.

e. **The $\{q, r\}$ parameter** In Fig. 7, we observe the evolution of the model in the $\{q, r\}$ profile. Initially, the model exhibits values in the range $q > 0$ and $r < 1$. This region corresponds to the quintessence region in cosmology. Quintessence is a hypothetical form of DE characterized by a time-varying EoS. The positive q indicates an accelerating expansion of the Universe, while the $r < 1$ suggests that the energy density of quintessence dominates over the matter density, driving the accelerated expansion. As the evolution progresses, the model transitions to values in the range $q < 0$ and $r > 1$, which corresponds to the Chaplygin gas region. The Chaplygin gas is another DE model that exhibits exotic properties, such as a constant EoS. The negative q indicates a decelerating expansion of the Universe, and $r > 1$ suggests that the Chaplygin gas dominates over matter and causes the deceleration. Finally, the model deviates towards the de Sitter point located at $(-1, 1)$. The de Sitter Universe represents a special case where the expansion of the Universe is driven by a cosmological constant or vacuum energy, leading to a constant and accelerated expansion. Also, $q = -1$ signifies this acceleration, and $r = 1$ indicates the dominance of the cosmological constant over other energy components.

f. **The $\{s, r\}$ parameter** In Fig. 8, the evolution of the model in the $\{s, r\}$ profile provides valuable insights into the cosmological context. At early times, the model assumes values in the range $r < 1$ and $s > 0$, which corresponds to the

quintessence region. This indicates that during the early stages of the Universe, quintessence, a form of DE with a dynamic EoS, dominates the energy content. As the Universe evolves, the model crosses the fixed point $\{r, s\} = \{1, 0\}$, indicating a transition to the ΛCDM regime. In this late-time phase, $1 > 1$ while s becomes negative. This behavior is characteristic of the Chaplygin gas region, which is a model that incorporates a generalized form of DE. Therefore, the $\{s, r\}$ pair of Statefinder diagnostic provides crucial information about the cosmological evolution. It illustrates the transition from the quintessence region to the ΛCDM regime, highlighting the dominance of different forms of DE throughout cosmic history.

g. **Om diagnostic parameter** Fig. 9 shows us how the evolution of $Om(z)$ varies at different redshifts (z) for the BAZS Type Model. Notably, $Om(z)$ consistently stays below the current matter density parameter Ω_{m0}, suggesting that the model remains in the phantom region throughout the Universe's evolution at all redshifts.

h. **Information criteria** Based on the Table 3. The $\chi^{2\ min}_{\text{tot}}$ for the BAZS type model is lower than that of the ΛCDM model, indicating a better fit in terms of minimizing the total χ^2. The χ^2_{red} for the BAZS type model is also lower, suggesting a better fit when accounting for the degrees of freedom. The AIC_c for the BAZS type model is higher than that of the ΛCDM model, suggesting that the ΛCDM model is favored when considering both the goodness of fit and the complexity of the model. The calculated ΔAIC_c of 1.29 suggests that the BAZS model is less supported as compared to the ΛCDM model. This value falls within the range of models that are less strongly supported according to AIC. The BIC for the BAZS type model is higher than that of the ΛCDM model, further supporting the notion that the ΛCDM model is preferred in terms of model complexity and goodness of fit according to BIC. The calculated ΔBIC of 16.8188 indicates that the ΛCDM model is significantly more favored as compared to the BAZS type model. This substantial difference suggests strong evidence in favor of the ΛCDM model according to BIC.

10 Discussions and conclusions

The present study delves into an intriguing scenario of accelerated expansion in a flat FLRW Universe, encompassing DE, radiation, and DM. We aimed to develop a robust and credible DE model capable of comprehensively describing the entire evolutionary history of the Universe. Rather than adopting a conventional EoS parameterization, we proposed a novel approach by parametrizing the deceleration parameter. This unique choice allows us to explore the dynamics of the Universe without being bound by specific gravitational theories while capturing both early deceleration and late-time acceleration. Our selected parametrization, denoted as $q(z)$, involves three parameters: q_0, q_1, and β. Although it is possible to include more terms in the parametrization, we opted for a concise formulation to ensure feasibility in constraining the parameters using available observational data. To obtain the best-fit values of the model parameters, we have applied the MCMC algorithm, utilizing diverse datasets such as $H(z)$,

SNIa, BAO, GRB, and quasars. Through rigorous analysis, we explored the evolution of the deceleration and jerk parameters based on the obtained best-fit values. The graphical representations of the deceleration parameter revealed a distinct transition from deceleration to acceleration, signifying a departure from the conventional ΛCDM model. Furthermore, we observed the evolving nature of the jerk parameter, indicating a propensity for the Universe to exhibit behaviors beyond the standard framework. In addition, we discussed the Statefinder and $Om(z)$ diagnostic approaches, providing further insights into the viability of our models. By using the information criteria, we assessed the compatibility and adequacy of our proposed parametrization in the context of the accelerating Universe. In conclusion, the parametrization of the deceleration parameter investigated in this study offers a significant contribution to our understanding of the accelerating Universe. Through a comprehensive analysis of various observational datasets, we have established a robust framework capable of describing the intricate dynamics of the cosmos. We hope that this research will inspire further investigations and pave the way for more profound insights into the mysteries of our evolving Universe.

Acknowledgements G. Mustafa is very thankful to Prof. Gao Xianlong from the Department of Physics, Zhejiang Normal University, for his kind support and help during this research. Further, G. Mustafa acknowledges Grant No. ZC304022919 to support his Postdoctoral Fellowship at Zhejiang Normal University, China.

Author Contributions HC and SM: Conceived of the presented idea, Verified the analytical methods, Discussed the results and contributed to the final manuscript. AB: Developed the theory, Investigate the analysis and supervised the findings of this work, Discussed the results and contributed to the final manuscript. UD: Conceived of the presented idea, Discussed the results and contributed to the final manuscript. G. Mustafa: Performed the computations, Verified the analytical methods, Discussed the results and contributed to the final manuscript.

Declarations

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

References

1. Riess, A.G., et al.: Observational evidence from supernovae for an accelerating universe and a cosmological constant. Astron. J. **116**, 1009–1038 (1998). https://doi.org/10.1086/300499. arXiv:astro-ph/9805201
2. Perlmutter, S., et al.: Measurements of the cosmological parameters Omega and Lambda from the first 7 supernovae at z> = 0.35. Astrophys. J. **483**, 565 (1997). https://doi.org/10.1086/304265
3. Perlmutter, S., et al.: Measurements of Ω and Λ from 42 high redshift supernovae. Astrophys. J. **517**, 565–586 (1999). https://doi.org/10.1086/307221
4. Vollick, D.N.: 1/R curvature corrections as the source of the cosmological acceleration. Phys. Rev. D **68**, 063510 (2003). https://doi.org/10.1103/PhysRevD.68.063510
5. Nojiri, S., Odintsov, S.D.: Modified gravity with negative and positive powers of the curvature: unification of the inflation and of the cosmic acceleration. Phys. Rev. D **68**, 123512 (2003). https://doi.org/10.1103/PhysRevD.68.123512
6. Tonry, J.L., et al.: Cosmological results from high-z supernovae. Astrophys. J. **594**, 1–24 (2003). https://doi.org/10.1086/376865. arXiv:astro-ph/0305008

7. Riess, A.G., et al.: Type Ia supernova discoveries at z > 1 from the Hubble Space Telescope: evidence for past deceleration and constraints on dark energy evolution. Astrophys. J. **607**, 665–687 (2004). https://doi.org/10.1086/383612
8. Clocchiatti, A., et al.: Hubble space telescope and ground-based observations of type Ia supernovae at redshift 0.5: cosmological implications. Astrophys. J. **642**, 1–21 (2006). https://doi.org/10.1086/498491
9. Caldwell, R.R., Komp, W., Parker, L., Vanzella, D.A.T.: A sudden gravitational transition. Phys. Rev. D **73**, 023513 (2006). https://doi.org/10.1103/PhysRevD.73.023513. arXiv:astro-ph/0507622
10. Sahni, V., Starobinsky, A.A.: The Case for a positive cosmological Lambda term. Int. J. Mod. Phys. D **9**, 373–444 (2000). https://doi.org/10.1142/S0218271800000542
11. Padmanabhan, T.: Cosmological constant: the weight of the vacuum. Phys. Rep. **380**, 235–320 (2003). https://doi.org/10.1016/S0370-1573(03)00120-0. arXiv:hep-th/0212290
12. Peebles, P.J.E., Ratra, B.: The cosmological constant and dark energy. Rev. Mod. Phys. **75**, 559–606 (2003). https://doi.org/10.1103/RevModPhys.75.559
13. Copeland, E.J., Sami, M., Tsujikawa, S.: Dynamics of dark energy. Int. J. Mod. Phys. D **15**(11), 1753–1935 (2006)
14. Amendola, L., Tsujikawa, S.: Dark Energy: Theory and Observations. Cambridge University Press, Cambridge (2010)
15. Steinhardt, P.J., Wang, L., Zlatev, I.: Cosmological tracking solutions. Phys. Rev. D **59**(12), 123504 (1999)
16. Weinberg, S.: The cosmological constant problem. Rev. Mod. Phys. **61**, 1–23 (1989). https://doi.org/10.1103/RevModPhys.61.1
17. Bouali, A., Albarran, I., Bouhmadi-López, M., Ouali, T.: Cosmological constraints of phantom dark energy models. Phys. Dark Univ. **26**, 100391 (2019). https://doi.org/10.1016/j.dark.2019.100391. arXiv:1905.07304
18. Bouali, A., Albarran, I., Bouhmadi-Lopez, M., Errahmani, A., Ouali, T.: Cosmological constraints of interacting phantom dark energy models. Phys. Dark Univ. **34**, 100907 (2021). https://doi.org/10.1016/j.dark.2021.100907. arXiv:2103.13432
19. Mhamdi, D., Bargach, F., Dahmani, S., Bouali, A., Ouali, T.: Comparing phantom dark energy models with various diagnostic tools. Gen. Relativ. Gravit. **55**(1), 11 (2023). https://doi.org/10.1007/s10714-022-03055-7
20. Dahmani, S., Bouali, A., El Bojaddaini, I., Errahmani, A., Ouali, T.: Constraining neutrino properties and smoothing the Hubble tension via the LSBR model. Gen. Relativ. Gravit. **55**(1), 22 (2023). https://doi.org/10.1007/s10714-023-03066-y
21. Dahmani, S., Bouali, A., Bojaddaini, I.E., Errahmani, A., Ouali, T.: Smoothing the H_0 tension with a dynamical dark energy model. arXiv:2301.04200
22. Bouali, A., Chaudhary, H., Mehrotra, A., Pacif, S.K.J.: Model-independent study for a quintessence model of dark energy: analysis and observational constraints. arXiv:2304.02652
23. Capozziello, S., Cardone, V.F., Elizalde, E., Nojiri, S., Odintsov, S.D.: Observational constraints on dark energy with generalized equations of state. Phys. Rev. D **73**, 043512 (2006). https://doi.org/10.1103/PhysRevD.73.043512. arXiv:astro-ph/0508350
24. Escamilla-Rivera, C., Capozziello, S.: Unveiling cosmography from the dark energy equation of state. Int. J. Mod. Phys. D **28**(12), 1950154 (2019). https://doi.org/10.1142/S0218271819501542
25. Debnath, U.: Gravitational waves for variable modified Chaplygin gas and some parametrizations of dark energy in the background of FRW universe. Eur. Phys. J. Plus **135**(2), 135 (2020). https://doi.org/10.1140/epjp/s13360-020-00219-9
26. del Campo, S., Duran, I., Herrera, R., Pavon, D.: Three thermodynamically-based parameterizations of the deceleration parameter. Phys. Rev. D **86**, 083509 (2012). https://doi.org/10.1103/PhysRevD.86.083509. arXiv:1209.3415
27. Cunha, J., Lima, J.A.S.: Transition redshift: new kinematic constraints from supernovae. Mon. Not. R. Astron. Soc. **390**(1), 210–217 (2008)
28. Cunha, J.V.: Kinematic constraints to the transition redshift from SNe Ia union data. Phys. Rev. D **79**, 047301 (2009). https://doi.org/10.1103/PhysRevD.79.047301. arXiv:0811.2379
29. Riess, A.G., et al.: Type Ia supernova discoveries at z > 1 from the Hubble Space Telescope: evidence for past deceleration and constraints on dark energy evolution. Astrophys. J. **607**, 665–687 (2004). https://doi.org/10.1086/383612

30. Xu, L.-I., Zhang, C.-W., Chang, B.-R., Liu, H.-Y.: Constraints to deceleration parameters by recent cosmic observations. Mod. Phys. Lett. A **23**, 1939–1948 (2008). https://doi.org/10.1142/S0217732308025991
31. Xu, L., Lu, J.: Cosmic constraints on deceleration parameter with Sne Ia and CMB. Mod. Phys. Lett. A **24**, 369–376 (2009). https://doi.org/10.1142/S0217732309027212
32. Nair, R., Jhingan, S., Jain, D.: Cosmokinetics: a joint analysis of standard candles, rulers and cosmic clocks. J. Cosmol. Astropart. Phys. **2012**(01), 018 (2012)
33. Akarsu, O., Dereli, T., Kumar, S., Xu, L.: Probing kinematics and fate of the Universe with linearly time-varying deceleration parameter. Eur. Phys. J. Plus **129**, 22 (2014). https://doi.org/10.1140/epjp/i2014-14022-6
34. Santos, B., Carvalho, J.C., Alcaniz, J.S.: Current constraints on the epoch of cosmic acceleration. Astropart. Phys. **35**(1), 17–20 (2011)
35. Gong, Y.-G., Wang, A.: Reconstruction of the deceleration parameter and the equation of state of dark energy. Phys. Rev. D **75**, 043520 (2007). https://doi.org/10.1103/PhysRevD.75.043520. arXiv:astro-ph/0612196
36. Turner, M.S., Riess, A.G.: Do SNe Ia provide direct evidence for past deceleration of the universe? Astrophys. J. **569**, 18 (2002). https://doi.org/10.1086/338580. arXiv:astro-ph/0106051
37. Al Mamon, A., Das, S.: A divergence free parametrization of deceleration parameter for scalar field dark energy. Int. J. Mod. Phys. D **25**(03), 1650032 (2016). https://doi.org/10.1142/S0218271816500322
38. Gadbail, G.N., Mandal, S., Sahoo, P.K.: Parametrization of deceleration parameter in f (q) gravity. Physics **4**(4), 1403–1412 (2022)
39. Bouali, A., Shukla, B.K., Chaudhary, H., Tiwari, R.K., Samar, M., Mustafa, G.: Cosmological tests of parametrization $q = \alpha - \beta$ H in f(Q) FLRW cosmology. Int. J. Geom. Methods Mod. Phys. (2023). https://doi.org/10.1142/s0219887823501529
40. Khurana, M., Chaudhary, H., Mumtaz, S., Pacif, S., Mustafa, G.: Analyzing a higher order $q(t)$ model and its implications in the late evolution of the universe using recent observational datasets. arXiv preprint arXiv:2309.14222
41. Bouali, A., Chaudhary, H., Debnath, U., Sardar, A., Mustafa, G.: Data analysis of three parameter models of deceleration parameter in frw universe. arXiv preprint arXiv:2304.13137
42. Bouali, A., Chaudhary, H., Mehrotra, A., Pacif, S.: Model-independent study for a quintessence model of dark energy: analysis and observational constraints. arXiv preprint arXiv:2304.02652
43. Arora, D., Chaudhary, H., Pacif, S.K.J.: Diagnostic and comparative analysis of dark energy models with $q(z)$ parametrizations. Available at SSRN 4543124
44. Shekh, S., Chaudhary, H., Bouali, A., Dixit, A.: Observational constraints on teleparallel effective equation of state. Gen. Relativ. Gravit. **55**(8), 95 (2023)
45. Chaudhary, H., Arora, D., Debnath, U., Mustafa, G., Maurya, S.K.: A new cosmological model: exploring the evolution of the universe and unveiling super-accelerated expansion. arXiv preprint arXiv:2308.07354
46. Shukla, B.K., Bouali, A., Chaudhary, H., Tiwari, R.K., San Martin, M.: Cosmographic studies of $q = \alpha - \beta$ h parametrization in f (t) framework. Int. J. Geom. Methods Mod. Phys. (2023). https://doi.org/10.1142/S0219887824500075
47. Chaudhary, H., Bouali, A., Debnath, U., Roy, T., Mustafa, G.: Constraints on the parameterized deceleration parameter in frw universe. Phys. Scr. **98**(9), 095006 (2023)
48. Bouali, A., Chaudhary, H., Mumtaz, S., Mustafa, G., Maurya, S.: Observational constraining study of new deceleration parameters in frw universe. Fortschr. Phys. (2023). https://doi.org/10.1002/prop.202300033
49. Bouali, A., Shukla, B., Chaudhary, H., Tiwari, R.K., Samar, M., Mustafa, G.: Cosmological tests of parametrization q= α- β h in f (q) flrw cosmology. Int. J. Geom. Methods Mod. Phys. (2023). https://doi.org/10.1142/S0219887823501529
50. Chaudhary, H., Kaushik, A., Kohli, A.: Cosmological test of $\sigma\theta$ as function of scale factor in f (r, t) framework. New Astron. **103**, 102044 (2023)
51. Khurana, M., Chaudhary, H., Debnath, U., Molla, N.U., Mustafa, G.: Cosmological test of dark energy parametrizations in Horava–Lifshitz gravity. arXiv preprint arXiv:2310.07410
52. Mamon, A.A., Das, S.: A parametric reconstruction of the deceleration parameter. Eur. Phys. J. C **77**(7), 495 (2017)

53. Jamil, M., Momeni, D., Myrzakulov, R.: Observational constraints on non-minimally coupled Galileon model. Eur. Phys. J. C **73**(3), 2347 (2013)
54. Bouali, A., Chaudhary, H., Debnath, U., Roy, T., Mustafa, G.: Constraints on the parameterized deceleration parameter in FRW universe. arXiv:2301.12107
55. Shekh, S.H., Bouali, A., Mustafa, G., Pradhan, A., Javed, F.: Observational constraints in accelerated emergent $f(Q)$ gravity model. Class. Quantum Gravity **40**(5), 055011 (2023). https://doi.org/10.1088/1361-6382/acb631
56. Gadbail, G.N., Harshita, Bouali, A., Sahoo, P.K.: Statistical and cosmological analysis of Weyl-type $f(Q,T)$ models using Pantheon+ dataset. arXiv:2305.11190
57. Barboza, E., Jr., Alcaniz, J., Zhu, Z.-H., Silva, R.: Generalized equation of state for dark energy. Phys. Rev. D **80**(4), 043521 (2009)
58. Kundu, R., Debnath, U., Pradhan, A.: Studying the optical depth behaviour of parametrized deceleration parameter in non-flat universe. Int. J. Geom. Methods Mod. Phys. **20**, 2350110 (2023)
59. Bandyopadhyay, T., Debnath, U.: Fluid accretion upon higher-dimensional wormhole and black hole for parameterized deceleration parameter. Int. J. Geom. Methods Mod. Phys. **19**(12), 2250182 (2022)
60. Foreman-Mackey, D., Hogg, D.W., Lang, D., Goodman, J.: emcee: the MCMC hammer. Publ. Astron. Soc. Pac. **125**(925), 306 (2013)
61. Brooks, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. **7**(4), 434–455 (1998)
62. Lewis, A., Bridle, S.: Cosmological parameters from CMB and other data: a Monte Carlo approach. Phys. Rev. D **66**(10), 103511 (2002)
63. Gilks, W.R., Richardson, S., Spiegelhalter, D.: Markov Chain Monte Carlo in Practice. CRC Press, Boca Raton (1995)
64. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)
65. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications
66. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis. CRC Press, Boca Raton (2013)
67. Verde, L., Peiris, H., Spergel, D., Nolta, M., Bennett, C., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., et al.: First-year Wilkinson microwave anisotropy probe (wmap)* observations: parameter estimation methodology. Astrophys. J. Suppl. Ser. **148**(1), 195 (2003)
68. Trotta, R.: Applications of Bayesian model selection to cosmological parameters. Mon. Not. R. Astron. Soc. **378**(1), 72–82 (2007)
69. Gaztanaga, E., Bonvin, C., Hui, L.: Measurement of the dipole in the cross-correlation function of galaxies. J. Cosmol. Astropart. Phys. **2017**(01), 032 (2017)
70. Stern, D., Jimenez, R., Verde, L., Kamionkowski, M., Stanford, S.A.: Cosmic chronometers: constraining the equation of state of dark energy. I: H (z) measurements. J. Cosmol. Astropart. Phys. **2010**(02), 008 (2010)
71. Lee, S.: Constraining minimally extended varying speed of light by cosmological chronometers. arXiv preprint arXiv:2301.06947
72. Gaztanaga, E., Cabre, A., Hui, L.: Clustering of luminous red galaxies IV: baryon acoustic peak in the line-of-sight direction and a direct measurement of H(z). Mon. Not. R. Astron. Soc. **399**, 1663–1680 (2009). https://doi.org/10.1111/j.1365-2966.2009.15405.x
73. Chimento, L., Forte, M.I.: Unified model of baryonic matter and dark components. Phys. Lett. B **666**, 205–211 (2008). https://doi.org/10.1016/j.physletb.2008.07.064
74. Oka, A., Saito, S., Nishimichi, T., Taruya, A., Yamamoto, K.: Simultaneous constraints on the growth of structure and cosmic expansion from the multipole power spectra of the SDSS DR7 LRG sample. Mon. Not. R. Astron. Soc. **439**, 2515–2530 (2014). https://doi.org/10.1093/mnras/stu111
75. Wang, Y., et al.: BOSS. Mon. Not. R. Astron. Soc. **469**(3), 3762–3774 (2017)
76. Moresco, M., Cimatti, A., Jimenez, R., Pozzetti, L., Zamorani, G., Bolzonella, M., Dunlop, J., Lamareille, F., Mignoli, M., Pearce, H., et al.: Improved constraints on the expansion rate of the universe up to z 1.1 from the spectroscopic evolution of cosmic chronometers. J. Cosmol. Astropart. Phys. **2012**(08), 006 (2012)
77. Chuang, C.-H., Wang, Y.: Modeling the anisotropic two-point galaxy correlation function on small scales and improved measurements of $H(z)$, $D_A(z)$, and $\beta(z)$ from the Sloan digital sky survey DR7 luminous red galaxies. Mon. Not. R. Astron. Soc. **435**, 255–262 (2013). https://doi.org/10.1093/mnras/stt1290

78. Alam, S.: BOSS. Mon. Not. R. Astron. Soc. **470**(3), 2617–2652 (2017). https://doi.org/10.1093/mnras/stx721
79. Blake, C., et al.: The WiggleZ Dark Energy Survey: joint measurements of the expansion and growth history at z < 1. Mon. Not. R. Astron. Soc. **425**, 405–414 (2012). https://doi.org/10.1111/j.1365-2966.2012.21473.x
80. Zhang, C., Zhang, H., Yuan, S., Liu, S., Zhang, T.-J., Sun, Y.-C.: Four new observational h (z) data from luminous red galaxies in the Sloan digital sky survey data release seven. Res. Astron. Astrophys. **14**(10), 1221 (2014)
81. Chuang, C.-H., Prada, F., Cuesta, A.J., Eisenstein, D.J., Kazin, E., Padmanabhan, N., Sánchez, A.G., Xu, X., Beutler, F., Manera, M., et al.: The clustering of galaxies in the SDSS-III Baryon oscillation spectroscopic survey: single-probe measurements and the strong power of f (z) σ8 (z) on constraining dark energy. Mon. Not. R. Astron. Soc. **433**(4), 3559–3571 (2013)
82. Moresco, M., Pozzetti, L., Cimatti, A., Jimenez, R., Maraston, C., Verde, L., Thomas, D., Citro, A., Tojeiro, R., Wilkinson, D.: A 6% measurement of the hubble parameter at z 0.45: direct evidence of the epoch of cosmic re-acceleration. J. Cosmol. Astropart. Phys. **2016**(05), 014 (2016)
83. Lee, S.: Cosmic distance duality as a probe of minimally extended varying speed of light. arXiv preprint arXiv:2108.06043
84. Delubac, T., Bautista, J.E., Rich, J., Kirkby, D., Bailey, S., Font-Ribera, A., Slosar, A., Lee, K.-G., Pieri, M.M., Hamilton, J.-C., et al.: Baryon acoustic oscillations in the lyα forest of boss dr11 quasars. Astron. Astrophys. **574**, A59 (2015)
85. Bautista, J.E., Guy, J., Rich, J., Blomqvist, M., Des Bourboux, H.D.M., Pieri, M.M., Font-Ribera, A., Bailey, S., Delubac, T., Kirkby, D., et al.: Measurement of baryon acoustic oscillation correlations at z= 2.3 with sdss dr12 lyα-forests. Astron. Astrophys. **603**, A12 (2017)
86. Font-Ribera, A., Kirkby, D., Busca, N., Miralda-Escude, J., Ross, N.P., Slosar, A., Rich, J., Aubourg, E., Bailey, S., Bhardwaj, V., et al.: Quasar-Lyman α forest cross-correlation from boss dr11: baryon acoustic oscillations. J. Cosmol. Astropart. Phys. **2014**(05), 027–027 (2014)
87. Ratsimbazafy, A.L., Loubser, S.I., Crawford, S.M., Cress, C.M., Bassett, B.A., Nichol, R.C., Väisänen, P.: Age-dating luminous red galaxies observed with the Southern African large telescope. Mon. Not. R. Astron. Soc. **467**(3), 3239–3254 (2017). https://doi.org/10.1093/mnras/stx301
88. Scolnic, D.M., Jones, D., Rest, A., Pan, Y., Chornock, R., Foley, R., Huber, M., Kessler, R., Narayan, G., Riess, A., et al.: The complete light-curve sample of spectroscopically confirmed SNe Ia from pan-starrs1 and cosmological constraints from the combined pantheon sample. Astrophys. J. **859**(2), 101 (2018)
89. Conley, A., Guy, J., Sullivan, M., Regnault, N., Astier, P., Balland, C., Basa, S., Carlberg, R., Fouchez, D., Hardin, D., et al.: Supernova constraints and systematic uncertainties from the first three years of the supernova legacy survey. Astrophys. J. Suppl. Ser. **192**(1), 1 (2010)
90. Demianski, M., Piedipalumbo, E., Sawant, D., Amati, L.: Cosmology with gamma-ray bursts-II. Cosmography challenges and cosmological scenarios for the accelerated universe. Astron. Astrophys. **598**, A113 (2017)
91. Roberts, C., Horne, K., Hodson, A.O., Leggat, A.D.: Tests of λcdm and conformal gravity using grb and quasars as standard candles out to $z \sim 8$. arXiv preprint arXiv:1711.10369
92. Percival, W.J., Reid, B.A., Eisenstein, D.J., Bahcall, N.A., Budavari, T., Frieman, J.A., Fukugita, M., Gunn, J.E., Ivezi'c, Z., Knapp, G.R., et al.: Baryon acoustic oscillations in the Sloan digital sky survey data release 7 galaxy sample. Mon. Not. R. Astron. Soc. **401**(4), 2148–2168 (2010)
93. Beutler, F., Blake, C., Colless, M., Jones, D.H., Staveley-Smith, L., Campbell, L., Parker, Q., Saunders, W., Watson, F.: The 6df galaxy survey: baryon acoustic oscillations and the local hubble constant. Mon. Not. R. Astron. Soc. **416**(4), 3017–3032 (2011)
94. Delubac, T., Rich, J., Bailey, S., Font-Ribera, A., Kirkby, D., Le Goff, J.-M., Pieri, M.M., Slosar, A., Aubourg, É., Bautista, J.E., et al.: Baryon acoustic oscillations in the lyα forest of boss quasars. Astron. Astrophys. **552**, A96 (2013)
95. Anderson, L., Aubourg, E., Bailey, S., Bizyaev, D., Blanton, M., Bolton, A.S., Brinkmann, J., Brownstein, J.R., Burden, A., Cuesta, A.J., et al.: The clustering of galaxies in the SDSS-III baryon oscillation spectroscopic survey: baryon acoustic oscillations in the data release 9 spectroscopic galaxy sample. Mon. Not. R. Astron. Soc. **427**(4), 3435–3467 (2012)
96. Seo, H.-J., Ho, S., White, M., Cuesta, A.J., Ross, A.J., Saito, S., Reid, B., Padmanabhan, N., Percival, W.J., De Putter, R., et al.: Acoustic scale from the angular power spectra of SDSS-III dr8 photometric luminous galaxies. Astrophys. J. **761**(1), 13 (2012)

97. Ross, A.J., Samushia, L., Howlett, C., Percival, W.J., Burden, A., Manera, M.: The clustering of the SDSS dr7 main galaxy sample-I. A 4 per cent distance measure at z= 0.15. Mon. Not. R. Astron. Soc. **449**(1), 835–847 (2015)
98. Tojeiro, R., Ross, A.J., Burden, A., Samushia, L., Manera, M., Percival, W.J., Beutler, F., Brinkmann, J., Brownstein, J.R., Cuesta, A.J., et al.: The clustering of galaxies in the SDSS-III baryon oscillation spectroscopic survey: galaxy clustering measurements in the low-redshift sample of data release 11. Mon. Not. R. Astron. Soc. **440**(3), 2222–2237 (2014)
99. Bautista, J.E., Vargas-Magaña, M., Dawson, K.S., Percival, W.J., Brinkmann, J., Brownstein, J., Camacho, B., Comparat, J., Gil-Marín, H., Mueller, E.-M., et al.: The SDSS-IV extended baryon oscillation spectroscopic survey: baryon acoustic oscillations at redshift of 0.72 with the dr14 luminous red galaxy sample. Astrophys. J. **863**(1), 110 (2018)
100. De Carvalho, E., Bernui, A., Carvalho, G., Novaes, C., Xavier, H.: Angular baryon acoustic oscillation measure at z= 2.225 from the SDSS quasar survey. J. Cosmol. Astropart. Phys. **2018**(04), 064 (2018)
101. Ata, M., Baumgarten, F., Bautista, J., Beutler, F., Bizyaev, D., Blanton, M.R., Blazek, J.A., Bolton, A.S., Brinkmann, J., Brownstein, J.R., et al.: The clustering of the SDSS-IV extended baryon oscillation spectroscopic survey dr14 quasar sample: first measurement of baryon acoustic oscillations between redshift 0.8 and 2.2. Mon. Not. R. Astron. Soc. **473**(4), 4773–4794 (2018)
102. Abbott, T., Abdalla, F., Alarcon, A., Allam, S., Andrade-Oliveira, F., Annis, J., Avila, S., Banerji, M., Banik, N., Bechtol, K., et al.: Dark energy survey year 1 results: Measurement of the baryon acoustic oscillation scale in the distribution of galaxies to redshift 1. Mon. Not. R. Astron. Soc. **483**(4), 4866–4883 (2019)
103. Molavi, Z., Khodam-Mohammadi, A.: Observational tests of Gauss–Bonnet like dark energy model. Eur. Phys. J. Plus **134**(6), 254 (2019)
104. Benisty, D., Staicova, D.: Testing late-time cosmic acceleration with uncorrelated baryon acoustic oscillation dataset. Astron. Astrophys. **647**, A38 (2021)
105. Kazantzidis, L., Perivolaropoulos, L.: Evolution of the f σ 8 tension with the Planck 15/λ CDM determination and implications for modified gravity theories. Phys. Rev. D **97**(10), 103503 (2018)
106. Hogg, N.B., Martinelli, M., Nesseris, S.: Constraints on the distance duality relation with standard sirens. J. Cosmol. Astropart. Phys. **2020**(12), 019 (2020)
107. Martinelli, M., Martins, C.J.A.P., Nesseris, S., Sapone, D., Tutusaus, I., Avgoustidis, A., Camera, S., Carbone, C., Casas, S., Ilić, S., et al.: Euclid: forecast constraints on the cosmic distance duality relation with complementary external probes. Astron. Astrophys. **644**, A80 (2020)
108. Sahni, V., Saini, T.D., Starobinsky, A.A., Alam, U.: Statefinder: a new geometrical diagnostic of dark energy. J. Exp. Theor. Phys. Lett. **77**, 201–206 (2003)
109. Visser, M.: Jerk, snap and the cosmological equation of state. Class. Quantum Gravity **21**(11), 2603 (2004)
110. Luongo, O.: Dark energy from a positive jerk parameter. Mod. Phys. Lett. A **28**, 1350080 (2013). https://doi.org/10.1142/S0217732313500806
111. Dunajski, M., Gibbons, G.: Cosmic jerk, snap and beyond. Class. Quantum Gravity **25**(23), 235012 (2008)
112. Sahni, V., Saini, T.D., Starobinsky, A.A., Alam, U.: Statefinder: a new geometrical diagnostic of dark energy. J. Exp. Theor. Phys. Lett. **77**, 201–206 (2003)
113. Alam, U., Sahni, V., Deep Saini, T., Starobinsky, A.: Exploring the expanding universe and dark energy using the statefinder diagnostic. Mon. Not. R. Astron. Soc. **344**(4), 1057–1074 (2003)
114. Sami, M., Shahalam, M., Skugoreva, M., Toporensky, A.: Cosmological dynamics of a nonminimally coupled scalar field system and its late time cosmic relevance. Phys. Rev. D **86**(10), 103532 (2012)
115. Myrzakulov, R., Shahalam, M.: Statefinder hierarchy of bimetric and Galileon models for concordance cosmology. J. Cosmol. Astropart. Phys. **2013**(10), 047 (2013)
116. Aviles, A., Klapp, J., Luongo, O.: Toward unbiased estimations of the statefinder parameters. Phys. Dark Univ. **17**, 25–37 (2017)
117. Sahni, V., Shafieloo, A., Starobinsky, A.A.: Two new diagnostics of dark energy. Phys. Rev. D **78**(10), 103502 (2008)
118. Zunckel, C., Clarkson, C.: Consistency tests for the cosmological constant. Phys. Rev. Lett. **101**(18), 181301 (2008)
119. Shahalam, M., Pathak, S., Verma, M., Khlopov, M.Y., Myrzakulov, R.: Dynamics of interacting quintessence. Eur. Phys. J. C **75**, 1–9 (2015)

120. Agarwal, A., Myrzakulov, R., Pacif, S., Shahalam, M.: Cosmic acceleration from coupling of baryonic and dark matter components: analysis and diagnostics. Int. J. Mod. Phys. D **28**(06), 1950083 (2019)
121. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
122. Liddle, A.R.: How many cosmological parameters. Mon. Not. R. Astron. Soc. **351**(3), L49–L53 (2004)
123. Nesseris, S., Garcia-Bellido, J.: Is the Jeffreys' scale a reliable tool for Bayesian model comparison in cosmology? J. Cosmol. Astropart. Phys. **2013**(08), 036 (2013)

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

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

Authors and Affiliations

Himanshu Chaudhary[1,2,3] · Saadia Mumtaz[4] · Amine Bouali[5] · Ujjal Debnath[6] · G. Mustafa[7,8]

Himanshu Chaudhary
himanshuch1729@gmail.com

Saadia Mumtaz
saadia.icet@pu.edu.pk

Amine Bouali
a1.bouali@ump.ac.ma

Ujjal Debnath
ujjaldebnath@gmail.com

1 Department of Applied Mathematics, Delhi Technological University, Delhi 110042, India

2 Pacif Institute of Cosmology and Selfology (PICS) Sagara, Sambalpur, Odisha 768224, India

3 Department of Mathematics, Shyamlal College, University of Delhi, Delhi 110032, India

4 Institute of Chemical Engineering and Technology, University of the Punjab, Quaid-e-Azam Campus, Lahore 54590, Pakistan

5 Laboratory of Physics of Matter and Radiation, Mohammed I University, BP 717 Oujda, Morocco

6 Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711 103, India

7 Department of Physics, Zhejiang Normal University, Jinhua 321004, People's Republic of China

8 New Uzbekistan University, Movarounnahr street 1, Tashkent 100007, Uzbekistan

Solanki *et al. Int. J. Air-Cond. Ref.* (2023) 31:26
https://doi.org/10.1007/s44189-023-00042-8

International Journal of
Air-Conditioning and Refrigeration

RESEARCH Open Access

Performance enhancement and environmental analysis of vapor compression refrigeration system with dedicated mechanical subcooling

Naveen Solanki[1*], Akhilesh Arora[2] and Raj Kumar Singh[2]

Abstract

The primary focus of this study is on the energy, exergy, and environmental (3E) analysis of a dedicated mechanical subcooled vapor compression refrigeration (DMS-VCR) system for applications involving commercially available water chillers that employ R134a (in both subcooler and main cycle). For a cooling capacity of 100-kW water chillers, the mathematical model of the DMS-VCR system is built to determine the performance parameter of the system. The DMS-VCR system reduces electricity usage by 15.52% and increase in COP by 9.5%, which results in a significant reduction in CO_2 emissions of about 15.55%. When compared to equivalent vapor compression refrigeration system (VCRS), the system's exergetic efficiency is also increased by 8%. Since the computer simulation results will undoubtedly give design engineers a better option, the subcooling and superheating of the vapor compression refrigeration system become alluring in this study. Consequently, the DMS-VCR system performs better as per the combined 3E study.

Keywords Energy, Exergy, Environment, Exergetic efficiency

1 Introduction

The vapor compression refrigeration system (VCRS) is an established technology, and the majority of the world's total energy consumption is accounted for by "heating and cooling" processes like water heating, space cooling, refrigeration, air-conditioning, food and beverage processes, and heat demands in industrial processes. The demand for electricity (high-grade energy) is enormous in many developing nations, especially during the summer when the load on refrigeration and air-conditioning increases. Effective refrigeration systems are necessary to meet the demand for high-grade energy (electricity) during peak load. High-quality energy is becoming more and more expensive. Because of this, some people currently call for the creation of effective and sustainable VCRS. In order to save costs, scientists are anticipating sustainable and energy-efficient technologies. Many researchers have already demonstrated that mechanical subcooling of the vapor compression refrigeration system improves its performance. A DMS-VCR system can partially satisfy the need since even little changes in electricity can result in considerable global savings of high-grade energy. Two vapor compression cycles—one main and one subcooled—are part of a DMS-VCR cycle. A large quantity of electric energy can be saved by subcooling the refrigerant during the main compression cycle, which essentially improves system efficiency. In the literature survey, the energy savings is examined for dedicated and

*Correspondence:
Naveen Solanki
naveensolanki@mait.ac.in
[1] Department of Mechanical Engineering, Maharaja Agrasen Institute of Technology, PSP Area Sector-22, Rohini, Delhi 110086, India
[2] Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulat Pur, Bawana Road, Delhi 110042, India

Springer

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

integrated mechanical subcooling systems in CO_2 supporter systems in many supermarket applications [1]. The mechanical subcooling of VCR systems is investigated in the condition of optimal operation and management [2, 3]. They also highlighted some recent work in this direction and experimental work to prove the concept of the dedicated subcooling cycle. The effect of combining various refrigerants is investigated on DMS-VCR system performance [4]. They discovered that when R22 is subcooled by 5–8 °C in the main cycle, the evaporator's load-bearing capacity increases by 0.5 kW. Many researchers conducted an energy and exergy study to improve the mechanical subcooled system's performance and failed to consider economic and environmental factors, which are critical when deciding how to use energy resources responsibly. An artificial neural network approach is employed for thermodynamic properties of refrigerants, to simulate thermo-economic optimization based on exergy method (heat exchangers—can be optimized on their own) for three refrigerants: R134a, R22, and R407c in a superheated and subcooled VCR system [5]. Researchers theoretically analyzed the use of expander for additional compression work considerations to drive the mechanical subcooling cycle [6]. They claim that the hybrid VCR system outperforms the conventional VCR system, the conventional expansion power recovery system, and the conventional mechanical subcooling system in terms of coefficient of performance (COP), with maximum COP increases of 67.76%, 17.73%, and 19.27%, respectively, in systems using R744 as the main cycle refrigerant. Experimental and theoretical observations result into a 9% increase in efficiency with optimal subcooling [7]. They also discovered that the best subcooling varies depending on the operating conditions. The effect of condenser subcooling is investigated in R134a and R1234yf air-conditioning systems with and without internal heat exchangers, revealing that R1234yf had an increase in COP up to 18%, but for R134a, it is just 9%, indicating that R1234yf outperforms R134a [8]. They also discovered that combining internal heat exchangers with condenser subcooling leads to a more effective R1234yf air-conditioning system. The performance of a transcritical CO_2 refrigeration plant is improved by using a DMS system with zeotropic refrigerant mixtures, both conceptually and experimentally [9]. They claimed that these mixtures reduced irreversibilities in the cycle. The influence of compressor isentropic efficiency (0.1–1.0) and subcooler effectiveness (0.3–1.0) showed that the DMS vapor compression refrigeration cycle outperforms the simple VCRC [10].

The majority of DMS-VCR system studies are mostly introductory, focusing on the system's energy and exergy analysis. The inclusion of a thorough environmental investigation of DMS-VCR system performance based on 3E analysis viewpoint makes this work a pioneering effort in the area of DMS-VCR systems. The system's energy, exergy, and environmental (3E) performance are predicted by the proposed model. The DMS-VCR system concept is used in the current analysis to apply to a 100kW industrial water chillers. After creating a model on engineering equation solver (EES) software, the system is subjected to a 3E analysis for a variety of operating conditions (see Table 1). It highlights the importance of environment analysis in DMS-VCR systems, particularly for high cooling capacity systems using specific refrigerants like R134a. In addition, it is worth noting that a higher COP not only signifies improved exergetic efficiency but also indicates reduced energy consumption. This outcome is both economically advantageous and environmentally favorable.

In conclusion, the originality of conducting a 3E analysis for a refrigeration system utilizing the DMS system lies in its exclusion from conventional assessment techniques that primarily focus on COP. Instead, it provides a more

Table 1 Process data of DMS-VCR system (for refrigerant R134a)

Parameters		Values
Thermal parameters [11]	Evaporator cooling capacity ($\dot{Q}_{evap}$, kW)	100
	Evaporator temperature (T_{evap} °C)	0 (range − 5 to 5)
	Inlet temperature of evaporator coolant ($T_{evap,in}$ °C)	10
	Outlet temperature of evaporator coolant ($T_{evap,out}$ °C)	5
	Condenser temperature (T_{cond} °C)	40 (range 35 to 45)
	Condenser coolant inlet temperature ($T_{cond,in}$ °C)	30
	Condenser coolant outlet temperature ($T_{cond,out}$ °C)	35
	Overlap degree ($T_{overlap}$, °C)	5
	Degree of subcooling ($T_{subcooling}$, °C)	5, 18 (optimized)
	Degree of superheating ($T_{superheat}$, °C)	10
	Isentropic efficiency of compressors (o_{isen})	0.68, 0.75
	Environment temperature (T_o, °C)	25
	Environment pressure (P_o, kPa)	101.325
	Refrigerants (for both the cycles)	R134a

comprehensive evaluation of the system's performance. The finding of the optimal degree of subcooling (i.e. 18 °C) takes into account several crucial factors, including the overall COP, exergetic efficiency, irreversibility, and total work, which leads to an decrease in the system's irreversibility and a subsequent increase in COP upto optimal point and reverse pattern are observed above optimal point. This finding further spotlights the uniqueness of this research, as it goes beyond conventional approaches and provides a holistic assessment of the system's performance.

2 Description and modeling of DMS-VCR system

2.1 Dedicated mechanical subcooled refrigeration system

Figure 1 depicts the schematic form of primary components of a DMS-VCR system. These components include two condensers, two compressors, two expansion valves, one evaporator, and one subcooler. And Fig. 2 shows temperature-entropy curve for various operations in the main and subcooling auxiliary unit. The cycle works between the condenser and evaporator temperatures, respectively.

3 Thermodynamic modeling of DMS-VCR system

3.1 Energy and exergy analysis

Thermodynamic modeling of the DMS-VCR system for energy and exergy analysis based on first and second law respectively is carried out on EES with thermophysical properties of refrigerants derived from inbuilt EES functions at various state points of the system. Moving fluid streams with a finite mass flow rate and their specific heat contributes significantly to the heat transfer rate in heat exchangers (evaporator, condenser-1, condenser-2, and subcooler). As a result, Eq. (1) indicates the evaporator's heat transfer rate.

$$\dot{Q}_{evap} = (\dot{m}_{ef}.c_{sp})_{evap}(T_{in,evap} - T_{out,evap}) = \dot{m}_1(h_6 - h_5) \quad (1)$$

$\dot{Q}_{\text{evap}}$ can be written as shown below in Eq. (2), in terms of effectiveness (ϵ), minimum heat capacity (C_{min}), and the temperature difference.

$$\dot{Q}_{evap} = (\epsilon C)_{evap}(T_{in,evap} - T_{evap}), \text{ where } T_{evap} = T_5 \quad (2)$$

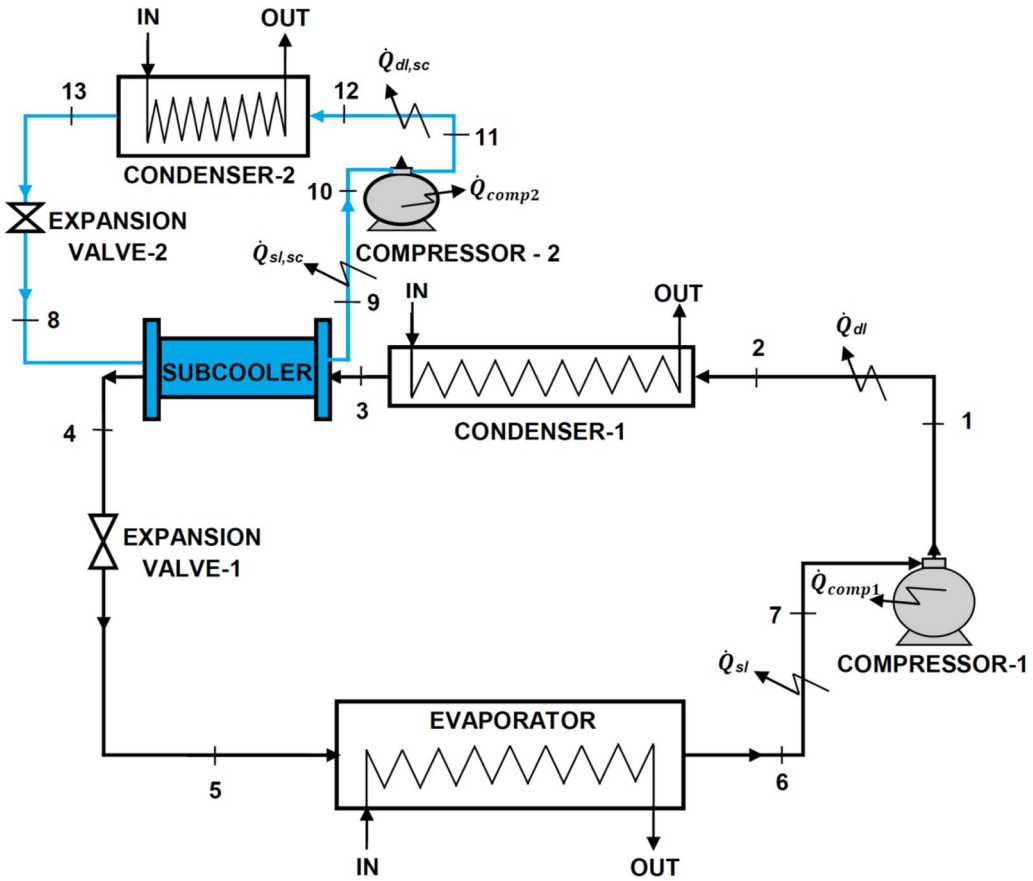


Fig. 1 Schematic diagram of DMS-VCR system. The subcooling auxiliary cycle (8–9–10–11–12–13) is attached well with the main cycle (1–2–3–4–5–6–7) at the condenser-1 outlet. These components of the two cycles (main and subcooler) are attached through a piping system in a folded loop with a provision for heat transfer to the environment. This designed system can use the same or different refrigerants in two cycles

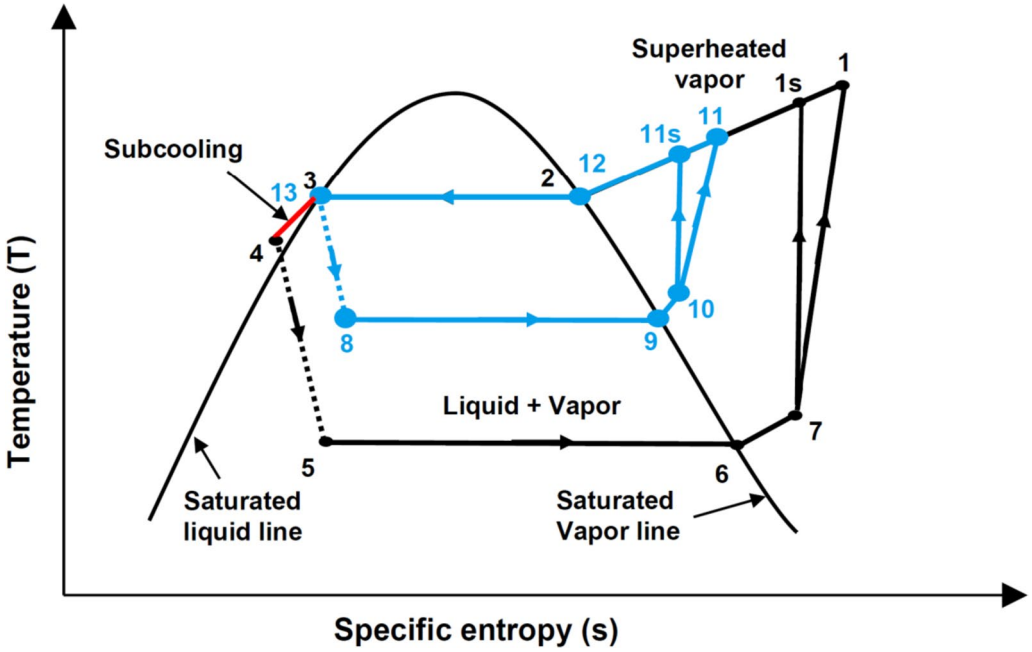


Fig. 2 Temperature-entropy diagram

The temperature of the refrigerant entering the evaporator is T_5, and the temperature of the outside fluid entering the evaporator is $T_{in,evap}$.

The heat transfer rates in condenser-1 and condenser-2 are also stated in the equations below.

Condenser-1

$$\dot{Q}_{cond1} = (\epsilon C)_{cond1}\left(T_{cond1} - T_{in,cond1}\right), where\ T_{cond1} = T_3 \tag{3}$$

$$\dot{Q}_{cond1} = \dot{m}_1(h_2 - h_3) = (\dot{m}_{ef}.c_{sp})_{cond1}(T_{out,cond1} - T_{in,cond1}) \tag{4}$$

Here, T_3 is the temperature of the saturated liquid refrigerant exiting the condenser-1, and $T_{in,cond1}$ is the temperature of the outside fluid approaching the condenser to cool the refrigerant in the condenser-1.

Condenser-2

$$\dot{Q}_{cond2} = (\epsilon C)_{cond2}\left(T_{cond2} - T_{in,cond2}\right), where\ T_{cond2} = T_{13} \tag{5}$$

$$\dot{Q}_{cond2} = \dot{m}_2(h_{12} - h_{13}) = (\dot{m}_{ef}.c_{sp})_{cond2}(T_{out,cond2} - T_{in,cond2}) \tag{6}$$

T_{13} represents the temperature of the saturated liquid refrigerant exiting the condenser-2, whereas $T_{in,cond2}$ represents the temperature of the outside fluid entering the condenser to cool the refrigerant in the condenser-2.

The heat transfer rate in the subcooler portion can be calculated using the formulas below.

$$\dot{Q}_{SC} = \in_{sc}.\dot{Q}_{SC,max} = \in_{sc}.\dot{m}_1.c_{sp}.(T_3 - T_8) \tag{7}$$

And an energy balance at the subcooler section provides the following relation.

$$\dot{m}_2(h_9 - h_8) = \dot{m}_1(h_3 - h_4) \tag{8}$$

where $T_9 = T_4 - \Delta T_{overlap}$ and T_9 are the saturated vapor temperature (sub-cooler section).

Equations (9, 10, 11 and 12) demonstrate the necessary power for the main cycle and subcooling auxiliary cycle compressors in terms of the isentropic efficiency [12, 13].

Compressor-1

$$\dot{W}_{comp1} = \frac{\dot{W}_{comp1,isen}}{\eta_{isen1}} \tag{9}$$

where $\dot{W}_{comp1,isen} = \dot{m}_1(h_{1s} - h_7)$ and $\eta_{isen1} = 0.85\text{-}0.046667 \times (\frac{P_{cond1}}{P_{evap}})$

$$\dot{Q}_{comp1} - \dot{W}_{comp1} = \dot{m}_1(h_1 - h_7) \text{ and } \dot{Q}_{sl} = \dot{m}_1(h_7 - h_6) \tag{10}$$

Compressor-2

$$\dot{W}_{sc} = \frac{\dot{W}_{sc,isen}}{\eta_{isen2}} \tag{11}$$

$$where\ \dot{W}_{sc,isen} = \dot{m}_2(h_{11s} - h_{10}) \text{ and } \eta_{isen2} = 0.85\text{-}0.046667 \times \left(\frac{P_{cond2}}{P_{sc}}\right)$$

$$\dot{Q}_{comp2} - \dot{W}_{sc} = \dot{m}_2(h_{11} - h_{10})\ and\ \dot{Q}_{sl,sc} = \dot{m}_2(h_{10} - h_9) \tag{12}$$

The isentropic condition of refrigerant vapor at the compressor's exit is represented by points 1 s and 11 s. The steady-state flow equation is used to express the compressor's work input. The heat transfer rates from the compressors to the environment are $\dot{Q}_{cp}$ and $\dot{Q}_{sc}$ (here, for analysis purpose, the heat losses or heat gains in different components are neglected).

The refrigerant effect to compressor power ratio is known as the coefficient of performance (COP), which is given by the following:

$$COP = \frac{\dot{Q}_{evap}}{\dot{W}_{total}},\ where\ \dot{W}_{total} = \dot{W}_{comp1} + \dot{W}_{sc} \tag{13}$$

Fan power consumption is very low in commercial and residential systems. So, it is not considered in the calculation of COP.

Exergy analysis is the most accurate way for calculating actual losses owing to irreversibility in the process. Irreversibility ($\dot{I}$) is the difference between reversible and useful work (W_{useful}) during the process and is expressed as a potential reduction in work.

For a particular system component, thermodynamic modeling includes the below-mentioned governing Eq. (14) [14, 15]:

$$\text{Exergy balance, } \dot{I} = E^{in} - E^{out} + E^{Q} - E^{W} \tag{14}$$

Here, $\dot{I}$ is the total irreversibility or exergy destruction (loss), and E^Q and E^W are the exergy flow related to heat and work transfer respectively [15]. In $E = \dot{m} * \psi$, ψ is the chemical, potential, kinetic, and physical exergy of matter, and E^{in} and E^{out}, which are matter exergies at the inlet and outlet of the control volume. The physical specific exergy is calculated after removing chemical, potential, and kinetic exergies.

$$\psi = (h - T_o s) - (h_o - T_o s_o) \tag{15}$$

$$\dot{I} = \sum \dot{E}_{in} - \sum \dot{E}_{out} + \sum \dot{Q}\left(1 - \frac{T_o}{T}\right) - \sum \dot{W} \tag{16}$$

where T_o, refers to the temperature of the reference environment.

By employing the above fundamental equations to all the elements of the DMS-VCR system, the system's entire irreversibility is indicated by the equation below (17).

$$\dot{I}_{total} = \dot{I}_{evap} + \dot{I}_{cond1} + \dot{I}_{cond2} + \dot{I}_{sc} + \dot{I}_{comp1} + \dot{I}_{comp2} + \dot{I}_{ev1} + \dot{I}_{ev2} \tag{17}$$

The exergetic efficiency of DMS-VCR System is given by Eq. (18) below.

$$\eta_{II} = \left(1 - \frac{\dot{I}_{total}}{\dot{W}_{total}}\right) \tag{18}$$

3.2 Environment analysis

Many environmentalists contend that global warming problems can be resolved by making CO_2 emissions subject to a price on carbon [11]. The " annual penalty cost" is the CO_2 emissions into the atmosphere caused by the production of power from fossil fuels. Equation (19) can be used to calculate this cost.

$$C_{env} = \left(\frac{m_{co_2}}{1000}\right) \times C_{co_2} \tag{19}$$

Here,C_{co_2}, which varies among nations, is used to represent the cost of the CO_2 emissions penalty. According to [11] and [16], the value ranges from 22 to 117 $/tons. The penalty cost of CO_2 emissions is assumed in the current work to be US $100 per tons of CO_2 emission. Additionally, mass of CO_2 emission (m_{co_2}) is given by Eq. (20):

$$m_{co_2} = \lambda \times \dot{W}_{total} \times t_{op} \tag{20}$$

where the DMS-VCR system's annual operational hours (t_{op}) and the conversion factor of emissions (λ) are assumed to be 5000 h and 0.968 kg/kWh, respectively.

4 Model validation

Numerical analysis constituted the foundation for the original work on the DMS-VCR system. Using the work of [17], the existing model is verified. The condenser-1 and condenser-2 temperatures are 50 °C, the net refrigerating effect (Q_{evap}) is 3.5167 kW, the evaporator-1 temperature (T_{evap}, DMS) is − 10 °C, there is up to 5 °C of subcooling, the compressors' isentropic efficiency (η_{comp1}, η_{comp2}) is 80%, the subcooler's efficiency ($\in_{sc}$) is 0.8, and the ambient temperature (T_o) and pressure (P_o) are 25 °C and 101.325 kPa, respectively. The simulation software is modified in response to the model validation process to make the simulation results consistent with the findings of theoretical study. Here, model validation is being used to assess a DMS-VCR system on the EES program. The results demonstrate that the EES model accurately predicts all significant quantities (compressor power, exergetic efficiency, and COP), with a maximum

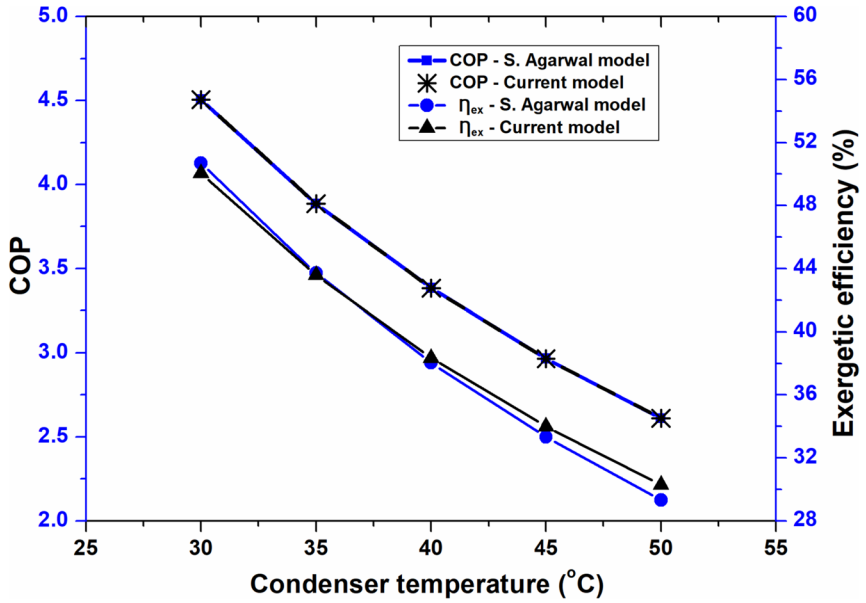


Fig. 3 Performance comparison of current model with [17] for the effect of condenser temperature on refrigerant R134a ($T_{evap} = -10$ °C)

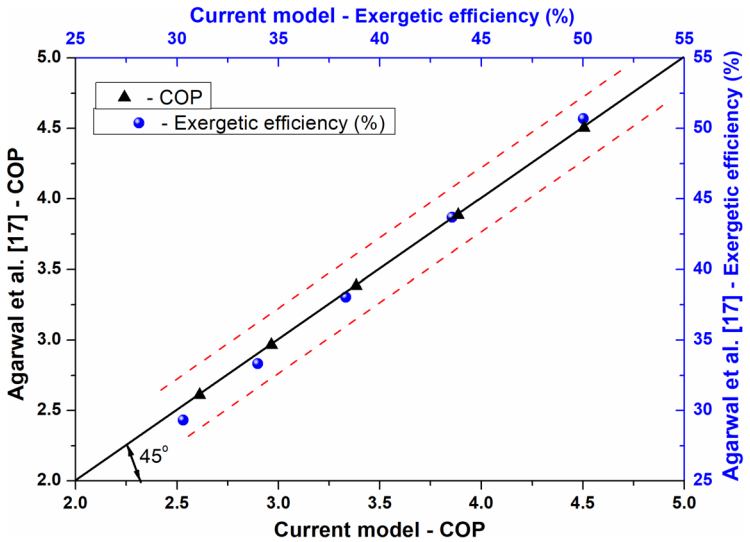


Fig. 4 Accuracy performance analysis on the test set of COP and exergetic efficiency

encountered error of 4.14% and the majority of errors being less than 2% (see Fig. 3).

Figure 4 illustrates a detailed analysis of the accuracy performance on the test set of COP and exergetic efficiency in comparison to the results presented in [17]. The *X*-axis displays COP and exergetic efficiency values obtained from current model's simulations, while the *Y*-axis represents the corresponding results reported by [17]. The proximity of data points to this trend line indicates the similarity of current model's performance to the benchmark set by [17]. This graph serves as a valuable tool for evaluating the similarity of current model's results with the established findings in the literature, aiding in the assessment of its accuracy and reliability within the context of COP and exergetic efficiency analysis.

5 Results and discussion

This study involves solving nonlinear mathematical model equations on an EES software using R134a as the working fluid in both the subcooler and main cycles. When compared to an equivalent VCRS, the DMS-VCR

system power consumption is reduced by 15.52%, and its COP rises by 9.5%. Prior to beginning the assessment of the mathematical model's performance from energy, exergetic, and environmental point of view, a mathematical model validation was completed. Numerous researchers advised a 5 °C to 10 °C of subcooling in a thorough literature review on the topic. The overall COP and exergetic efficiency rise up to 18 °C, but drops beyond that, when the degree of subcooling is varied from 1°C to 25 °C keeping other performance parameters constant. As a result, for the DMS-VCR system to work at its best, the abovementioned optimum value is advised. The following set of input conditions are used to obtain desired results with the DMS-VCR system (see Table 1).

5.1 Energetic analysis results

For the same cooling capacity, it is seen that the mass flow rate of the external fluid (water) and refrigerant (R134a) at the condenser-1 section of the DMS-VCR system arrangement is lowered by approximately 16% and 17%, respectively. Due to heat rejection in the subcooler section at 17 °C, the mass flow rate of refrigerant decreases. In comparison to an equivalent VCRS setup, the DMS-VCR system configuration's lower mass flow rate of refrigerant reduces the heat load in condenser-1 by 15.5%. Additionally, a reduced heat load in the condenser results in a lower mass flow rate of water needed to cool it. When compared to an equivalent VCRS system, the compressor-1's power consumption is lowered by 15.51% as a result of the DMS-VCR system's lower mass flow rate of refrigerant. It is determined that the equivalent VCRS and DMS-VCR systems have overall COPs of 3.58 and 3.92, respectively (see Table 2).

5.2 Exergetic analysis results

Table 3 shows how the DMS-VCR system and the equivalent VCRS differ qualitatively when using an exergy analysis approach. The irreversibility rate of each system component is determined by using irreversibility Eq. (16) [15]. The non-isentropic compression of compressor-1 causes a higher irreversibility rate in this system. Condensers and evaporator contribute significantly to the total irreversibility rate because of the temperature difference between the heat source and the working fluid. For VCRS and DMS-VCR system, the overall irreversibility rate is 17.93 and 15.69, respectively. The DMS-VCR system has a higher exergetic efficiency (i.e., 38.58%) than the equivalent VCRS (i.e., 35.74%), which measures the degree of thermodynamic performance.

Table 2 Comparative energy performance analysis of VCRS and DMS-VCR system

S. No.	Performance parameters		VCRS	DMS-VCRS
1	First law parameters	COP_{VCRS}	3.58	-
		$COP_{DMS\text{-}VCRS}$	-	3.92
		$\dot{Q}_{cond1}$(kW)	134.2	113.4
		$\dot{Q}_{cond2}$(kW)	-	18.47
		$\dot{Q}_{evap}$(kW)	100	100
		$\dot{Q}_{sc}$(kW)	-	15.51
		$\dot{W}_{comp1}$(kW)	27.9	23.57
		$\dot{W}_{comp2}$(kW)	-	1.97
		$\dot{m}_{ref,1}$(kg/s)	0.70	0.59
		$\dot{m}_{ref,2}$(kg/s)	-	0.10
		$\dot{m}_{ef,ev}$ (kg/s)	4.76	4.76
		$\dot{m}_{ef,cd1}$ (kg/s)	6.42	5.42
		$\dot{m}_{ef,cd2}$ (kg/s)	-	0.88

Table 3 Comparative exergy performance analysis of VCRS and DMS-VCR system

S. No.	Performance parameters		VCRS	DMS-VCRS
1	Second law parameters	$\dot{I}_{cond,1}$(kW)	3.92	3.31
		$\dot{I}_{cond,2}$(kW)	-	0.48
		$\dot{I}_{comp,1}$(kW)	7.80	6.58
		$\dot{I}_{comp,2}$(kW)	-	0.43
		$\dot{I}_{evap}$(kW)	2.85	2.85
		$\dot{I}_{ev,1}$(kW)	3.36	1.11
		$\dot{I}_{ev,2}$(kW)	-	0.17
		$\dot{I}_{sc}$(kW)	-	0.73
		$\dot{I}_{total}$(kW)	17.93	15.69
		η_{II} (%)	35.74	38.58

5.3 Environment analysis results

The environmental impact of the DMS-VCR system must be taken into account due to growing environmental concerns, especially those pertaining to problems with global warming brought on by greenhouse gases. In this way, the environmental impact of DMS-VCR systems is assessed using CO_2 emissions. Figure 5 provides a visual representation of the carbon dioxide (CO_2) emissions from two different systems: the DMS-VCR system and an equivalent VCRS. The DMS-VCR system emits approximately 114 tons of CO_2 annually, while the equivalent VCRS emit around 135 tons of CO_2 per year. It demonstrates that by using the DMS-VCR system instead of an equivalent VCRS, there is a significant reduction in carbon dioxide emissions. Specifically, the DMS-VCR system emits approximately 21 tons less CO_2 per year compared to the equivalent VCRS. It emphasizes the

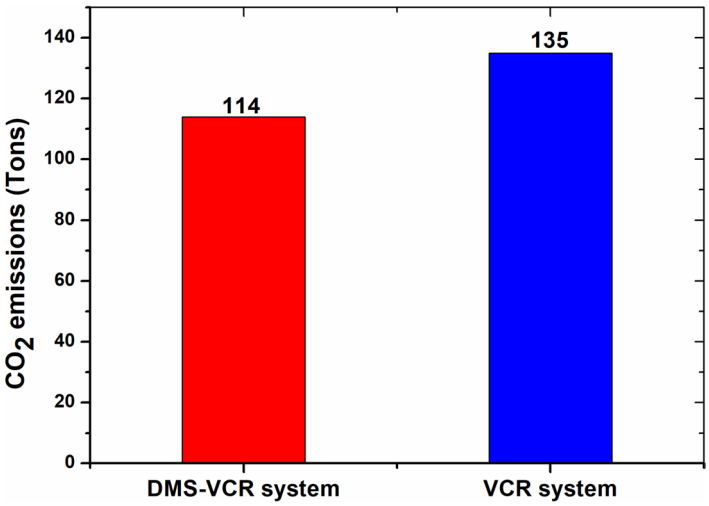


Fig. 5 Carbon-dioxide emission from DMS-VCR system and equivalent VCR system

Table 4 Environmental cost for different countries

S. No.	Country	Co_2 penalty cost (per ton of CO_2 emission)	Environmental cost of VCRS system	Environmental cost of DMS-VCR system
1	Uruguay	US $137	US $18,495	US $15,618
2	India	Rs. 400	Rs. 54,000	Rs. 45,600
3	Japan	¥ 289	¥ 39,015	¥ 32,946
4	New Zealand	NZ$ 30.83	NZ$ 4162	NZ$ 3515
5	Singapore	S$ 5	S$ 675	S$ 570
6	Republic of Korea	US $27.62	US $3729	US $3149
7	China	RMB 51.23	RMB 6916	RMB 5840

importance of making more sustainable choices in technology and equipment to reduce the carbon footprint and mitigate the effects of climate change.

The compressor uses less energy as a result; it considerably lowers the CO_2 emissions. The annual penalty costs for the corresponding VCRS and DMS-VCR systems are US $13,503 and US $11,407, respectively. C_{CO2} is a measure of the penalty cost of CO_2 emissions, and it varies by countries (see Table 4). According to researchers [11, 16], the price per tons ranges from 22 to 117 dollars. The penalty cost of CO_2 emissions is assumed to be US $100 per tons of CO_2 emission. To provide emitters with a grace period to implement required changes, Singapore's national climate change secretariat sets the carbon price at US $5 per tons of greenhouse gas emissions (tCO_2e—tons of CO_2 emission) from 2019 to 2023. In order to achieve climate objective, the carbon tax rate will be raised to US $25/$tCO_2e$ in 2024 and 2025, US $45/$tCO_2e$ in 2026 and 2027, and US $50–80/$tCO_2e$ by 2030 [18, 19]. Other countries have also planned for their carbon prices to continue until 2030. The DMS-VCR system is therefore a more promising decarbonizing cooling system than a simple VCRS.

5.4 Parametric analysis

5.4.1 Effect of evaporator and condenser temperature

Figure 6 illustrates the impact of evaporator temperature on key parameters of the DMS-VCR system, including total irreversibility, exergetic efficiency, and overall coefficient of performance. It is worth noting that the external fluid temperatures for both the condenser and evaporator should remain within a narrow 5 °C range (e.g., with the external fluid entering the evaporator at 10 °C and exiting at 5 °C, and similarly for the condenser, where it ranges from 30 °C to 35 °C). The approaching temperature of the leaving water for the heat exchangers that are currently in use is set at 5 °C. During this analysis, the evaporator temperature of the DMS-VCR system was varied between −5 °C and 5 °C, while keeping other design parameters, as well as the input and output temperatures of the external fluid constant.

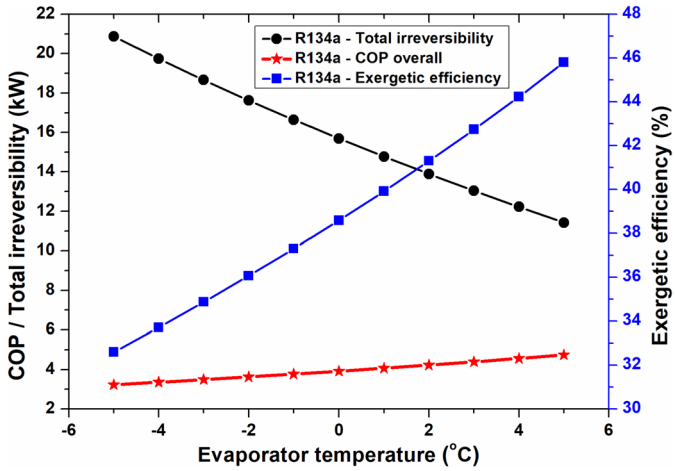


Fig. 6 Effect of evaporator temperature on total irreversibility, exergetic efficiency, and COP

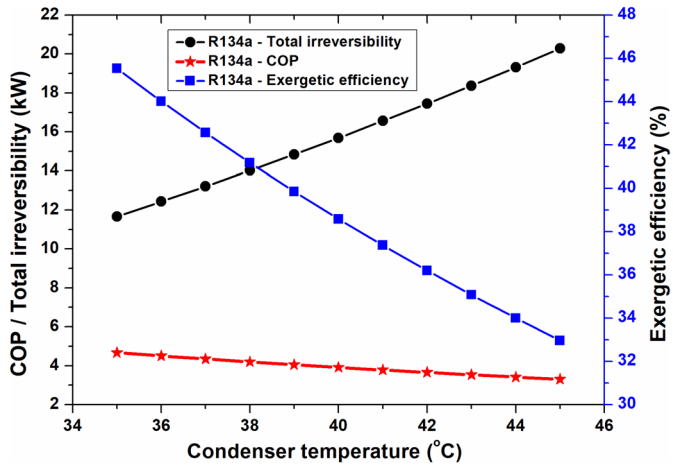


Fig. 7 Effect of condenser temperature on total irreversibility, exergetic efficiency, and COP

An increase in evaporator pressure from 243.5 kPa to 349.9 kPa resulted in a significant 32% reduction in power consumption within the compression section. This reduction can be attributed to the decrease in the specific volume of the refrigerant. Consequently, the system's overall COP increased from 3.23 to 4.75. Furthermore, the overall irreversibility of the system decreased from 20.87 kW to 11.42 kW. This reduction in irreversibility can be solely attributed to a decrease in the thermal gradient between the evaporator and the external fluid, which is brought about by the increase in evaporator temperature. From an energetic and exergetic perspective, a higher evaporator temperature is deemed necessary. This is not only advantageous from an environmental standpoint due to the decreasing trend in electricity consumption, resulting in reduced CO_2 emissions, but also it has the effect of reducing the annual penalty cost.

Figure 7 provides a detailed visual representation of several critical performance parameters of the DMS-VCR system, including the COP, total irreversibility, and exergetic efficiency. This also explores how these parameters are interconnected with changes in the condenser temperature. The DMS-VCR system's condenser temperature is intentionally adjusted across a range from 35 °C to 45 °C while maintaining constant values for inlet and outlet temperatures and other crucial design parameters. Upon careful examination of the data, it becomes evident that as the condenser temperature increases, there is a notable rise in the compressor work required. Specifically, there is a 41.38% increase in compressor work when the condenser pressure varies from 887.5 kPa to 1161 kPa. This increased compressor work results in higher emissions of CO_2 into the atmosphere, which is clearly undesirable from an environmental perspective.

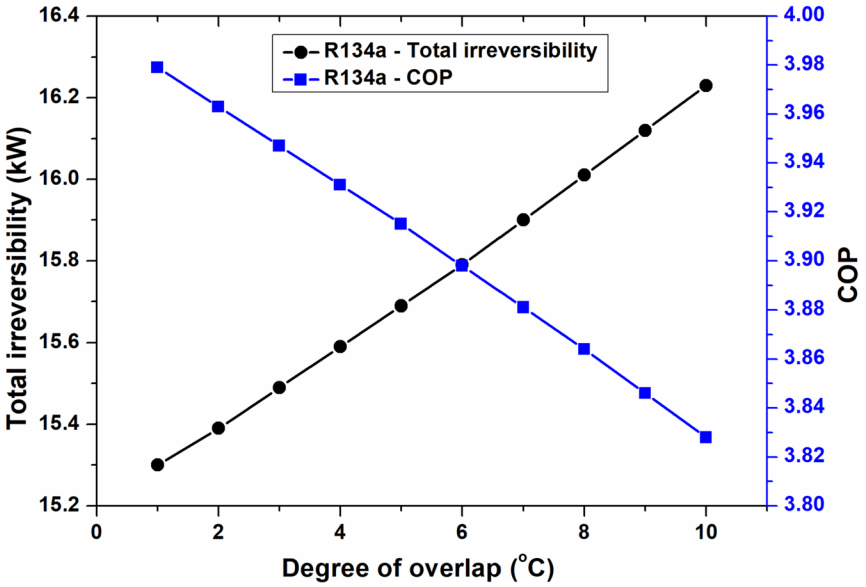


Fig. 8 Effect of overlap degree on overall irreversibility and COP rate of the system

As a direct consequence of this increase in condenser temperature and the associated rise in CO_2 emissions, the system's COP experiences a decrease. In simpler terms, the system becomes less energy efficient as the condenser temperature rises. Furthermore, Fig. 7 reveals a substantial increase in the overall irreversibility of the system, climbing from 11.66 kW to 20.29 kW as the condenser temperature increases. This increase in irreversibility signifies a loss of efficiency and an increase in wasted energy within the system, further emphasizing the negative implications of higher condenser temperatures. When evaluating the DMS-VCR system from various angles, including energy efficiency, exergetic efficiency, and environmental impact, the data presented in Fig. 7 does not support the use of high condenser temperatures.

5.4.2 Effect of degree of overlap

The difference in temperature between the fluid exiting the main cycle (state 4) after subcooling and the fluid exiting the subcooler unit (state 9) from the subcooler is referred to as the degree of overlap. Figure 8 investigates how varying the degree of overlap, ranging from 1 °C to 10 °C, affects the effectiveness of the subcooler section. Effectiveness here represents the efficiency of heat transfer, and it plays an essential role in determining the overall performance of the system. As the degree of overlap increases, moving from 1°C to 10 °C, the effectiveness of the subcooler section decreases notably, dropping from 95% to 64%. This decline in effectiveness suggests that the subcooler becomes less efficient in its heat transfer function as the degree of overlap increases. Several factors contribute to this reduced efficiency of the subcooler. These include limitations in its heat transfer capacity, the temperature approach effect, inadequate subcooling, and reduced surface area available for subcooling. The effectiveness of heat exchangers remains unchanged in case of evaporator and condenser-1, while it increases by 8.35% in case of condenser-2 and decreases by 32.14% in subcooler. This results in a 5.4% increase in heat load over the subcooler cycle condenser (condenser-2) by keeping the subcooler temperature and other design factors constant.

In terms of the overall system performance, as the subcooler cycle compressor work increases by nearly 63%, the total irreversibility experiences a 6.1% increase. This, in turn, leads to a decrease in the system's overall COP from 3.98 to 3.82 (see Fig. 8). The main compressor section's performance remains unchanged in this context. Ultimately, the figure emphasizes the critical importance of balancing the degree of overlap while considering various factors, including the system's needs, refrigerant characteristics, cooling medium properties, and heat exchanger design. To ensure efficient subcooling and achieve the desired performance in the DMS-VCR system, it is essential to carefully design and size the subcooler section. From both an energy and exergy standpoint, it is advisable to aim for moderate overlap values to maintain optimal system performance.

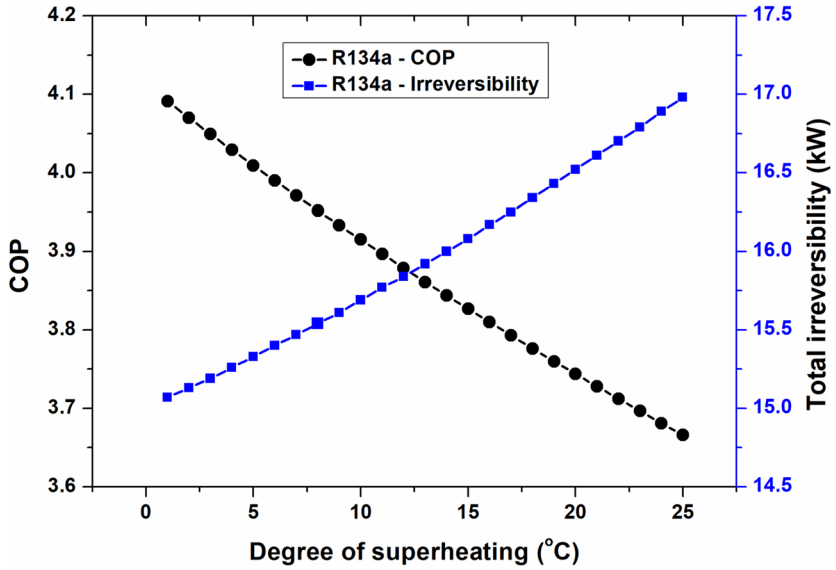


Fig. 9 Variation of COP and total irreversibility with degree of superheating

6 Effect of degree of superheating and subcooling

In a DMS-VCR system, superheating refers to the critical process of raising the temperature of the refrigerant vapor over its saturation point. This occurs as a result of the refrigerant's ability to absorb heat even after it has completely vaporized. In order to protect the system's components from possible harm, superheating is essential for maintaining the consistent and effective operation of the system. Additionally, superheating improves the performance of the entire system by raising the temperature of the refrigerant as it leaves the evaporator, a step required for cooling items or the surrounding environment. Thus, it indirectly helps in the compressor's lifetime and safety. Figures 9 and 10 offer a detailed analysis of how the degree of superheating affects various crucial parameters within a DMS-VCR system, including the COP, compressor work, irreversibility, and exergetic efficiency. The figures illustrate trends that emerge as the degree of superheat varies across a range from 1 °C to 25 °C. One key finding is that as the degree of superheat

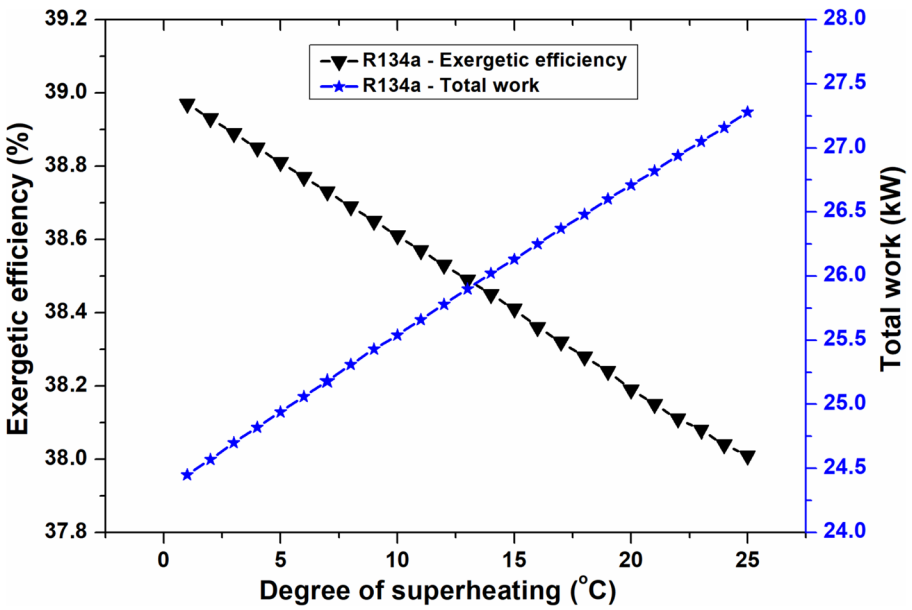


Fig. 10 Variation of exergetic efficiency and total work with degree of superheating

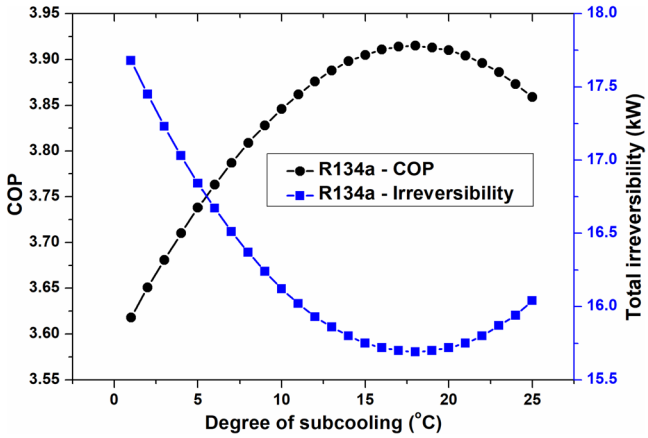


Fig. 11 Variation of COP and total irreversibility with degree of subcooling

increases, it demands more work from the compressor. This higher compressor work requirement is significantly greater than the increase in cooling capacity, resulting in a combined effect that leads to a decrease in the overall COP of the system. In practical terms, this means that superheating the refrigerant vapor within a DMS-VCR system tends to reduce the COP, causing it to drop from 4.09 to 3.66. Additionally, the compressor work increases notably, by approximately 11.57%.

Furthermore, as the degree of superheat increases, there is an associated rise in irreversibility by approximately 12.67%. This increase in irreversibility leads to a decrease in exergetic efficiency, dropping from 38.37% to 37.74%. Given these implications from both an energy and exergy perspective, it is advisable to minimize the degree of superheat within the system. To mitigate these effects and optimize the performance of the DMS-VCR system, it is recommended that manufacturers and design engineers prioritize the use of quality insulating packing around the pipeline. This measure can help maintain lower superheat levels, enhancing system efficiency and reducing energy consumption.

The degree of subcooling has a significant effect on the effectiveness, exergetic efficiency, and overall cooling capacity of the system. Engineers and technicians can increase the cooling efficiency of the system and guarantee its optimal performance by optimizing the degree of subcooling. The degree of subcooling is varied from 1 °C to 25 °C keeping the other design parameters constant (see Table 1). Through various simulations, it has been concluded that the ideal degree of subcooling is 18 °C. According to Figs. 11 and 12, at this subcooling level, the system's overall COP and exergetic efficiency increase, while its total irreversibility rate and total compressor work decrease. Below or above the optimal value of subcooling, the reverse patterns are observed.

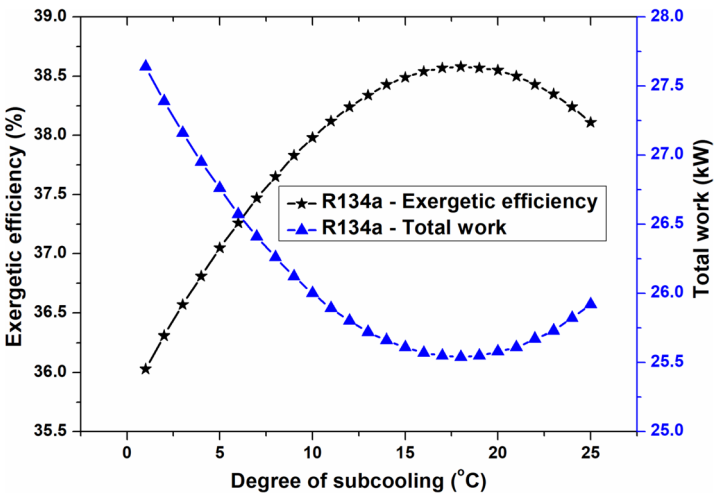


Fig. 12 Variation of exergetic efficiency and total work with degree of subcooling

After numerous simulations on EES, 18 °C is determined to be the optimal degree of subcooling on varying it up to 25 °C because, up to this level, COP and exergetic efficiency are increasing, but beyond this, a decreasing pattern is observed (see Figs. 11 & 12), so this is considered as the optimal degree of subcooling in this study. By identifying and addressing sources of irreversibility, we have concentrated on reducing exergy destruction inside the system. Changes have been made to reduce exergy losses associated with various components, such as valves, compressors, and heat exchangers (evaporator and condenser), by thorough exergy analysis. As a result, the system's total exergetic efficiency has significantly increased. The improvement in COP and exergetic efficiency shown in Figs. 11 and 12 has significant implications. Up to ideal degree of subcooling, the DMS-VCR system compressor-1 work decreases from 27.61 kW to 23.57 kW, whereas in subcooler system compressor work increases from 0.029 kW to 1.97 kW, resulting in a decrease in the total compressor work from 27.64 kW to 25.54 kW. It increases the system's overall COP from 3.62 to 3.92 (an improvement of 8.29%). As a result, the system's entire irreversibility drops by 11.26%.

A higher COP indicates better exergetic efficiency and less energy use, which is economical and benefits the environment. The optimal degree of subcooling is determined by the overall COP, exergetic efficiency, irreversibility, and total work. Hence, optimum level of subcooling basically improves DMS-VCR system performance because going above this limit increases the system's irreversibility and decreases COP, which is an add-on to the originality of the work.

7 Conclusions

In conclusion, this paper presents a comprehensive analysis of a dedicated mechanical subcooled vapor compression refrigeration (DMS-VCR) system using R134a as the working fluid in both the subcooler and main cycles. The study primarily focuses on energy, exergy, and environmental (3E) parameters, assessing the system's performance and environmental impact. The DMS-VCR system shows promising results, significantly reducing electricity usage by 15.52% and increasing the COP by 9.5%. This improved efficiency results in a substantial reduction in CO_2 emissions by approximately 15.55% compared to an equivalent VCRS. Exergetic analysis reveals that the DMS-VCR system also exhibits an 8% increase in exergetic efficiency, highlighting its enhanced thermodynamic performance. Furthermore, the study emphasizes the importance of key design parameters, such as the degree of overlap, superheating, and subcooling. Analyzing these parameters is crucial for achieving the system's optimal performance. The research identifies an ideal degree of subcooling at 18 °C, leading to increased COP, exergetic efficiency, and reduced irreversibility within the system.

In terms of environmental impact, the DMS-VCR system demonstrates a significant advantage, emitting approximately 21 tons less CO_2 annually compared to an equivalent VCRS. The yearly penalty cost of CO_2 emissions in this analysis for DMS-VCR system is roughly 16% lower than it would be in an equivalent VCRS; this variation in results is because of decrease in compressor power consumption. The paper also discusses the economic aspects, considering penalty costs associated with CO_2 emissions in different countries. In summary, the DMS-VCR system represents a promising solution for more efficient and environmentally friendly cooling applications. It not only enhances energy performance but also contributes to the reduction of greenhouse gas emissions, aligning with the global effort to mitigate climate change. The findings of this study provide valuable insights for design engineers and policymakers seeking sustainable cooling technology alternatives.

8 Nomenclature

C_{env}	Annual environment damage cost ($/year)
C_{co_2}	Penalty cost of CO_2 emissions ($/tons of CO_2 emission)
C	Thermal capacitance rate (kW/K)
C_{sp}	Specific heat (kJ/kg.K)
E	Exergy flow
h	Specific enthalpy (kJ/kg)
$\dot{I}$	Rate of irreversibility (kW)
$\dot{m}$	Flow rate of mass (kg/s)
m_{co_2}	Annual emission of carbon dioxide (ton)
P	Pressure (kPa)
PR	Pressure ratio between high- and low-pressure side
$\dot{Q}$	Heat transfer (kW)
s	Specific entropy (kJ/kg.K)
T	Temperature (K)
t_{op}	Per-year operation period (h)
$\dot{W}$	Power input rate (kW)

Greek symbols

$\in$	Effectiveness
η	Efficiency
Ψ	Chemical, potential, kinetic, and physical exergy of matter
λ	Conversion factor of CO_2 emission (kg/kWh)

Superscripts

in	Inlet
out	Outlet
Q	Heat
W	Work

Subscripts

ref	Refrigerant
o	Environment condition
dl	Discharge line
comp	Compressor
cond	Condenser
ef	External fluid
ev	Expansion valve
evap	Evaporator
isen	Isentropic
sc	Subcooling/subcooler
sl	Suction line
II	Second law
1, 2, 3….	State point

Abbreviations

COP	Coefficient of performance
DMS	Dedicated mechanical subcooled/subcooler
DOS	Degree of subcooling
DOO	Degree of overlap
EES	Engineering equation solver
VCRS	Vapor compression refrigeration system
VCRC	Vapor compression refrigeration cycle

Authors' contributions

Naveen Solanki (Corresponding Author): Coding for the EES software. Writing and drafting the paper. Proofreading and finalizing the manuscript. Formulating and applying mathematical equations. Conducting the result and discussion analysis. Validating the research outcomes. Akhilesh Arora: Assisting in coding for the EES software. Contributing to the writing and drafting of the paper. Collaborating on proofreading and finalizing the manuscript. Participating in the formulation and application of mathematical equations. Engaging in result and discussion analysis. Raj Kumar Singh: Providing support in coding for the EES software. Offering insights for writing and drafting the paper. Participating in proofreading and finalizing the manuscript. Contributing to the formulation and application of mathematical equations. Collaborating on result and discussion analysis. We believe that these contributions encompass the various aspects of the research process, ensuring a comprehensive and collaborative effort from all authors.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 25 July 2023 Accepted: 7 November 2023
Published online: 29 November 2023

References

1. Catalán-Gil, J., Llopis, R., Sánchez, D., Nebot-Andrés, L., & Cabello, R. (2019). Energy analysis of dedicated and integrated mechanical subcooled CO 2 boosters for supermarket applications. *International Journal of Refrigeration, 101*, 11–23.
2. Qureshi, B. A., & Zubair, S. M. (2012). The effect of refrigerant combinations on performance of a vapor compression refrigeration system with dedicated mechanical sub-cooling. *International Journal of Refrigeration, 35*(1), 47–57.
3. Qureshi, B. A., & Zubair, S. M. (2013). Mechanical sub-cooling vapor compression systems: current status and future directions. *International Journal of Refrigeration, 36*(8), 2097–2110.
4. Qureshi, B. A., Inam, M., Antar, M. A., & Zubair, S. M. (2013). Experimental energetic analysis of a vapor compression refrigeration system with dedicated mechanical sub-cooling. *Applied Energy, 102*, 1035–1041.
5. Selbaş, R., Kizilkan, Ö., & Şencan, A. (2006). Thermoeconomic optimization of subcooled and superheated vapor compression refrigeration cycle. *Energy, 31*(12), 2108–2128.
6. She, X., Yin, Y., & Zhang, X. (2014). A proposed subcooling method for vapor compression refrigeration cycle based on expansion power recovery. *International Journal of Refrigeration, 43*, 50–61.
7. Koeln, J. P., & Alleyne, A. G. (2014). Optimal subcooling in vapor compression systems via extremum seeking control: Theory and experiments. *International Journal of Refrigeration, 43*, 14–25.
8. Pottker, G., & Hrnjak, P. (2015). Experimental investigation of the effect of condenser subcooling in R134a and R1234yf air-conditioning systems with and without internal heat exchanger. *International Journal of Refrigeration, 50*, 104–113.
9. Llopis, R., Toffoletti, G., Nebot-Andrés, L., & Cortella, G. (2021). Experimental evaluation of zeotropic refrigerants in a dedicated mechanical subcooling system in a CO2 cycle. *International Journal of Refrigeration, 128*, 287–298.
10. Agarwal, S., Arora, A., & Arora, B. B. (2020). *Exergy analysis of dedicated mechanically subcooled vapour compression refrigeration cycle using HFC-R134a, HFO-R1234ze and R1234yf*. Singapore: Springer.
11. Aminyavari, M., Najafi, B., Shirazi, A., & Rinaldi, F. (2014). Exergetic, economic and environmental (3E) analyses, and multi-objective optimization of a CO2/NH3 cascade refrigeration system. *Applied Thermal Engineering, 65*(1–2), 42–50.
12. Sayyaadi, H., & Nejatolahi, M. (2011). Multi-objective optimization of a cooling tower assisted vapor compression refrigeration system. *International Journal of Refrigeration, 34*(1), 243–256.
13. Jain, V., Sachdeva, G., & Kachhwaha, S. S. (2015). NLP model based thermoeconomic optimization of vapor compression-absorption cascaded refrigeration system. *Energy Convers. Manag., 93*, 49–62.
14. Arora, A., Arora, B. B., Pathak, B. D., & Sachdev, H. L. (2007). Exergy analysis of a vapour compression refrigeration system with R-22, R-407C and R-410A. *International Journal of Exergy, 4*(4), 441–454.
15. Nag, P. K. (2002). *Basic and applied thermodynamics by PK Nag: Thermodynamics by PK Nag* (p. 781)
16. Carlsson, B., Meir, M., Rekstad, J., Preiß, D., & Ramschak, T. (2016). Replacing traditional materials with polymeric materials in solar thermosiphon systems - Case study on pros and cons based on a total cost accounting approach. *Solar Energy, 125*, 294–306.
17. Agarwal, S. (2019). Thermodynamic performance analysis of dedicated mechanically subcooled vapour compression refrigeration system. *J. Therm. Eng., 5*(4), 222–233.
18. ADB. (2021). *Carbon Pricing for Green Recovery and Growth*.
19. European Union. (2020). *Carbon emissions pricing Some points of reference*. March.

Publisher's Note

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

MAPAN-Journal of Metrology Society of India
https://doi.org/10.1007/s12647-023-00694-2

CASE REPORT

$PM_{2.5}$ Observations of Outdoor and Indoor Environment at Dwarka, New Delhi, India

S. Malik[1]*, A. S. Rao[1] and S. K. Dhaka[2]

[1]Department of Applied Physics, Delhi Technological University, New Delhi, India

[2]Radio and Atmospheric Physics Lab, Rajdhani College, University of Delhi, New Delhi 110015, India

Received: 06 June 2023 / Accepted: 07 October 2023

© Metrology Society of India 2023

Abstract: Measurement of $PM_{2.5}$ is carried out simultaneously outdoors as well as indoors using a Compact and Useful $PM_{2.5}$ Instrument (CUPI) from Nov 2018 to Oct 2019. A high-time resolution data ($\sim$ 2 min interval) was recorded at Dwarka (West Delhi), New Delhi, to investigate the difference between outdoor and indoor ambient particulate matter. Year-long observations provide an opportunity to decipher that indoor $PM_{2.5}$ concentration increases significantly from mid-November 2018 to mid-December 2018. During this high concentration occurrence of $PM_{2.5}$, outdoor concentration was higher in the range of 100–200 $\mu g/m^3$ than indoor. Indoor pollution was less by 30–50% in comparison to the outdoor environment. Analyses reveal that during a high pollution period, intermittently for a short period, outdoor $PM_{2.5}$ was less, which coincides with clear weather days suggesting that indoor pollution did not disperse due to a closed environment. Diurnal variation of $PM_{2.5}$ clearly shows high concentration stays until 12:00 noon in both indoor and outdoor environments, which reduces in the afternoon and again picks up in the evening. Diurnal variation of solar radiation showed a consistent increase from 9 am to 2 pm, while relative humidity declined considerably from 10 am to 4 pm (local time); these two factors correspond to less concentration in both indoor and outdoor environments. However, there is no significant difference between outdoor and indoor concentration during summer and monsoon season.

1. Introduction

Particle pollution is an emerging area of concern as regards to human health in the present times. The instances of high particulate pollution have become quite common in Delhi-NCR region, India, especially during winter period [1]. Air pollution over north India and more specifically over Delhi has attracted both public and scholastic group. Outdoor air pollution has detrimental health effects and is responsible for more than 4 million deaths every year globally [1], resulting in substantial global and regional decrements in life expectancy. India experiences high ambient air pollution, with an annual population-weighted $PM_{2.5}$ (particulate matter with diameter less than 2.5 μm) mean of 74 $\mu g\ m^{-3}$ and experiences the highest number of deaths [2]. Submicron aerosol composition in the world's most polluted megacity from ambient air pollution among all countries in the world, Delhi (population of about 28 million), is identified the world's most polluted megacity, with recent annual-average $PM_{2.5}$ concentrations of $\sim$ 140 $\mu g\ m^{-3}$ [2, 3].

Therefore, the ambient $PM_{2.5}$ exposure to the human being is quite large that affects our daily life. The sources of $PM_{2.5}$ are present both indoors and outdoors. Indoor activities such as cooking, frying, tobacco smoke, kerosene heaters, etc. produce pollutants. Since the size range of $PM_{2.5}$ is very small, thus it permits to pave way into the respiratory tract, thereby reaching the lungs. Even its short-time exposure is capable of eye, nose, throat irritation, shortness of breath, and many more such effects. Prolonged exposure to $PM_{2.5}$ can prove hazardous as it can vary from showing mild symptoms to severe health conditions such as asthma. Hence, it is important to study $PM_{2.5}$ both indoor and outdoor, which is not done systematically partly due to the poor understanding that indoor air quality is better and safe. Indoor measurements of $PM_{2.5}$ is equally important, and there is a need to provide the real-time status of the air quality.

*Corresponding author, E-mail: shristymalik94@gmail.com

Published online: 31 October 2023

Anthropogenic emissions have been widely accepted as the dominant driver for $PM_{2.5}$ concentrations, whereas meteorological conditions also exert a strong influence on long-term $PM_{2.5}$ variations [4, 5–7]. Thus, it important to study $PM_{2.5}$ both indoors and outdoors. Influence of meteorological conditions on $PM_{2.5}$ concentrations is essential to understand its diurnal variability; in this regard, solar radiation and relative humidity analysed during winter to seek the influence on $PM_{2.5}$.

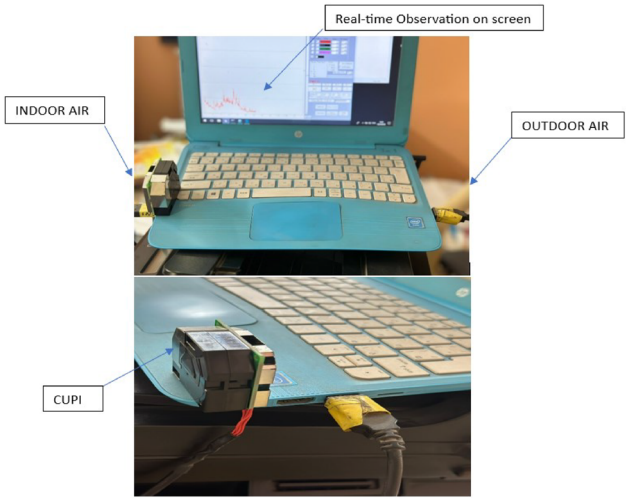
In this paper, the concentration of $PM_{2.5}$ presented both indoor as well as outdoor using a Compact and Useful $PM_{2.5}$ Instrument (CUPI) from Nov 2018 to Oct 2019. A high-time resolution data ($\sim$ 2 min interval) was recorded at Dwarka (West Delhi), New Delhi, to investigate the difference in the concentration of ambient particulate matter outdoor as well as indoor. CUPI instrument records data at a very fine time scale [8].

2. Data and Methodology

$PM_{2.5}$ observations were taken at Dwarka (South–West Delhi), India, using a Compact and Useful $PM_{2.5}$ Instrument (CUPI), which is a Panasonic sensor of small proportion including data logger. In order to ascertain the typical standards required for air quality measurements, this instrument has been duly tested and verified on behalf of Panasonic and deemed suitable for air environment monitoring. CUPI is capable of air quality measurements including HEMS (home energy management system) and BEMS (building energy management system) [9]. The volume of $PM_{2.5}$ is estimated by this sensor in terms of the counts and intensities of scattered signals from single particles. This total volume once observed was then converted to mass concentration after multiplying by a factor of 1.4. This factor was determined and explained in earlier study by Nakayama et al. [10]. Dhaka et al. [8] in a recent study, during lockdown period, used this CUPI instrument and demonstrated its capability of taking observations at a fine time resolution under different environmental conditions.

The physical environment in the premise of CUPI is located on the first floor, which is about 4 m above the ground. During observations, it was occupied by 2 members, with open window. The size of the room is 50 sq m. There was no kitchen usage during observations, however, it is the residential area wherein kitchen usage was common in all floors. During day, a good ventilation provision of air flow was there based on the prevailing meteorological conditions [9]. The details of the instrument are given at Nakayama et al. [10]. For the sake brevity, we mention here the highlight of the CUPI instrument and data quality.

Following is the instrument set up for the real-time observations.



Instrument set up at the observation site An onsite measurement device "CUPI" is shown in the figure. Upper panel depicts indoor device, and for the measurement of outdoor air $PM_{2.5}$ concentration, a USB cable is seen, which is connected to the CUPI unit located outside the house in Dwarka sub city, New Delhi. Size of the CUPI can be visualized as it is placed on the corner of the laptop. This is a compact size, having length 5.2 cm, width 4.5 cm, and the thickness being 2.2 cm.

2.1. Data Source, Uncertainly, Errors, and Novelty of the Observations

Under high relative humidity conditions ($> 70\%$), the sensor has a tendency to overestimate the $PM_{2.5}$ mass concentrations compared to those measured by the standard instruments such as BAM instrument (Nakayama et al. 2018). The CUPI instrument also tested in the Indian environment, and the device did not saturate even up to 800 $\mu g/m^3$ (measurement done during Diwali has shown the concentration close to 1000 $\mu g/m^3$).

The comparison of the CUPI measurement at Dwarka location with the standard instrument (BAM-1020 instrument) at US Embassy in New Delhi have shown that the observations are in good agreement. Data comparison with the BAM instrument was shown during Oct–Nov 2019 [See Fig. 3 of our recent study] [11]. Further, the sensitivity (Unit-to-unit variation) of the sensor is typically less than 10%, temporal resolution being less than 2 min.

This new device has a novelty because of its compact size and low power consumption. The power consumption of CUPI is about 0.4 W at 5 V. It is easy to carry and install at any place as it requires minimum space. The life span of

the CUPI in the most polluted environment is about nearly one year. Observations are quite reliable as mentioned above.

Geographical location of the observation site is at Dwarka, New Delhi, which is the capital and a chief metropolitan city of India. It is situated between the latitude of 28° 34′ N and the longitude of 77° 12′ E at an elevation of around 213.3–305.4 m above sea level. Atmospheric conditions of Delhi are subtropical and is characterized by a semiarid climate with summer (April–June), moderate rainfall (July–September), post-monsoon (October–November), and winter (December–February). Temperature of Delhi is characterized by a maximum of ∼ 45–48 °C in June during the summer season and minimum of ∼ 1–2 °C in January during the winter season, and the average rainfall is around 611 mm [10].

High-resolution data of $PM_{2.5}$ were acquired using CUPI (one indoor and the other one outdoor), which were deployed to produce hourly and daily averaged data over a period from Nov 2018 to Oct 2019. Meteorological data were also used from the nearby central pollution control board (CPCB) and Delhi pollution control committee (DPCC) stations during this period. CPCB provides data quality assurance by following rigorous protocols for the sampling, analysis, and calibration. Continuous ambient air quality monitoring (CAAQM) systems are used. Multipoint calibration using automatic dilution system for the calibration or/and auto calibration are in practice [12]. For the analysis of horizontal wind '*u* and *v*' and vertical wind '*w*', which is required to understand the large-scale convergence of winter season, the monthly mean ERA-5 reanalysis data for a period of 10 years from 2010 to 2019 have been used. This is an improved dataset in comparison to earlier ones like ERA interim. Moreover, ERA 5 dataset surmounts the high-resolution topical analysis on 31 km horizontal resolution, thereby permitting an evolution of weather systems.

3. Results

3.1. Diurnal Variation of $PM_{2.5}$ in Outdoor and Indoor Environment During All Seasons from Nov 2018 to Oct 2019

Hourly data of $PM_{2.5}$ acquired using CUPI for both outdoor and indoor environment are shown in Fig. 1a. Diurnal (00 Hrs to 24 Hrs) variation is quite evident; pollution remains quite dominant during winter season both for indoor as well as outdoor air. Concentration of $PM_{2.5}$ were found in the range of 100–600 μg m^{-3}, which maximizing during morning and late evenings. In the afternoon, diurnal cycle shows a decline. However, during winter when indoor and outdoor concentration is compared, significant difference is noted outdoor remains on higher side. In order to investigate the difference in the concentration, data are examined on a fine scale especially during winter. For this purpose, the difference in outdoor and indoor concentration of $PM_{2.5}$ is plotted in Fig. 2. Upper panel shows diurnal variation of concentration of $PM_{2.5}$ throughout the year, while lower panel is focussed to show the difference during winter period, starting from Nov 2018 to Feb 2019. Owing to the fine scale data, we took the difference of indoor and outdoor daily averaged $PM_{2.5}$ concentration data to obtain fine scale view for winter season (see lower panel, Fig. 2).

CUPI sensor-based air quality monitoring, installed at Dwarka station, revealed that diurnal variation in the indoor environment is missing and not prominent as seen in the outdoor environment. During high concentration period (Nov–Dec), indoor pollution remains less in the range of 100–200 μg m^{-3} in comparison to outdoor. Thus, the effect of meteorology (solar radiation, wind speed, and relative humidity, etc.) in outdoor air pollution is much more influential than indoor environment. This is one of the key findings of this analysis. Another important finding is the intermittence presence (for about a week period) of low values of $PM_{2.5}$ (by an amount of 50–100 μg m^{-3}) in outdoor environment, which coincides with sunny days and fine weather conditions. Moreover, the 24-h averaged graphs of both indoor and outdoor data along with the correlation coefficient between the two datasets have been incorporated for an increased comprehensibility of the comparison results [9]. The graph reflects a significant variation in the outside and inside data. The correlation coefficient (R_{xy}) between the two series is '0.71', which suggests a high correlation as seen in time variation of $PM_{2.5}$ as shown in Fig. 1b. First half of the graph shows less concentration of indoor, while after mid-December indoor concentration was marginally larger than outdoor concentration. This finding is important, which suggests during crop burning period [11] there is extra load due to transport of the pollutants from Punjab and Haryana with low wind conditions; whereas in the latter half of the winter this kind of activity stops and meteorological conditions taking the scenario of dispersion due to wind and solar radiation (See Online annexure S1 for vertical wind variability).

Meteorological instruments, which are deployed by CPCB and DPCC, precisely measure with high accuracy and resolutions all the atmospheric parameters such as wind speed, wind direction, ambient temperature, relative humidity, solar radiation, and atmospheric pressure using state of the art technology. These processes involve automated transfer of data every 15 min [8]. CUPI is located in sector 4, Dwarka, which is at about 4 km away from DPCC Dwarka monitoring site.

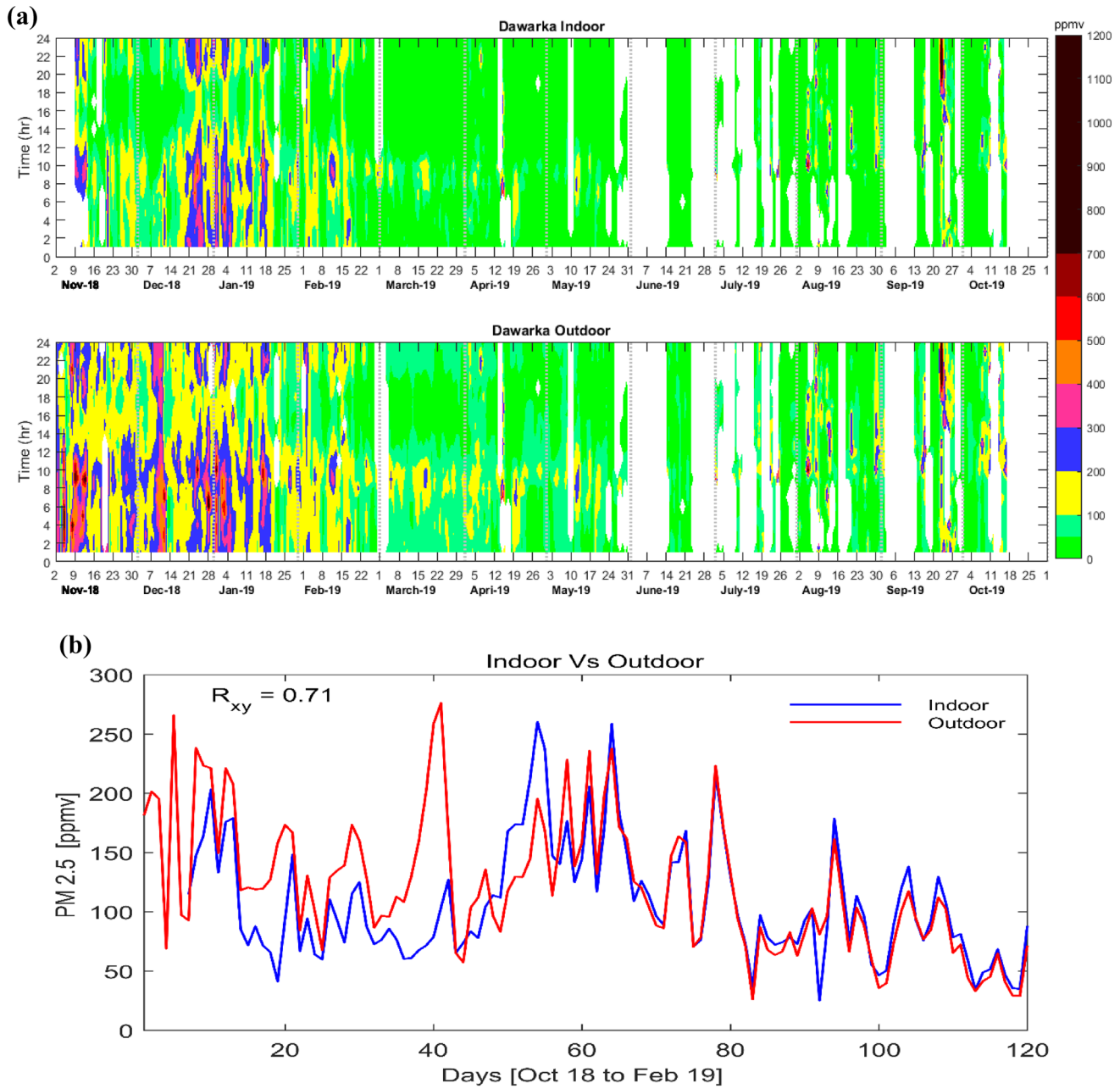


Fig. 1 **a** Hourly averaged $PM_{2.5}$ concentration from November 2018 till October 2019 over Dwarka, New Delhi, showing both Indoor–outdoor Contour. Alongside scale depicts the colour code of the $PM_{2.5}$ concentration. On *x*-axis, year-long data are plotted, while on *y*-axis, data displayed from 00 to 24 Hrs each day. **b** Correlation plot between the 24-h averaged $PM_{2.5}$ indoor and outdoor datasets highlighting a significant correlation of 0.71 for the period from Oct 2018 to Feb 2019

3.2. Diurnal variation of solar radiation and its structure from Nov 2018 to Oct 2019

Solar radiation is one of the key parameters, which plays an important role in dispersing the pollutants. High concentration of $PM_{2.5}$ has been observed during winter partly due to less reaching solar radiation on the ground. Figure 3 represents diurnal variation of solar radiation from Nov 2018 to Feb 2019, covering fairly a winter period. Late winter, during Jan and Feb 2019, solar radiation amounting $\geq$ 300 W/m^2 reaching on ground. Note that lower panel of Fig. 2 shows not much difference to indoor and outdoor pollution. Dhaka et al. [8] have shown in the CUPI observations, during March and April 2020 (lockdown period), that solar radiation crossing 250 W/m^2 is sufficient to evaporate the tiny droplets of the morning prevalent haze. It is apparent that during Dec 2018, less solar radiation ($\leq$ 250 W/m^2) were present hence not sufficient to

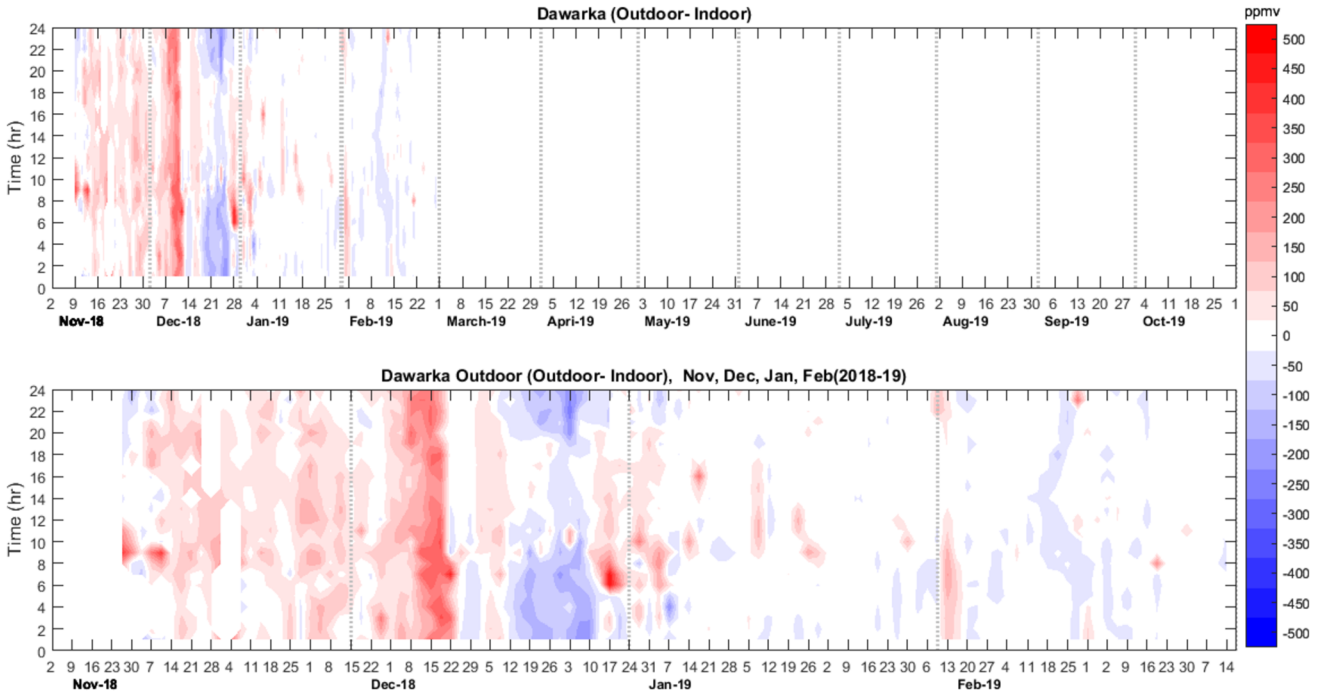


Fig. 2 Difference of outdoor and indoor (outdoor–indoor) daily averaged $PM_{2.5}$ concentration is shown from Nov 2018 to Oct 2019 over Dwarka, New Delhi, showing both Indoor-outdoor Contour. In the bottom panel, scale is expanded (difference is highlighted during winter, i.e. from Nov 2019 to Feb 2019)

disperse the pollutants. Such situation of lesser solar radiation favours more trapping of pollutants.

Lower panel of Fig. 3 shows that high relative humidity ($\geq$ 70%) is more susceptible to formation and retention of high concentration of $PM_{2.5}$. As soon as relative humidity declines after 10–11 am in the diurnal cycle during winter (that corresponds to increasing of solar radiation), the CUPI observations show less $PM_{2.5}$. Decrease in relative humidity ($\leq$ 40–50%) corresponds to 100–200 μg m^{-3} dip in the concentration (see Fig. 1, upper panel). Thus, both solar radiation and relative humidity are the key atmospheric parameters, which play a significant role in the formation and dispersion of the pollutants during winter [5, 7].

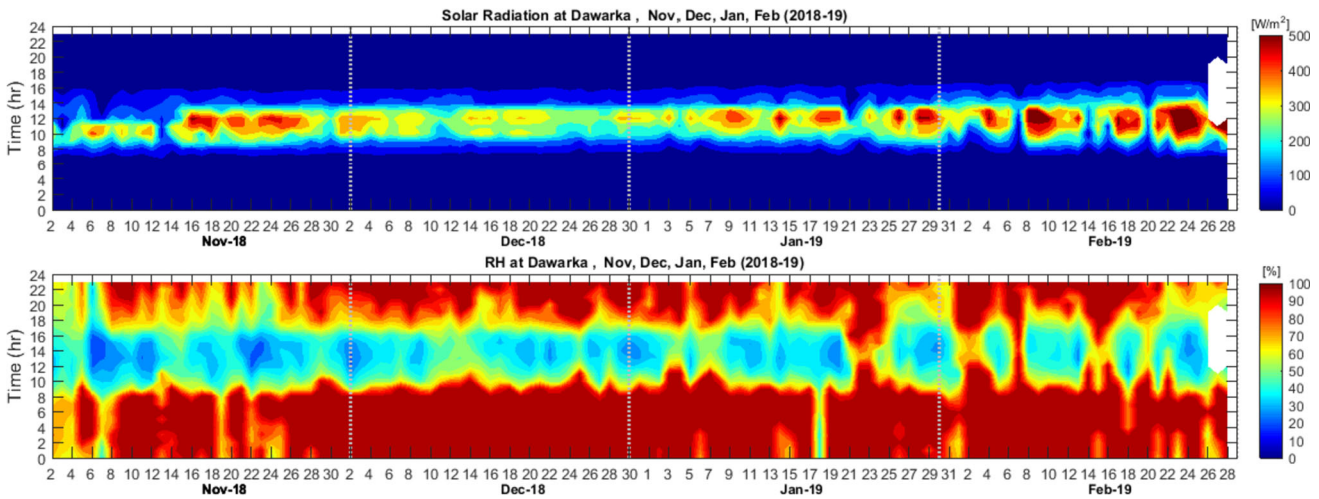


Fig. 3 Diurnal variation of solar radiation (upper panel) and relative humidity (lower panel) during winter (Nov-18 to Feb-19), a peak season of high $PM_{2.5}$ concentration

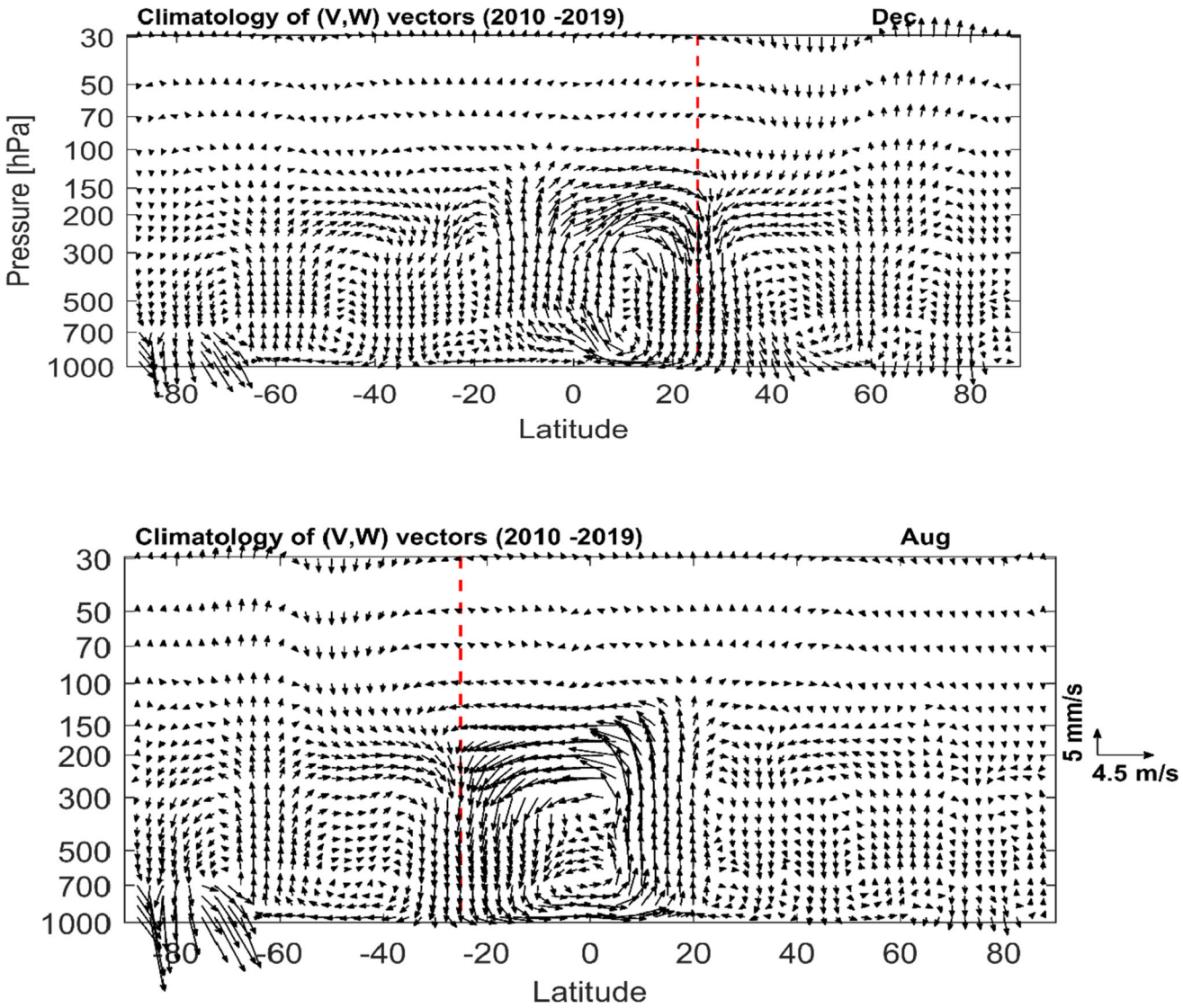


Fig. 4 Representation of meridional and vertical wind (Hadley cell) for the month of December (upper panel) and August (lower panel). Wind vector (V, W) is shown from Northern pole to Southern pole from surface to 30 mb pressure level. The vertical red dashed line is shown at the latitudinal position of Delhi in the upper panel. During winter, downward motions (convergence) are quite evident over Northern India, which is a favourable condition for increasing pollutants

3.3. Role of Hadley Cell in Determining Dynamics of Atmosphere During Winter in Northern India

As mentioned above, there are meteorological factors, which prominently describes day to day weather conditions and temporal variation in the concentration of pollutants. Winter conditions in northern India experiences low wind speed, foggy environment, less solar insolation, high relative humidity, comparatively high atmospheric pressure, and low atmospheric boundary layer [1]; all these factors are favourable for developing high concentration of $PM_{2.5}$. Large-scale motions including Hadley cell is also examined to check the convergence or divergence of the air mass so that one can understand the mechanism on a wider scale during winter in northern India. Monthly mean ERA-5 reanalysis data for a period of 10 years from 2010 to 2019 have been used to generate zonally averaged wind from northern hemisphere to southern hemisphere to show meridional and vertical wind structure with altitude (pressure). Above mentioned features in wind are shown in Fig. 4. Climatology of meridional and vertical wind is shown during Dec (winter) in the upper panel, while lower panel depicts the wind scenario during Aug (summer).

Convergence of air mass during winter seems prominent, this natural variability provides a way for high concentration of pollutants over northern Indian region. Red dotted vertical line represents location of Delhi, which corresponds to situation of downward wind. This is an

additional feature of large-scale circulation, which creates downward motion and precisely unfavourable for dispersion of pollutants during winter season. Air mass convergence shifts in the southern hemisphere around 20–25°S during Aug (lower panel), while there is an upward motion during summer in the northern hemisphere creating less accumulation of the pollutants.

Thus, large circulation is a mechanism, which explains the tropospheric wind structure and movements which in turn acted as a tool for us to understand the stagnant atmospheric conditions prevailing over Delhi during winter [13–16] seen from Fig. 4, we may conclude that the rate of vertical dispersion is very slow and weak thus whatever movement happens it will be aligned horizontally.

4. Conclusion

Measurement of $PM_{2.5}$ is carried out simultaneously outdoors as well as indoors using a Compact and Useful $PM_{2.5}$ Instrument (CUPI) from Nov 2018 to Oct 2019. Following are the important findings of the indoor and outdoor environmental pollution.

(1) Year-long observations provide an opportunity to decipher that indoor $PM_{2.5}$ concentration increases significantly from mid-November 2018 to mid-December 2018. During this high concentration occurrence of $PM_{2.5}$, outdoor concentration was higher in the range of 100–200 $\mu g/m^3$ than indoor. Indoor pollution was less by 30–50% in comparison to the outdoor environment.

(2) Analyses reveal that during a high pollution period, intermittently for a short period, outdoor $PM_{2.5}$ was less, which coincides with clear weather days suggesting that indoor pollution did not disperse due to a closed environment.

(3) Diurnal variation of $PM_{2.5}$ clearly shows high concentration stays until 12:00 noon in both indoor and outdoor environments, which reduces in the afternoon and again picks up in the evening.

(4) Diurnal variation of solar radiation showed a consistent increase from 9 am to 2 pm, while relative humidity declined considerably from 10 am to 4 pm (local time); these two factors correspond to less concentration in both indoor and outdoor environments. However, there is no significant difference between outdoor and indoor concentration during summer and monsoon season.

(5) As seen during Dec 2018, less solar radiation ($\leq$ 250 W/m^2) were present hence not sufficient to disperse the pollutants. Such situation of lesser solar radiation favours more trapping of pollutants. On the other hand, high relative humidity ($\geq$ 70%) is more susceptible to formation and retention of high concentration of $PM_{2.5}$.

(6) During winter, convergence of air mass of (Hadley cell) taking place during winter in the northern India, thus showing a downward motion creating a condition of less dispersion of the pollutants. This is the additional large-scale circulation feature, in addition to weak wind, low solar radiation, high relative humidity, and lower atmospheric boundary layer.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12647-023-00694-2.

Acknowledgements Authors acknowledge the Delhi Pollution Control Committee (DPCC) and Central Pollution Control Board (CPCB) for sharing meteorological data from their sites. All the authors thankfully acknowledge Prof. S. Hayashida, Prof. Y. Matsumi, and Prof. T. Nakayama, Dr. Vinay Kumar, and Ms. Chetna for their comments and help. Instrument (CUPI) provided for the observations gratefully acknowledged. ERA-5 reanalysis data have been used to generate wind components.

References

1. S.K. Chetna Dhaka, G. Longiany, V. Panwar, V. Kumar, S. Malik, A.S. Rao, N. Singh, A.P. Dimri, Y. Matsumi, T. Nakayama and S. Hayashida, Trends and variability of $PM_{2.5}$ at different time scales over Delhi: long-term analysis 2007–2021. Aerosol Air Qual. Res., *15* (2022) 20–30. https://doi.org/10.4209/aaqr.220191.
2. Y. Chen et al., Local characteristics of and exposure to fine particulate matter (PM2.5) in four indian megacities. Atmos. Environ. X, *5* (2020) 100052. https://doi.org/10.1016/j.aeaoa.2019.100052.
3. J.L. Schnell, V. Naik, L.W. Horowitz, F. Paulot, J. Mao, P. Ginoux, M. Zhao and K. Ram, Exploring the relationship between surface PM2.5 and meteorology in Northern India. Atmos. Chem. Phys., *18*(14) (2018) 10157–10175. https://doi.org/10.5194/acp-18-10157-2018.
4. Z. Chen, D. Chen, C. Zhao, M.P. Kwan, J. Cai, Y. Zhuang, B. Zhao, X. Wang, B. Chen, J. Yang and R. Li, Influence of meteorological conditions on PM2.5 concentrations across China: a review of methodology and mechanism. Environ. Int., *139* (2020) 105558. https://doi.org/10.1016/j.envint.2020.105558.
5. S. Gani et al., Submicron aerosol composition in the world's most polluted megacity: the Delhi aerosol supersite study. Atmos. Chem. Phys., *19*(10) (2019) 6843–6859. https://doi.org/10.5194/acp-19-6843-2019.
6. A. Malik, S.G. Aggarwal, B. Kunwar, D.K. Deshmukh, K. Shukla, R. Agarwal, K. Singh, D. Soni, P.R. Sinha, S. Ohata and T. Mori, Physical and chemical properties of PM1 in Delhi: a comparison between clean and polluted days. Sci. Total Environ., *892* (2023) 164266.
7. A. Malik, S.G. Aggarwal, S. Ohata, T. Mori, Y. Kondo, P.R. Sinha, P. Patel, B. Kumar, K. Singh, D. Soni and M. Koike, Measurement of black carbon in Delhi: evidences of regional transport, meteorology and local sources for pollution episodes. Aerosol Air Qual. Res., *22*(8) (2022) 220128.

8. S.K. Dhaka, Chetna, V. Kumar, V. Panwar, A.P. Dimri, N. Singh, P.K. Patra, Y. Matsumi, M. Takigawa, T. Nakayama and K. Yamaji, PM2.5 diminution and haze events over Delhi during the COVID-19 lockdown period: an interplay between the baseline pollution and meteorology. Sci. Rep., *10*(1) (2020) 13442. https://doi.org/10.1038/s41598-020-70179-8.
9. X. Liu, ASTM and ASHRAE standards for the assessment of indoor air quality, Handbook of Indoor Air Quality, (2021), pp. 1–36.
10. T. Nakayama, Y. Matsumi, K. Kawahito and Y. Watabe, Development and evaluation of a palm-sized optical PM 2.5 sensor. Aerosol. Sci. Technol., *52* (2018) 2–12.
11. M. Takigawa, P.K. Patra, Y. Matsumi, S.K. Dhaka, T. Nakayama, K. Yamaji, M. Kajino and S. Hayashida, Can Delhi's pollution be affected by crop fires in the Punjab region? SOLA, *16* (2020) 86–91. https://doi.org/10.2151/sola.2020-015.
12. B.P. Singh and P. Kumar, Spatio-temporal variation in fine particulate matter and effect on air quality during the COVID-19 in New Delhi, India. Urban Clim., *40* (2021) 101013. https://doi.org/10.1016/j.uclim.2021.101013.
13. Z. Liu, L. Shen, C. Yan, J. Du, Y. Li and H. Zhao, Analysis of the influence of precipitation and wind on PM2.5 and PM10 in the Atmosphere. Adv. Meteorol. (2020). https://doi.org/10.1155/2020/5039613.
14. I.M. Dima and J.M. Wallace, Notes and correspondence on the seasonality of the Hadley cell. J. Atmos. Sci., *12* (2002) 1522–1527.
15. T. Xian, J. Xia, W. Wei, Z. Zhang, R. Wang, L.P. Wang and Y.F. Ma, Is Hadley cell expanding? Atmosphere, *12*(12) (2021) 1699. https://doi.org/10.3390/atmos12121699.
16. R.V. Martin, M. Brauer, A. van Donkelaar, G. Shaddick, U. Narain and S. Dey, No one knows which city has the highest concentration of fine particulate matter. Atmos. Environ. X, *3* (2019) 100040. https://doi.org/10.1016/j.aeaoa.2019.100040.

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

Solar Compass 8 (2023) 100064

Contents lists available at ScienceDirect

Solar Compass

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

Potential assessment of solar industrial process heating and CO_2 emission reduction for Indian cement industry

Niranjan Sahoo [a,b], Anil Kumar [b,c,*]

[a] Galgotias College of Engineering and Technology, Greater Noida, India
[b] Department of Mechanical Engineering, Delhi Technological University, Delhi, India
[c] Centre for Energy and Environment, Delhi Technological University, Delhi, India

ARTICLE INFO

Keywords:
Solar energy
Industrial process heating
Cement
Carbon emission
Calcination

ABSTRACT

For the Indian cement sector, a simple approach is presented to evaluate the potential of solar industrial process heating (SIPH) and the resulting decrease in CO_2 emissions. The first step was to identify the locations of cement plants with their annual installed capacity and their annual actual cement production. After that, the yearly process heating requirement for the calcination process at each cement plant has been estimated using the actual annual cement production value. Finally, using the concept of a top-of-tower (TT) solar plant design, the total thermal energy that could be saved has been estimated as 771.35 Petajoule (PJ) /annum. Finally, the adoption of the SIPH system with storage is expected to mitigate CO_2 emissions by 45.2 MT (Megatons) annually. By utilizing renewable energy sources in the cement manufacturing process, energy consumption and CO_2 emissions will be reduced.

1. Introduction

India contributed more than 7 % of the world's installed capacity for cement production, placing it as the second largest manufacturer in the world. In 2021, India's domestic cement production was nearly 294.4 MT, while its installed capacity was 500 MTPA (Mega tons per annum). India got a growth of 12.1 % in cement production from September 2021 to September 2022. It is backed up by the robust real estate market and significant government investment in smart city projects and urban infrastructure. It is supported by the high level of activity going on in real estate and high Government spending on smart cities and urban infrastructure. It is expected that the cement demand will reach 419.92 MT by 2027 [1]. India currently has 210 large plants and 350 mini plants with cumulative installed capacities of 410 MT and 11.10 MT respectively. Most of the large cement plants are situated in three states (Andhra Pradesh, Rajasthan and Tamil Nadu). The state-wise installed capacity and actual production of cement for the year 2020–21 are shown in Fig. 1(Source: Indian Minerals Yearbook 2021) [2].

Energy is the most fundamental need of any industry to sustain and grow. A considerable amount of thermal energy is utilized during the process heating while most of the processes operate in the temperature range between 400 °C and 1400 °C. Production of cement needs a tremendous amount of energy and it is in the form of thermal and electrical respectively. Thermal energy is utilized for heating various processes such as blending and homogenization, precalcination, pyro processing, etc., whereas electricity is used to operate air compressors, coolers, grinding units, lighting systems, etc. The thermal and electrical energy required for the different processes of cement production is shown in Fig. 2.

Cement use is rising quickly on an international scale. A huge quantity of thermal energy is needed to produce cement, which is in high demand. Most of the cement industries use fossil fuels for generating thermal energy. The most commonly used fossil fuels are coke, coal, biomass, natural gas, etc. Consumption of these fossil fuels generates a lot of greenhouse gases into the atmosphere which degrades the atmosphere day by day [4]. Hence, the use of fossil fuels should be limited and there must be some alternatives so that the cement sectors should not be disturbed. One of the best ways to come out of the problem is the use of renewable sources of energy. Solar energy can be effectively used to provide thermal energy for the different processes of cement manufacturing. Calcination is an important process of cement manufacturing and previous research suggested that concentrated solar energy could replace conventional fossil fuel. Solar reactors are used for calcining the raw materials which can be fixed at two different positions as shown in Fig. 3 [5–7].

Many researchers have investigated the potential of solar energy in cement manufacturing, as evidenced in a literature review [9–13].

* Corresponding author at: Department of Mechanical Engineering, Delhi Technological University, Delhi, India.
E-mail address: anilkumar76@dtu.ac.in (A. Kumar).

https://doi.org/10.1016/j.solcom.2023.100064
Received 12 July 2023; Received in revised form 30 September 2023; Accepted 8 November 2023
Available online 10 November 2023
2772-9400/© 2023 The Author(s). Published by Elsevier Ltd on behalf of International Solar Alliance. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Nomenclature

$ATED_{CP}$	annual thermal energy demand of clinker production
$STED_{CP}$	specific thermal energy demand of clinker production
$APCC_i$	annual production of cement clinker in the ith state of the nation
Q_{rxn}	energy flux required for raw material calcination
Q_{solar}	energy flux needed for the process
Q_l	energy flux losses at the reactor
η_{SF}	solar field efficiency
Q_{SR}	solar reactor power output

Gonzalez and Flamant checked the technical feasibility of using solar energy for the calcination process in cement manufacturing [7]. This investigation included concentrated solar energy as a potential replacement for 40 % to 100 % of the thermal energy needed in conventional cement production. It also included the economic analysis and the technical issues while implementing concentrated solar energy in cement production. The study revealed that the central power tower with a solar reactor positioned at the top of the preheater tower is the best design model for implementing concentrated solar energy. Implementation of concentrated solar energy in the calcination process of cement manufacturing can avoid 40 % of CO_2 emissions into the atmosphere. A silo was proposed for storing calcined raw material at 900 °C. Land requirements were calculated based on the criteria that about 20 % of the total land is used for the concentrated solar power (CSP) central tower plants. A 120MW solar power would require 192 ha of land. Payback time (PBT) and internal rate of return (IRR) calculations were used to evaluate the economic feasibility of CST technology in cement manufacturing. Payback timeframes ranged from six to ten years, which is much less than the typical 20-year payback times for CSP projects. IRR lies between 8.77 % and 11.36 % for all scenarios except those with 45 % thermal losses.

Another researcher assessed the technical feasibility of using concentrated solar energy for the calcination process in cement production [14]. The analysis suggested a design of the solar cement plant that uses the concept of solar thermal calciner built at German Aerospace Canter (DLR). An analysis of the solar calciner's energy balance and the impact of different scenarios is presented. The analysis revealed that 14–17 % of CO_2 emissions can be avoided by upgrading the conventional cement plant with a solar cement plant. Furthermore, the potential reduction of CO_2 emissions in Spain through the year 2050 is calculated. According to the analysis, 39 % of the plant were located in such areas where solar energy can be utilized for cement manufacturing. The Spanish cement industry could reduce emissions by 7 % up to 2050 through the use of solar calciners instead of fossil fuels in the conventional calciner. Costs of avoiding CO_2 were calculated using a conservative base scenario and ranged from 118 EUR/t to 74 EUR/t, depending on solar irradiance, reactor effectiveness, and solar multiple used.

A 3000-ton daily production capacity of a clinker plant was evaluated for concentrated solar energy applications [15]. A hybrid process (solar + fossil fuel) was suggested for supplying thermal energy in the calcination process. The minimum solar irradiation needed for the plant to operate in solar mode was 2000 kWh/m^2a and the efficiency of the solar calciner was assumed to be 86 %. The analysis did not include any thermal storage for storing calcined raw material. An author's estimate indicates that 9 % of CO_2 emissions and 28 % of total fossil fuel inputs can be avoided compared to a conventional cement factory.

The Indian government has initiated several programs to promote the use of solar energy in the nation's industrial sector. As part of its efforts to enhance the utilization of solar energy in the country's various industries, the nation collaborated with several international organizations like the Global Environment Facility (GEF), the United Nations Industrial Development Organization (UNIDO), and the United Nations Development Program (UNDP). UNDP assists countries and territories in reducing their greenhouse gas emissions and meeting climate change challenges through the use of concentrated solar energy for different processes in the industry. The goal of UNIDO is to create a world free from hunger and poverty, where business promotes low-emission economies, raises living standards, and protects the environment so that everyone can live in peace and comfort in the future. One essential element for the desired success in the aforementioned endeavours is having accurate knowledge about the potential for solar process heating in India as well as specific regions in the nation where early interventions may be made. Furthermore, a comparative analysis of commercially available solar reactors is crucial for delivering relevant information for an informed decision.

With this context in mind, this research makes a first effort to determine the potential of SIPH for the cement sector in India. Additionally, this research includes equivalent estimates of the decrease in

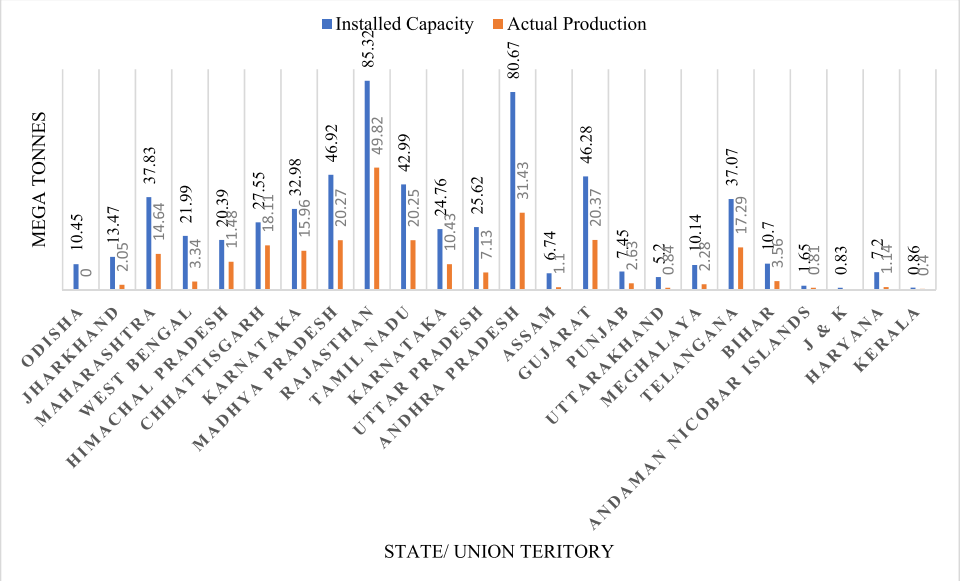


Fig. 1. Production of cement state-wise in India (2020–21).

CO_2 emissions attributable to solar process heating in the Indian cement sector. It should be emphasized that coal is primarily used as a fossil fuel in India's cement sector. Due to this, the present study focuses only on the potential reduction of CO_2 emissions from the use of concentrated solar energy in the cement industry in India, which mainly relies on coal.

In concern to the previous studies, a lot of work has been reported related to the use of solar energy for cement industries. However, as per the knowledge of the authors, no work has been reported that describes a systematic approach for evaluating the potential assessment of SIPH and CO_2 emission reduction for the cement industry. This study presents a thorough examination of concentrated solar energy systems for supplying thermal energy for the processes of cement manufacturing. This research is a considerably more precise method since it considers all relevant factors while estimating the potential of SIPH along with GHG emissions.

This study might be helpful for industries looking to replace thermal Energy with green energy. Additionally, it encourages the use of renewable energy for cement production while lowering the related carbon emissions to the environment. It also highlights the most recent findings, making it a valuable research resource.

The remainder of the work is structured as follows: Section 2 describes the cement manufacturing processes with energy utilized in cement production. The next Section 3 explains the methodology adopted for potential estimation and GHG emission reductions.

A description of each cement manufacturing process and typical energy used in each process is included in Section 2. The approach adopted for evaluating the potential of SIPH and CO_2 emission reductions is described in Section 3. Section 4 presents and discusses the findings of the research and Section 5 ends with outlining the conclusions and recommendations.

2. Cement manufacturing processes and typical energy consumption

The production of cement needed a series of processes. These include (i) Quarrying and crushing of limestone (ii) Addition of additives (iii) Raw meal preparation (iv) Blending and storing (v) Preheating and pre calcining (vi) clink erization (vii) Clinker cooling and final grinding and (viii) packaging and despatch. The next paragraphs provide a detailed description of the processes (Fig 2) that are involved in quarrying to packing.

2.1. Raw material extraction/ quarrying

The first process of cement manufacturing is quarrying, in which hard rocks (limestone) are extracted and broken into smaller pieces. For crushing of ores generally, two crushers are used. The primary crusher converted the rocks into sizes of about 6 inches. Then the secondary crusher further reduces the size up to 3 inches. Bulldozers and trucks are then used to transport the crushed limestones to the cement factory [16–18].

2.2. Raw material preparation

The next process after quarrying is raw material preparation, in which the crushed limestone is added with some ingredients like iron, silica, bauxite, quartzite, etc. All these ingredients are ground into fine powders and stored in the silos and fed to the kiln. The grinding of raw materials is done with the help of a vertical roller mill (VRM) or a ball mill. The ball mills typically contain balls of various sizes, and a classifying liner fixes each ball's place. A larger size of the ball is used for impact grinding, while a smaller size is used for attrition grinding. A VRM adopts the compression principle on the raw material for grinding.

2.3. Blending and storing silo

The composition of the raw material fed into the rotary kiln system has a significant impact on both the effectiveness of the kiln and the quality of the clinker that comes out of it. The components in the raw material should be properly blended and homogenized to prevent variability. For proper blending and homogenization of the raw components, continuous blending silos are utilized.

2.4. Preheating and pre-calcining of raw material

As the name suggests pre-heating means the heating of raw materials before they entered into the rotary kiln system. This process is done only for the dry process of cement manufacturing. The pre-heating is done in a tower known as a preheater tower which made of several vertical cyclones. Pre-heating the raw material is accomplished with the help of hot gases that are released from the kiln to reduce energy consumption and improve the environmental friendliness of the cement-making process. In order to heat the raw material to a temperature of around 1000–1200 °C, hot air is pulled into the cyclone from the opposite

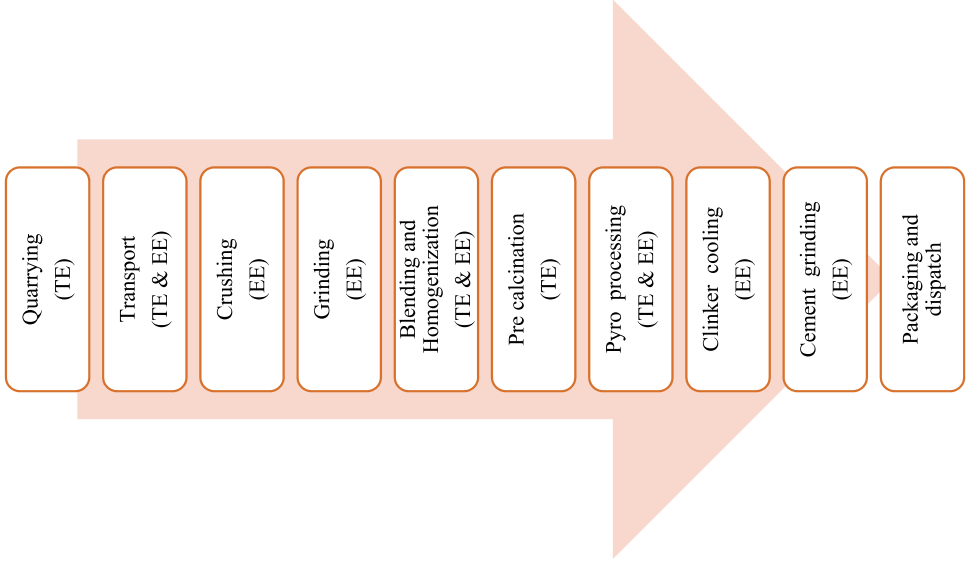


Fig. 2. Processes in cement production along with energy requirement [3].

direction of the material flow. The modern cement manufacturing plant contains a combustion chamber that is internal to the preheater tower. this is commonly known as pre-calciner.

2.5. Kiln phase

The kiln system is the most important part of a cement manufacturing unit. It consists of a large inclined steel tube that rotates between 0.5 and 4.5 revolutions every minute. The diameter of this rotating kiln ranges from 3 to 9 m, and its length varies from 50 to 200 m, respectively. The slope of the rotary kiln ranges between 2.5 and 4.5 % to the horizontal. For an efficient heat transfer, the raw material enters from the higher side of the kiln and moves slowly to the lower side. The firing is done at the lower end of the rotating kiln with the use of fossil fuels (coal) [19,20]. Different types of zones in rotating kilns with temperature variations are shown in Fig. 4. Rotary kilns can be designed for handling a broad range of capacities, from small, batch-scale units processing anywhere from 50 to 200 lb/hr, to commercial-scale units processing material in the range of 200 lb/hr to 20 TPH. Large commodity kilns such as those used in the cement industry can process up to 50 TPH, but these size kilns are less common for many of the lower-capacity processes in use today. Residence time, also known as retention time, is the amount of time in which the material is processed in the kiln. As with temperature, the residence time is determined solely by the requirements of the intended reaction.

2.6. Clinker cooling and final grinding

The clinker coming out from the rotary kiln is rapidly cooled down from 2000 °C to 100 °C–200 °C by passing cool air over it. Then various additives are mixed with the clinker to be ground in order to produce the final product, cement. Gypsum is the most commonly added additive which is added to regulate the compressive strength and setting time of the cement. The grinding operation in the cement mill is carried out with the help of metallic balls [22].

2.7. Packing and shipping

The last and final step of cement manufacturing is packing and shipping. Large, vertical silos are used to store cement. Most cement is often transported by trucks, trains, or ships and few quantities are bagged for retail sale for clients who need modest quantities. The transport methods are the same as those used for bulk transport cement.

3. Methodology used to estimate the potential and minimization of GHG

The steps taken for evaluating the potential of SIPH in India's cement manufacturing industry (Fig. 5) involve (a) identifying the nation's cement plants (b) finding out the annual production capacity and actual production of each plant (c) evaluating the annual thermal energy requirements for clink erization process of each plant (i.e. gross SIPH potential) (d) assessing the potential for solar process heating to mitigate CO_2 emissions. The subsequent sections provide a brief explanation of each of the aforementioned steps.

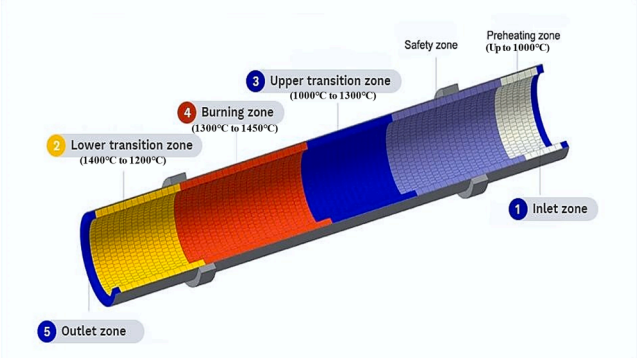


Fig. 4. Different types of zones in a rotating kiln [21].

3.1. Concept of solar cement plant

Based on earlier research, Fig. 6 illustrates the concept of using concentrated solar energy to form cement clinkers [7]. Traditional cement production involves preheating, pre-calcining, and clink erization are the orderly steps involved in the process of raw materials. However, when using solar energy, the calcination process is carried out in a solar calciner that is mounted on top of a solar tower. Solar incident power at the reactor provides the heat needed for calcination, and the output from the solar reactor is delivered to the rotating kiln or thermal storage unit. The raw material flow is shown through the help of a solid line whereas flue gas flow is presented with a dotted line. Thermal storage is used to store calcined material which can be used at night or when there is no proper sunlight for solar power generation.

Modifying the conventional cement plant to a hybrid mode requires installing some new equipment. It mainly includes transport systems for transferring the output from the preheater to the solar reactor and the output from the reactor to either the rotary kiln or the storage unit. In order to reach the high temperature needed for the calcination reaction, a concentrator (compound parabolic) must be installed on the roof of the solar tower. Thermal energy needed for operating the conventional cement plant in solar mode mainly depends on the heat losses in the solar reactor given in Eq. (1). Solar reactor design for a solar cement plant is as follows [6]:

$$\dot{Q}_{solar} = \frac{\dot{Q}_{rxn} + \dot{Q}_l}{\eta_{SF}\alpha_{SR}} \tag{1}$$

Heat and mass balance for the solar reactor is shown in Fig. 7. Assumptions and factors taken into consideration are as follows:

Ø Calcination of raw material starts within the calciner only.

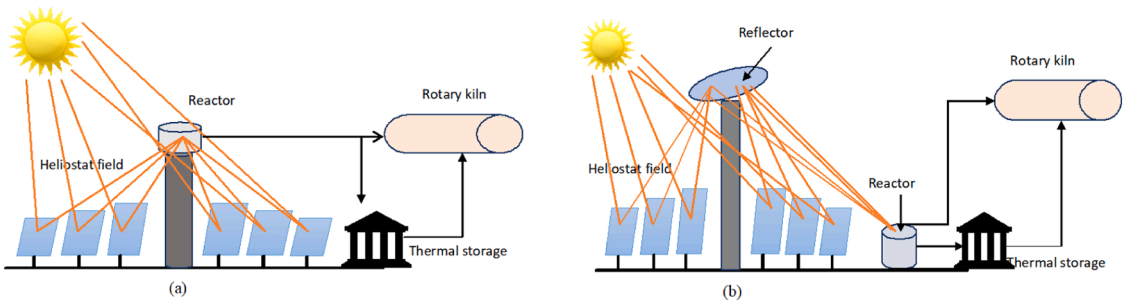


Fig. 3. Model of solar cement plant (a) solar reactor located on the top and (b) solar reactor located on the bottom [8].

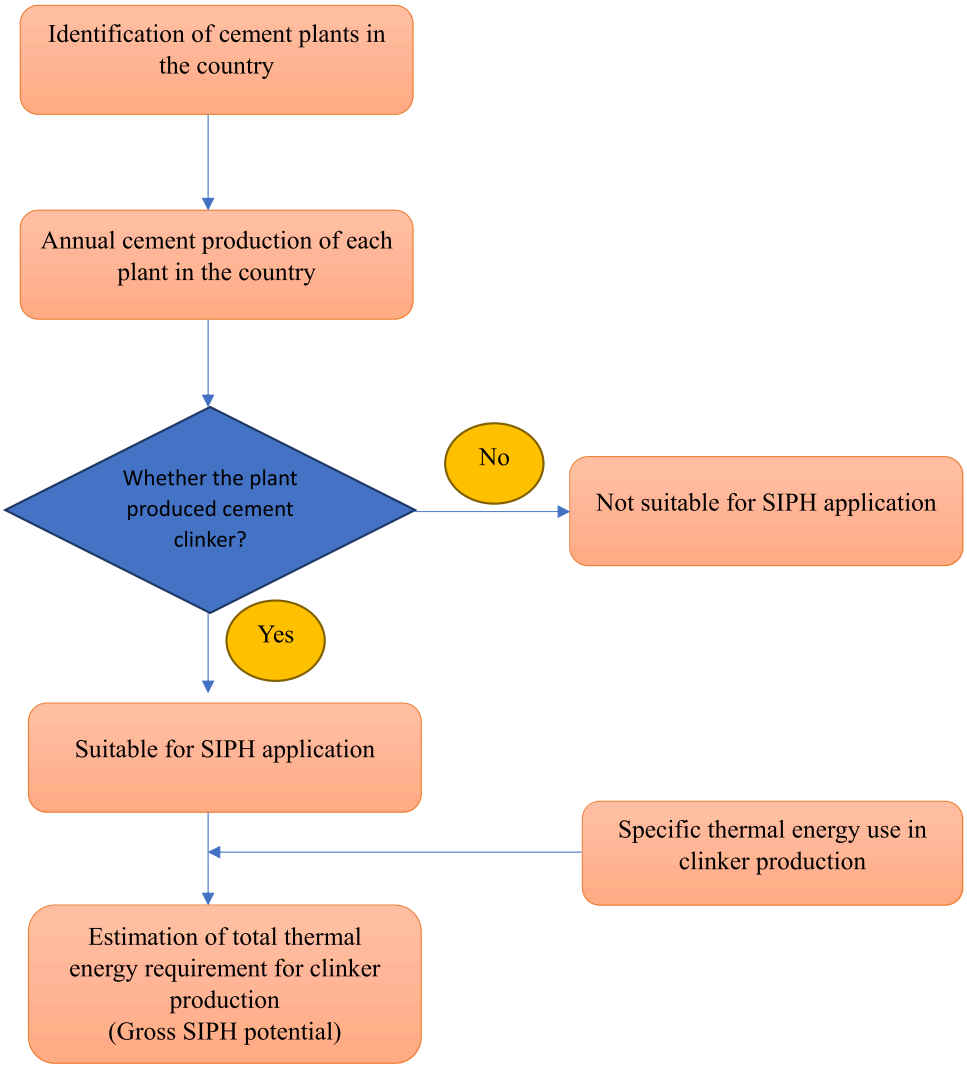


Fig. 5. Methods used in study to estimate SIPH potential.

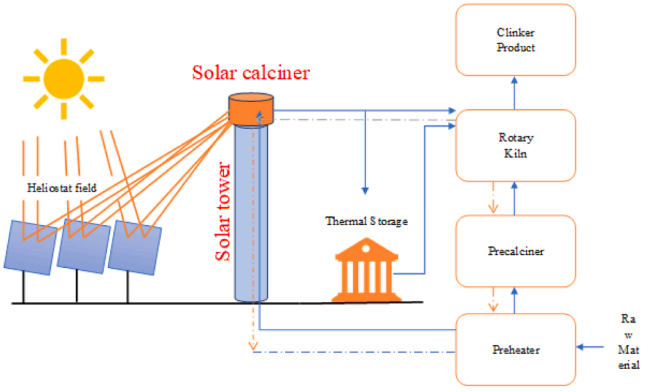


Fig. 6. Design of solar-power cement plant.

- Ø Kiln gas is thoroughly mixed with tertiary air, and no heat is lost from the mixing chamber.
- Ø Unreacted gases and the calcined raw material must exit the solar reactor at a uniform temperature.
- Ø Steady state equilibrium condition within the system is considered.

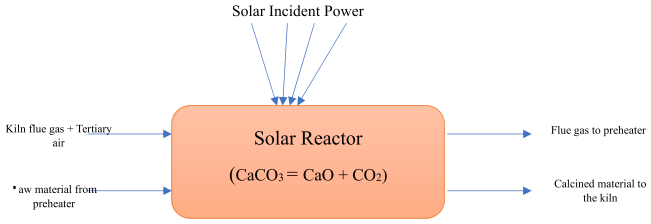


Fig. 7. Solar reactor heat and mass balance.

3.2. *Assessment of thermal energy for cement clinker production*

In cement manufacturing, clink erization is a highly energy-consumable process. The type of rotary kiln and the use of a preheater play a crucial role in determining the clinker's specific energy consumption. Table 1 shows the thermal energy utilized per ton of clinker for various rotary kilns with the number of cyclone preheater stages. For each ton of clinker produced, the dry process of making cement uses 3.40GJ of specific energy, whereas the wet process uses 5.29GJ/t. According to reports, several nations' specific energy consumption per ton of clinker is less than 2.95 GJ, however, India has a minimum value of 3.06 GJ [16–18]. Several factors may contribute to India's higher specific energy consumption, including harder raw materials and

Table 1
Typical thermal energy consumption per ton of clinker.

Rotary kiln types	Cyclone preheater stages	Fuel usage (GJ per ton clinker)
Wet process kiln	——	5.86–6.28
Dry process kiln	——	4.60
Dry process kiln	1	4.2
Dry process kiln	2	3.8
Dry process kiln	3	3.3
Dry process kiln	4	3.14
Dry process kiln	5	3.01
Dry process kiln	6	< 2.93

low-quality fuel [19].

The annual useful thermal energy requirement of clinker production (excluding cement grinding plant and secondary plants) is estimated by using 3.40GJ of specific energy per ton of clinker production and production data of cement as shown in Fig. 1. The clinker-to-cement ratio is assumed to be 0.95.

The country's annual thermal energy demand (PJ) $ATED_{CP}$ for clinker production may be calculated as

$$ATED_{CP} = STER_{CP} \sum_{i=1}^{n} APCC_i \quad (2)$$

Where, $STER_{CP}$ is the specific thermal energy requirement in clinker production and $APCC_i$ is the annual production of cement clinker in the i^{th} state of the nation.

The gross potential for SIPH in the manufacturing of cement in India is theoretically represented by the yearly thermal energy requirements in clinker production. A certain amount of cement clinker is produced rather than dry process of cement manufacturing. All the cement plants in India have not used dry processes for cement manufacturing. Since data regarding the use of rotary kiln for clinker production is not available so in the present analysis it is assumed that all the clinkers are produced by dry process of cement manufacturing. As mentioned earlier, cement grinding plant and secondary cement plants have not been included in this analysis though they have not produced cement clinkers.

3.3. Maximum CO_2 reduction potential

If all of the thermal energy needed for the calcination process is substituted by solar energy, the most substantial quantity of CO_2 reduction is achievable. The highest CO_2 reduction in this situation is equal to:

$$(\dot{m}_{co_2})_{\max.red} = (\dot{m}_{co_2})_{conv} - (\dot{m}_{co_2})_{sol\,\min} \quad (3)$$

The minimum quantity of CO_2 emissions $(\dot{m}_{co_2})_{sol\,\min}$ from a solar cement plant is calculated by the raw materials CO_2 concentration (Y_{rm}) co_2 and the amount of coal utilized in the rotary kiln $(m_{coal})kiln$

$$(\dot{m}_{co_2})_{sol\,\min} = (Y_{rm})_{co_2}\,\dot{m}_{rn} + (\dot{m}_{coal})_{\text{kiln}} LHV_{coal} C_{co_2} \quad (4)$$

The quantity of CO_2 released $(\dot{m}_{co_2})_{conv}$ in the traditional process is determined by the CO_2 content of the raw materials and the amount of coal used in the calciner and rotating kiln $(\dot{m}_{coal})_{calci}$:

$$(\dot{m}_{co_2})_{conv} = (Y_{rm})_{co_2}\,\dot{m}_{rn} + (\dot{m}_{coal})_{\text{kiln}} LHV_{coal} C_{co_2} + (\dot{m}_{coal})_{calci} LHV_{coal} C_{co_2} \quad (5)$$

Therefore, the decrease in CO_2 emissions caused by implementing the solar calciner in a conventional cement plant is computed using the amount of fuel (coal) saved.

$$(\dot{m}_{co_2})_{\max.red} = (\dot{m}_{coal})_{calci} LHV_{coal} C_{co_2} \quad (6)$$

Based on the above equations, the conventional case's CO_2 emissions are to be estimated as $(m_{co_2})conv/\ m_{clinker} = 842\ kgCO_2/t_{clinker}$. The calcination process is responsible for around 66 % of these emissions. Out of the remaining 34 % of emissions, 13 % is due to the fuel used in the clinkering, and 21 % is due to the fuel used in the conventional calciner [23,24]. By using a solar calciner, the CO_2 emissions could be reduced to a minimum of 665 kg per tonne of clinker i.e., (m_{co_2}) sol min/ $m_{clinker} = 665\ kgco_2/t_{clinker}$. Hence, the maximum 21 % of the total CO_2 emission can be avoided when 100 % of the calciner thermal energy required is replaced by solar Energy. Remarkably, the amount of CO_2 emissions produced due to the calcination process is nearly three times greater than those caused by the usage of fossil fuels. Therefore, if CO_2 from the solar calciner is separated in a controlled manner, the impact would grow by a factor of four to 87 %.

4. Results and discussion

This section presents typical results from an attempt to evaluate the potential of concentrated solar energy in the Indian cement sector and the accompanying potential for CO_2 emission reduction using the methodology described in Section 3. Based on conventional and suggested (solar) systems, Table 2 estimates the annual thermal energy needs of clinker production for each state.

Fig. 8 shows the estimated quantity of CO_2 emissions that can be mitigated by installing solar industrial process heating systems in clusters of cement plants in India. The conventional cement plants with existing thermal energy produce 21.498 MT of CO_2 per annum whereas the use of SIPH system in conventional cement plants reduces the CO_2 emission to 16.979 MT per annum. As a result, the potential for reducing CO_2 emissions annually is relatively greater. Additionally, 100 % replacement of fossil fuel with concentrated solar energy could save 45.193 MT of CO_2 emission annually (Fig. 8).

Table 2
Cement production and thermal energy requirement at preferred locations with clusters of cement plants in India.

Indian State/ UT	Annual Installed Capacity (MT)	Annual actual Production (MT)	**$ATED_{CP}$** Based on Conventional System (PJ)	**$ATED_{CP}$** Based on Proposed System (PJ)
Odisha	10.45	0	0	0
Jharkhand	13.47	2.05	6.6215	0.42845
Maharashtra	37.83	14.64	47.2872	3.05976
West Bengal	21.99	3.34	10.7882	0.69806
Himachal Pradesh	20.39	11.48	37.0804	2.39932
Chhattisgarh	27.55	18.11	58.4953	3.78499
Karnataka	32.98	15.96	51.5508	3.33564
Madhya Pradesh	46.92	20.27	65.4721	4.23643
Rajasthan	85.32	49.82	160.9186	10.41238
Tamil Nadu	42.99	20.25	65.4075	4.23225
Karnataka	24.76	10.43	33.6889	2.17987
Uttar Pradesh	25.62	7.13	23.0299	1.49017
Andhra Pradesh	80.67	31.43	101.5189	6.56887
Assam	6.74	1.1	3.553	0.2299
Gujarat	46.28	20.37	65.7951	4.25733
Punjab	7.45	2.63	8.4949	0.54967
Uttarakhand	5.2	0.84	2.7132	0.17556
Meghalaya	10.14	2.28	7.3644	0.47652
Telangana	37.07	17.29	55.8467	3.61361
Bihar	10.7	3.56	11.4988	0.74404
Andaman Nicobar Islands	1.65	0.81	2.6163	0.16929
J & K	0.83	0	0	0
Haryana	7.2	1.14	3.6822	0.23826
Kerala	0.86	0.4	1.292	0.0836
Total	605.06	255.33	824.7159	53.36397

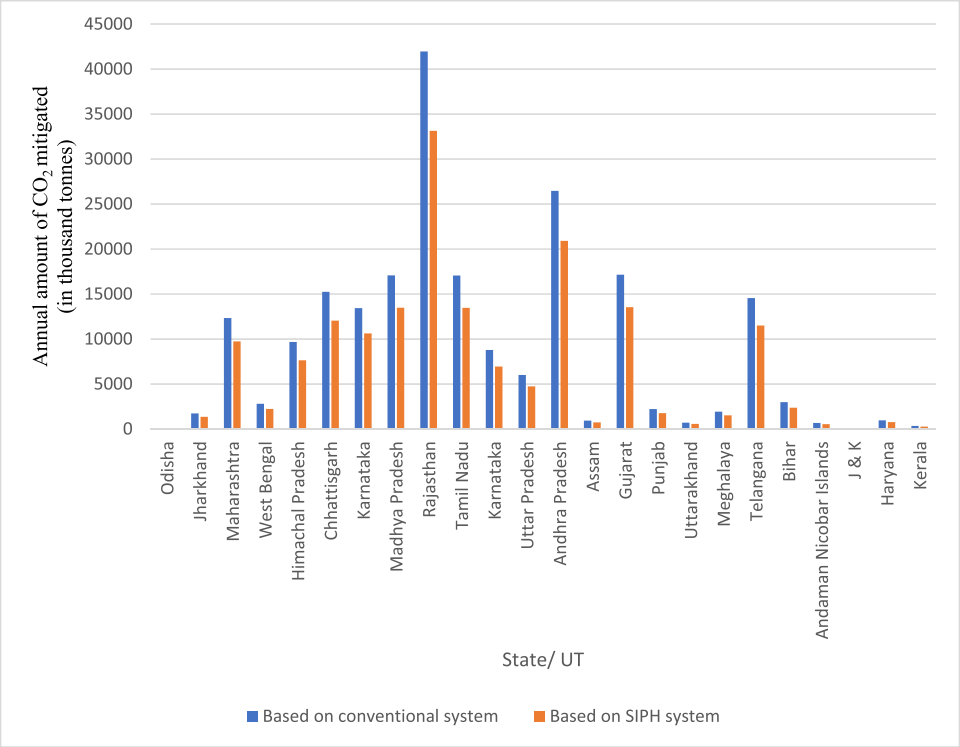


Fig. 8. State-wise reduction of CO_2 emissions at Indian cement plants implementing SIPH system.

5. Conclusion

This paper presented a preliminary analysis to explore SIPH technology in the cement industry. The potential of SIPH system for the clink erization process in cement manufacturing has been analyzed. The annual thermal energy savings by the use of solar calciner reactors was found to be 771.35 PJ. Furthermore, CO_2 emission reduction potential was also evaluated. When all of the calciner's required thermal energy is replaced by solar energy, a maximum of 21 % of the total CO_2 emission may be prevented. The usage of concentrated solar energy can prevent an estimated 45.193 MT of CO_2 emissions. It's important to keep in mind that the estimations in this study are predicated on the premise that there is sufficient space (roofs and/or open land) for the construction of solar reactors.

Declaration of Competing Interest

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

Acknowledgment

Authors are highly grateful to Dr. D. Yogi Goswami, Distinguished Professor, University of South Florida for supporting to creation Centre for Energy and Environment, Delhi Technological University Delhi (India), where this work is compiled.

References

[1] D.R. Sakhamuru, Techno-economic analysis and strategic decarbonization of the Indian cement industry (Master's thesis, Massachusetts Institute of Technology). (2022), https://hdl.handle.net/1721.1/144640 [Accessed 7 July 2023].
[2] Indian Minerals Yearbook 2021 (Part- III: mineral reviews) 60th Edition. https://ibm.gov.in/writereaddata/files/169052447764c35b3dd9249Cement_2021.pdf. [Accessed 7 July 2023].
[3] N. Sahoo, A. Kumar, Samsher, Review on energy conservation and emission reduction approaches for cement industry, Environ. Dev. 44 (2022), 100767, https://doi.org/10.1016/j.envdev.2022.100767.
[4] D.S. Renné, Progress, opportunities and challenges of achieving net-zero emissions and 100% renewables, Sol. Compass 1 (2022), 100007, https://doi.org/10.1016/j.solcom.2022.100007.
[5] A. Meier, N. Gremaud, A. Steinfeld, Economic evaluation of the industrial solar production of lime, Energy Convers. Manag. 46 (2005) 905–926, https://doi.org/10.1016/j.enconman.2004.06.005.
[6] A. Meier, E. Bonaldi, G.M. Cella, W. Lipinski, D. Wuillemin, Solar chemical reactor technology for industrial production of lime, Sol. Energy 80 (2006) 1355–1362, https://doi.org/10.1016/j.solener.2005.05.017.
[7] R.S. González, G. Flamant, Technical and economic feasibility analysis of using concentrated solar thermal technology in the cement production process: hybrid approach-a case study, J. Sol. Energy Eng. Trans. ASME 136 (2014), 025001, https://doi.org/10.1115/1.4026573.
[8] R. Pitz-Paal, R. Buck, P. Heller, T. Hirsch, W.D. Steinmann, Solar thermal power production, Transition to Renewable Energy Systems, wiley (2013) 307–338, https://doi.org/10.1002/9783527673872.ch17.
[9] N. Sahoo, A. Kumar, Samsher, Solar industrial process heating prospects in Indian cement industries, Advances in Mechanical and Energy Technology. ICMET 2021 in Lecture Notes in Mechanical, Engineering (2023), https://doi.org/10.1007/978-981-19-1618-2_36.
[10] N. Sahoo, A. Kumar, Samsher, A preliminary assessment of solar industrial process heating for cement industry, Advances in Mechanical and Materials Technology, EMSME 2020 in Lecture Notes in Mechanical, Engineering (2022), https://doi.org/10.1007/978-981-16-2794-1_15.
[11] G.J. 'Gus' Nathan, et al., Pathways to the use of concentrated solar heat for high temperature industrial processes, Sol. Compass 5 (2023) 10036, https://doi.org/10.1016/j.solcom.2023.100036.
[12] R. Pitz-Paal, R. Krüger, J. Hinsch, A. Lokurlu, M. Scheuerer, M. Schlecht, G. Schreiber, Decarbonizing the German industrial thermal energy use with solar, hydrogen, and other options–Recommendations for the world, Sol. Compass 3 (2022), 100029, https://doi.org/10.1016/j.solcom.2022.100029.
[13] D.Y. Goswami, Decarbonizing industrial process heat is essential to achieving net-zero goal, Sol. Compass 3–4 (2022), 100028, https://doi.org/10.1016/j.solcom.2022.100028.
[14] G. Moumin, M. Ryssel, L. Zhao, P. Markewitz, C. Sattler, M. Robinius, D. Stolten, CO_2 emission reduction in the cement industry by using a solar calciner, Renew. Energy 145 (2020) 1578–1596, https://doi.org/10.1016/j.renene.2019.07.045.
[15] A. Imhof, Solar Cement Plants - an interesting challenge for business and science, ZKG Int. 53 (2000) 448–457.
[16] N.A. Madlool, R. Saidur, M.S. Hossain, N.A. Rahim, A critical review on energy use and savings in the cement industries, Renew. Sustain. Energy Rev. 15 (2011) 2042–2060, https://doi.org/10.1016/j.rser.2011.01.005.

[17] N.A. Madlool, R. Saidur, N.A. Rahim, M.R. Islam, M.S. Hossian, An exergy analysis for cement industries: an overview, Renew. Sustain. Energy Rev. 16 (2012) 921–932, https://doi.org/10.1016/j.rser.2011.09.013.
[18] J.U. Ahamed, N.A. Madlool, R. Saidur, M.I. Shahinuddin, A. Kamyar, H.H. Masjuki, Assessment of energy and exergy efficiencies of a grate clinker cooling system through the optimization of its operational parameters, Energy 46 (2012) 664–674, https://doi.org/10.1016/j.energy.2012.06.074.
[19] J. Ravindran, S. Krishnan, Studies on thermal analysis of cement rotary kiln based on clinker coating materials on refractories, energy and monetary savings, in: Proceedings of the international congress on recent development in engineering and technology (RDET), Kuala Lumpur, Malasia, 2016.
[20] T. Engin, Thermal analysis of rotary kilns used in cement plants, in: Proceedings of the First Mechanical Engineering Congress (MAMKON'97), Istanbul Technical University, TURKIYE, 1997, pp. 4–6.
[21] A. Atmaca, R. Yumrutaş, Analysis of the parameters affecting energy consumption of a rotary kiln in cement industry, Appl. Therm. Eng. 66 (2014) 435–444.
[22] F. Zeman, Oxygen combustion in cement production, Energy Procedia 1 (2009) 187–194, https://doi.org/10.1016/j.egypro.2009.01.027.
[23] CEMCAP, CO_2 capture from cement production. D4.2 Design and performance of CEMCAP cement plant with MEA post combustion capture. https://www.sintef.no/globalassets/sintef-energi/cemcap/d4.2-design-and-performance-of.cemcap-cement-plant-with-mea-post-combustion-capture_rev1~1.pdf2016. [Accessed 10 July 2022].
[24] CEMCAP, CO_2 capture from cement production. D3.2 CEMCAP framework for comparative techno-economic analysis of CO2 capture from cement plants. https://www.sintef.no/globalassets/sintef-energi/cemcap/d3.2-cemcap-framework-for.comparative-techno-economic-analysis-of-co2-capture-from-cement-plants_.pdf2017. [Accessed 10 July 2022].

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17565-2

Recent advances in the discipline of text based affect recognition

Rajiv Kapoor[1] · **Manali Bhat**[1] · **Nikhil Singh**[1] · **Aarchishya Kapoor**[2]

Received: 8 August 2022 / Revised: 6 October 2023 / Accepted: 17 October 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Sentiment analysis is a part of natural language processing, along with text mining. Over the years, sentiment analysis has become a key study area for researchers and industries all over the world. The major goal of this review is to overview the papers that have found the sentiment of the data under study over the past few years. Also, the various techniques and methods that have enabled us to solve the problem of sentiment analysis are looked into and put forward briefly. The articles are looked into considering the area of sentiment analysis in which the contributions are made, be it sentiment detection, dataset creation, or transfer learning. This would help the researchers to curate advanced and accurate models and methods for analyzing sentiment. Additionally, they would be given a quick rundown of current developments in this area of research. With this in mind, the literature study is conducted taking into account various applications, various machine learning algorithms, the type of data utilized for analysis purposes, and various performance measurements. Challenges and gaps based on all the contributions studied are summarized.

Keywords Affect recognition · Machine learning · Sentiment · Sentiment analysis · Review

1 Introduction

Under natural language processing we have Sentiment analysis which aims at categorizing or classifying pieces of text according to the intuitional sentiment which is derived from it [78]. Sentiment is expression of emotion and opinions by individuals, communities or groups based upon events, topics or issues pertaining to different fields. Sentiment analysis is analysis of the text and its linguistics for identification and segregation of views put forward by a user / author. This would, thereby, summarize the major intent of the author pertaining to the context or the polarity of the context. Thereby being able to guide brands

✉ Rajiv Kapoor
rajivkapoor.dtu@gmail.com

[1] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

[2] Department of Computer Science and Engg, IIT Indore, Indore, India

Published online: 03 November 2023

of the sentiment for their product in the market, the brand reputation etc. The intent of the text could be a suggestion, assertion or judgment.

Opinion mining as well as sentiment analysis is many times used in place of each other. However, some argue it to be a slightly variant from one another. Any opinion that one holds is subjective, it could rely on facts or on feelings and judgments [66]. The process of extracting subjective information from the data or text is sentiment analysis [66]. While opinion mining suggest the process of extracting opinions and then analyzing them is different.

Analysis could be done for varied reasons as shown in the Fig. 1. The sentiment analysis which is based on aspect heavily depends on attributes or properties of an object, product or a service. This helps to understand performance of a particular aspect and the value it has generated. Emotion classification is done with the purpose of identifying the emotion. It is based on machine learning (ML) algorithms and lexicon analysis. Consumer Intent, this involves tracking consumer likes, dislikes to build a pattern which would develop an understanding of the kind of approach consumers have for a particular entity or topic, their behavioral patterns. This could aid businesses in branding their goods and services to appeal to a specific demographic and meet their needs for a more general goal. Polarity of text for understanding views of the author. This purely tries to divide a text into one of the binary or ternary options.

The views of many people for certain events, brands and product can be known by performing sentiment analysis. Thus, the data is gathered and modified for improvements by researchers to do the evaluations. This makes data for analysis a crucial part for analyzing the sentiment. Objective or Subjective data may or may not contain opinions or intent. Thus, in order to perform sentiment analysis, it becomes essential to recognize and extract the features from the source material. This makes feature selection and extraction an important aspect of analyzing sentiment.

Data could possibly be from surveys, social media, microblogging sites, user reviews from shopping sites like Amazon, Shopify, Flipkart etc. It is obtained by crawling the websites using the Application Programming Interface provided by the platforms or using the evaluated and labelled datasets especially for twitter posts or amazon products or movie reviews etc.

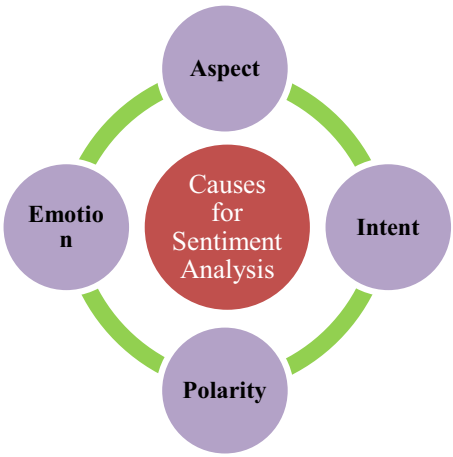


Fig. 1 Reasons for sentiment analysis

Data collected could be analyzed in varied quantities. By looking onto the entire document at once we would be able to derive the summarized polarity of the writer. This polarity could be classified in two or three categories namely, positive, negative (for datasets which are bipolar like yes / no, spam / not spam etc.) and/or both along with neutral. Taking a look at the different statements in the document block one could conclude whether there are or are not any sentiment polarities that oppose each other. As the analysis of only one statement is performed, one might not get a clear picture but would surely get a more detailed idea of what is been communicated within the sentence. Another way to study the data could be by fixating on certain parameters or aspects and considering only that part of the entire document that discusses it. This could help derive emotions people are associating to that particular parameter of the device or news etc.

Analyzing sentiments is challenging due to issues like use of slang, grammatical errors or spelling mistakes or even sarcasm. Understanding if the text is genuine or not is also a challenge. As there are various factors like social, political circumstances etc. that drive sentiments of a user or author.

The paper is divided in the sections as mentioned: 2. Motivation, 3. Classifying sentiment analysis 4. Generalized flow for analysis 5. Study related to sentiment analysis 6. Challenges and gaps.

2 Motivation

Text based sentiment analysis is the study of text and its linguistics in order to identify and categorize the many points of view expressed by a user or author. Sentiment and emotion analysis approaches can be classified into three categories: deep learning, lexicon and machine learning. Although the knowledge base for recognizing emotions from voice/speech and other multimodal approaches is extensive, there is a significant gap in study for texts [107, 123]. This is due to the fact that, unlike multimodal approaches, texts may not depict unusual emotional cues. Furthermore, the difficulty of recognizing emotions from brief texts, emoticons, and grammatical faults, when combined with the constant emergence of new words as a result of language dynamics, could be crippling [108]. A lack of suitable emotion terminology is another issue. There are currently no domain-specific emotion lexicons available [109]. Only a careful selection of tools can guarantee a precise and legitimate method to gauging sentiment from text, while domain knowledge and validation are still required components in sentiment analysis applications. To highlight such issues and recent trends various reviews on the sub-discipline of sentiment analysis have been published covering domains like context based [110], multimodal sentiment analysis [111, 112], Deep learning based [113], Deep learning for aspect-based [114], artificial intelligence based [115], review on specific languages [116, 117], sequential transfer learning [118]. As a result, it is critical to comprehend every component of text-based sentiment analysis in order to increase the quality of affective computing in the upcoming years. The following factors prompted us to conduct this study:

1. The first step is to investigate the various components of text-based sentiment analysis. The representation of the targeted area varies depending on the aspect. They each have their own set of advantages and disadvantages. This fact prompted us to investigate the qualities of several types of sentiment analysis schemes in order to identify relevant features for sentiment analysis.

2. As from Table 1 it can be clearly seen that recent advances in the domain of sentiment analysis uses deep learning schemes more as compared to other methods. In this review we are exploring and analyzing these tactics for text based affect recognition because of the introduction of innovative optimization functions and approaches.

3 Classifying sentiment analysis

Results from the study pertaining to sentiment analysis largely rely on the cause for which analysis is being performed. This makes categorizing sentiment analysis a vital task. Broadly, in order to categorize sentiment analysis, we have considered three major aspects viz., Method based, result based and application based. These categories are narrowed down by categorizing them further.

Opinion mining and affective analysis can be used to categorize sentiment analysis based on results. Opinion mining deals with the text's polarity. These polarities could be dual (positive or negative) or triple (positive, neutral, negative) Emotion based analysis is majorly based on the 8 emotions of human sentiments like happiness, sadness, anger, fear, anticipation, surprise, disgust, trust.

Sentiment analysis can also be classified as Lexicon-based, Machine learning based, and mixture of both the methods. The Fig. 2 shows all possible ways in which sentiment analysis could be categorized for better understanding.

Lexicon based classification considers words and marks their intensity and polarity. Approaches which are based on Lexicon have low utility when worked on unseen data [74]. Polarity defines if the word is biased towards a positive side or negative side or neutral side and intensity defines the degree to which the word is biased towards the polarity. For example, the words sad and very sad have the same polarity but the intensity of the words could differ.

When the text to be analyzed is fed, the words of the input text are compared and weighed to those of the sentiment lexicon to get the output sentiment. Glove [84], Word-2Vec [83], and FastText [85] can also be used to get the word vector features, embeddings. More often lexicon-based approach deals with opinion mining type of classification. English language has a huge number well-structured lexicons but that isn't the case for other commonly used languages. Recently, there have been studies pertaining to analysis for non-English languages particularly Arabic, Chinese, Urdu, Russian etc.

Machine Learning Models are divided as unsupervised and supervised algorithms. There could also be a blend of both of these algorithms to form a hybrid or a semi-supervised algorithm. The supervised algorithms are efficient enough for labelled training datasets. A huge amount of annotated data is required for performing supervised learning [53]. Various types of supervised machine learning methods such as Naive Bayes (NB) [79], Support Vector Machines (SVM) [82], Maximum Entropy (ME), Logistic Regression (LR) [81], etc. and feature combinations have been applied in sentiment analysis research [73]. Classifiers and Regression models all are part of supervised learning models with labelled data. Unsupervised algorithms perform well for real-time data that is not classified or labelled. Here, text is at first converted in form of vectors to obtain the features of the text. Clustering based models like K-means or K-clustering or dimensionality reduction techniques for feature selection and features extraction are unsupervised.

Table 1 Summary of different aspects of the literature studied

Ref	Year	Area of focus/Application	Scheme used	Dataset	Classification scheme	Performance parameters
[1]	2020	Topic level sentiment classification	Machine Learning	Twitter data—# Brexit	Topic-level Sentiment Classification	Accuracy- 92%
[2]	2020	Bipolar sentiment classification	Deep Learning & Lexicon based	dangdang.com, Chinese e-commerce website	BiGRU	Accuracy- 93.5% Precision- 93% Recall- 93.6% F1- 93.3%
[3]	2020	Reputation generation for online entities and visualize reputation for supporting customers though the decision-making process	Machine Learning	IMDb, TripAdvisor and Amazon websites are scrapped	–	Accuracy- 88.81%
[4]	2019	classifying tweets into several ordinal classes using machine learning classifiers	Machine Learning	Twitter dataset by NLTK corpora resources	SoftMax, SVR, Decision Tree, Random Forest	Accuracy- 91.81 MAE- 0.154 MSE- 0.155
[5]	2020	constructing a Danmaku sentiment dictionary and using it along with Naïve Bayes	Machine Learning	Danmaku video platform's live commenting function	Naïve Bayes	Accuracy- 0.882 Recall- 0.783 F1- 0.823 Precision- 0.867
[6]	2020	a feature ensemble model is built consisting of lexicons, semantics, position, and polarity	Deep Learning	Twitter dataset	CNN	Precision- 0.81 Recall- 0.82 F1- 0.81
[7]	2020	Aspect-Based Sentiment Analysis Using Inter-Aspect Relations	Deep Learning	SemEval-2014 Task 4	Recursive & Recurrent Neural Network	Accuracy- 81.38%
[8]	2019	An Aspect-Level Classification with the help of Attention in hierarchical form	Deep Learning	SemEval-2014 -4th task, SemEval-2015-5th and 12th task	CH-HEAT	Accuracy- 92.4%
[9]	2018	Analysis of semantic fuzziness to get opinion on Chinese texts	Machine Learning	Dataset from Virtual reviewers. Meant for exploring collaborative movie reviews	–	–

Table 1 (continued)

Ref	Year	Area of focus/Application	Scheme used	Dataset	Classification scheme	Performance parameters
[10]	2018	paper focuses on refining the pre-trained word vectors using intensity scores from lexicons	Deep Learning	Stanford Sentiment Treebank, SemEval dataset	CNN, DAN, Bi-LSTm, Tree LSTM	–
[11]	2018	sentiment lexicon in introduced into LSTM. Also attention mechanism without a target is added for general sentiment analysis	Deep Learning	IMDB, Yelp2013, MR, NB4000(Chinese), book4000(Chinese)	LE-LSTM	Accuracy- 80.8%
[12]	2019	Create an alarm when the performance of ML model starts deteriorating with new datasets	Machine Learning	Amazon datasets introduced by Ghaddar and Naoum-Sawaya [54]	Random Forest Class, Multinomial Naïve Bayes	–
[13]	2019	an extended dictionary consisting of the basic, fields and polysemic sentiment words is built	Lexicon Based	sentiment dictionary combines National Taiwan University's Dictionary of sentiment, Hownet's sentiment dictionary and popular sentiment words from internet, data crawled from Jingdong, Ctrip and eLong websites	Naïve Bayesian	Precision- 0.861 Recall- 0.843 F1-Value- 0.853
[14]	2019	constructing multiple sentiment dictionaries, sentiment classification in 3 categories	Lexicon Based	crawler software crawls the Chinese blogs with ''#Short Video Rectangle#'' and ''#'' I am not a drug god '' burst red lead social hot discussion#'' data	–	–
[15]	2020	Addressing issues of over-fitting, vanishing gradient and improve accuracy and proposing BiGRU and CNN based model	Deep Learning	Chinese product reviews of a shopping website and the Chinese hotel reviews from the travel app	BiGRU & CNN	Accuracy- 94%
[16]	2019	The dimensional hierarchy of the topic of tweets is found	Machine Learning	Twitter REST API	–	–

Table 1 (continued)

[17]	2019	multi-domain adversarial network (MDANet) is proposed to classify text	Deep Learning	word embeddings were trained from Sina Weib, Amazon product review corpus by [56], consumption intention corpus [55] and sentiment analysis corpus [56]	MDANet	Accuracy- 89.98%
[18]	2020	using context-dependent part-of-speech, sentiment lexicon is built	Lexicon Based	long text review dataset (LMRD) [57], (MRD)[58]	–	Accuracy- 82%
[19]	2019	Using sentiment specific word embeddings from Arabic tweets for sentiment classification using a feature ensemble model	Machine Learning	SemEval-2017 Arabic tweet dataset, AraSenTi-Tweet dataset [59]; ASTD dataset[60]	SVM with LibLinear	–
[20]	2020	weakly supervised learning is used for aspect-level sentiment analysis	Deep Learning	Coursera	CNN + Fast Text	F-Score- 93.3%
[21]	2018	Identification of emotions in Chinese metaphorical texts is done	Lexicon, Machine & Deep Learning	sentiment corpus of Dalian University of Technology	–	Accuracy- 84.1% Recall- 86.1% F-Value- 85.2%
[22]	2020	Fine grain aspects are mined using deep learning model	Deep Learning	Laptop, Restaurant dataset from SemEval2014 and Twitter datasets	BERT	Accuracy- 91.6% F1-Score- 88.01
[23]	2020	feature selection by modifying the fitness function of the Genetic algorithm and Wrapper approaches	Machine Learning	Scrapping twitter for apparel brands names as keywords	–	Accuracy- 92%
[24]	2019	Sarcasm detection using soft attention-based bi-LSTM and convolution neural network	Deep Learning	SemEval 2015- 11th Task, twitter's random tweets	sAtt-BLSTM	Accuracy- 97.87 & 93.71 Precision- 92.19%
[25]	2019	backdoor attacks happening on text classification based on LSTM	Deep Learning	IMDB movie reviews dataset	LSTM	Accuracy- 84.5%
[26]	2020	figuring out polarity of social media during elections using neural networks	Deep Learning	Scrapping twitter data using relevant hashtags 2016-US elections, 2018-Italy elections	IOM-NN	–

Table 1 (continued)

[27]	2018	prediction for stock market composite index movements	Deep Learning	stock market-data for 1.1.2015 to 31.12.2016, from Wind, financial information service provider in China. Xueqiu	–	Accuracy- 0.601 F1-Score- 0.621
[28]	2020	Classification of App Reviews using Convolutional Neural Network	Deep Learning	reviews dataset from Google store and Apple store of different categories	MLP-ANN, Decision Tree, Naïve Bayes	–
[29]	2018	Observing evolution of behaviour based on topics discussed	Machine Learning	Twitter datasets	–	–
[30]	2018	Sentiment classification of reviews and dealing with ambiguity of words with different polarity in different context	Deep Learning	Amazon reviews from categories like kitchen product, DVD, book, electronics	Cross modal hypergraph	Accuracy- 81.65%
[31]	2019	rumor detection in Chinese	Deep Learning	Sina Weibo datasets,	–	Accuracy- 95.2% Precision- 0.963 Recall- 0.940 F-measure- 0.952
[32]	2020	Aspect based fashion recommendation	Lexicon Based	the Clothing, Shoes & Jewellery data of Amazon 5-core and the reviews from US online retailer	–	–
[33]	2019	method to add prior information in the RNN models	Deep Learning	datasets Yelp 2013 and Dangdang	RNN, LSTM & Bi-LSTM	Accuracy- 56.2 MSE- 0.656
[34]	2019	Identifying factors influencing online product sales on online platform	Deep Learning	Taobao.com	–	R^2- 0.812 RMSE- 3.06E-02 MSE- 3.66E-03
[35]	2018	sentimental transfer learning for stock prediction model	Transfer Learning	Hong Kong Stock Exchange data (2003–08), news archive from FINET5 from Jan. 2003 to Mar. 2008	SVM	–

Table 1 (continued)

[36]	2020	tweet analysis platform to study psychological needs (during covid-19)	Machine Learning	tweets scrapped from Twitter	SVM	Accuracy- 93.56% F-score-0.93 Recall-0.93 Precision-0.93
[37]	2020	keyword matching for studying the impact of emotional behaviors to user Emotion scores	Lexicon Based	Twitter scraping (Japanese)	–	–
[38]	2020	Sarcasm detection using multi-dimension question answering (MQA) network and building the conversation context information using attention and BiLSTM	Deep Learning	Internet Argument Corpus (IACv2)	–	–
[39]	2018	Sarcasm detection using multi-head attention network based on Bi-LSTM	Deep Learning	self-annotated corpus SARC 1 dataset	MHA-BiLSTM	Precision- 72.63% Recall- 83.03% F1-Score-77.48%
[40]	2019	extract rule-based semantic relation using Chinese sentiment lexicon	Lexicon Based	Baidu interpretation, modern Chinese dictionary	–	Precision- 97.43%
[41]	2018	Aspect level sentiment classification using CNN + Bi-LSTM	Deep Learning	SemEval 2014, one is from Laptop domain and another is from Restaurant domain	CNN & Bi-LSTM	–
[42]	2020	Detecting hate speech using ML algorithms and developing English corpus from South African tweets	Machine Learning	twitter data from 5.5.2019 and 13.5.2019 using Twitter Archiver.8 (plugin for Google Sheets)	–	Accuracy- 67.1 TPF- 89.4
[43]	2018	Football-Specific Tweets for sentiment classification	Machine Learning	UEFA Champions League 2016/2017, FIFA World Cup 2014	SVM, MNB	Accuracy- 0.79 f-score- 0.80
[44]	2018	spam detection at tweet-level	Machine Learning	HSpam14 data set (15 days data collected)	Multi-Classifier	Precision- 95%

Table 1 (continued)

[45]	2020	Paper focuses on selecting semantic features to perform text classification	Deep Learning	Stanford Sentiment Treebank (SST1, SST2), MPQA dataset, IMDB, Movie Review, TREC datasets, Subjectivity dataset, customer reviews for products	SRCLA	–
[46]	2019	using target information with BERT for sentiment classification	Machine Learning	SemEval-2014 task 4 restaurant and laptop data, twitter dataset	BERT	Accuracy- 85.27% F1-Score- 79.61%
[47]	2018	Predicting Vehicle Sales	Machine Learning	Bureau of Economic Analysis, U.S. Commerce Department, Yahoo Finance, monthly Tweets using API	Naïve Bayes	MAPE- 3.96% WAPE- 4.008% NMAE- 0.108
[48]	2020	exploring the feasibility of deep semantic representation, improve the performance of aspect extraction	Rule Based	SemEval-2014 and SemEval-2015 datasets and also datasets relating to digital cameras, cell phone, MP3 player, and DVD player. Computer wireless router, and speaker	–	Precision- 89.2% Recall- 97.3% F-Score- 90.2
[49]	2018	Subjectivity classification using recurrent CNN with the multi-features combination	Deep Learning	STS-Test dataset, SemEval2014, 2015,2016 dataset, SE-Twitter dataset, SS-Twitter,	DRCNN	Accuracy- 88.92%
[50]	2018	Filtering Mechanism for Recommending Posts in a Social Networks using Linguistic Features	Lexicon Based	3 open Facebook groups	–	–
[51]	2019	Aims to classify text using Bi-LSTM	Deep Learning	crawled from Ctrip (https://www.ctrip.com/)	Feedforward Neural Network	Precision- 91.54% Recall- 92.82% F1-Score- 92.18%
[52]	2020	Aspect based sentiment analysis	Deep Learning	SemEval and ACL Twitter Dataset (2014)	LSTM	Accuracy- 81.96% F1- Measure- 75.02
[53]	2020	Aspect based sentiment analysis	Deep Learning	Task 4 of SemEval 2014, 2017 datasets	ADeCNN	Accuracy- 85%
[120]	2021	end-to-end sentiment analysis approach to handle negations,	Deep Learning	Amazon reviews dataset	SVM, ANN, RNN, NB	Accuracy- 95.67%

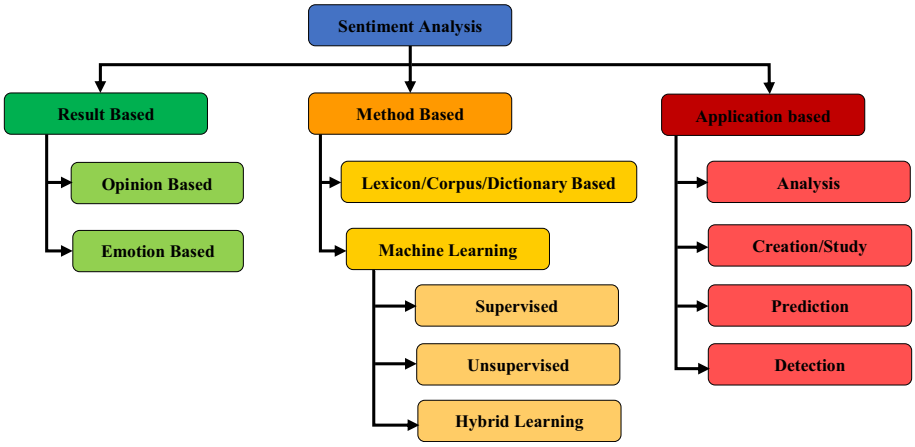


Fig. 2 Classification of sentiment analysis

Major drawback of the Lexicon-based classification is that there could be less values assigned or inappropriate values assigned to words. Although, Machine learning algorithms deal with this quite well they need to have a huge training data to correctly classify the sentiments.

Sentiment analysis has varied applications. Web 2.0 upgrade facilitated blogging and social networking sites which has increased the applications of sentiment analysis in many folds. Thus, by considering all of these as categories we would be able to view from a broader perspective the areas to which sentiment analysis is applied and the areas to which it could be applied in further studies.

Sentiment analysis has applications ranging in diverse areas [61, 62]. Namely, it's been used for marketing and sales by companies and brands for brand managements, understanding voice of consumer, market research, product analytics eyeing the competitors etc. Social media platforms and online shopping sites could be used for this purpose as these are platforms where consumers and potential consumers from different demographics come together and discuss their experience, intent regarding the product or service and influence each other to form an overall perception of the product or the company.

In addition to this, sentiment analysis can be used to determine how the general public feels about a certain piece of legislation, breaking news, etc. This would help the public to express their feelings and believes as well as the policy makers on how the new policies are being perceived. It could also be used to predict the election poll results and know the public choice well in advance. It can also be used to develop a smart society for public services [119] Further advanced study in this regard could even replace the traditional polling system the way we know it today.

Analyzing sentiments could also help in performing surveillance to detect criminal intelligence on various social media platforms. Currently, there is need for performing surveillance, detecting and isolating cyber-trafficking [63]. A study dealing with this issue monitors cyber-trafficking social media discussions, the authors proposed a system that collects, analysis, and visualizes social media content, the linkage among users and their temporal features. This would help in keeping cyber bullying and cybercrime in check. Also, at many times the social media users knowingly or unknowingly put up sensitive information which could be misused or traded by various organizations such as advertising

companies, or criminals. [64]. this makes it an essential task to focus on identifying privacy leaks in the content of texts put out on social media.

Additionally, sentiment analysis may be used to examine market patterns or the sentiment of a certain share. Alongside, it would be useful for stock forecasting and might assist purchasers choose the most promising stocks.

Thus, we could broadly classify sentiment analysis applications in four categories namely: analysis, creation/study, detection and prediction as mentioned in Fig. 4. Here, the creation / study deal with all the works which have put their focus on building of corpus / lexicon or some kind of a dataset. Analysis group deals with all the study performed to analyze some data and derives something out of it. Detection group deals with all the researches which have worked to detect emotions or sentiments and so on, whereas, Prediction categories all the study which culminates to forecasting and predicting.

Apart from all of this there is a method which lets researchers utilize the efforts of their fellow researchers. This is transfer learning [88] which takes one type of training data / knowledge and utilizes it for data of a different type which shares some commonalities. This helps in generalizing the outcome and lowers the labeling effort [64].

4 Generalized flow for analysis

Majorly the sentiment analysis happens in a format as shown in the flow diagram. However, based on kind of dataset used, model parameters and desired output there could be slight changes in the flow mentioned. The type of machine learning model also determines the flow of the process.

At first, data is procured from a well-known dataset or by the means of web scrapping. As per study of papers it is been found that a lot of papers heavily relied on the data gathered by user generated content (UGC) that comes from the social networking sites and social media. The UGC data on the social media and user reviews commonly showcase subjective texts as the writers freely post their opinions, the most popular one being twitter [76, 78]. Data collected from twitter is preferred for it being user generated, contains a variety of opinions from people of different demographics, huge choices of topics for data scrapping. Other than twitter, data pertaining to e-commerce websites, movie reviews etc. is chosen.

By making use of different data sources, some studies have managed to cover a wider range of materials [78]. The data which is gathered has to be converted to a usable format. For this purpose, it is cleaned, preprocessed to desirable format that can be used for further processing. This stage incorporates a number of minor tasks. Tasks like tokenizing viz., breaking text into pieces of smaller size text, stop word filtration (words that don't contribute to final goal example, repeating characters etc. are discarded) and stemming or lemmatizing (converting word to a base word) and splitting dataset. Dataset can be split into 2 to 3 kinds namely, training and testing dataset and / or validation dataset. In order to train the sentiment classifiers, sentiment reversal prediction model training dataset is used, for testing the performance of the classifiers as well as the sentiment reversal prediction model, the validation data is used, for evaluating the blind performance test dataset is used [67]. In short, the training dataset is used by the model for learning purposes while the testing dataset is used for getting an unbiased result from the trained or learned model. Validation dataset is used to guide while the model is getting tuned to the hyperparameters.

Now we discuss a crucial step for sentiment analysis which is responsible to decide what parameters to lookout while deciding on a sentiment. For supervised learning, we carefully choose features that would be enough to correctly distinguish between various categories of classification. While for unsupervised learning, feature extraction is performed prior to passing the data to the model. The features could be any of the parts of speech like the adjectives or the adverbs that would add on the emotion being put out, it could even be words that voice out opinions like don't, dislike, never, ugly, bad good, like etc. or even negation words or phrases which completely change meaning put forward by the statement, the phrases whose meaning could even suggest irony (Fig. 3).

Before one could do feature selection a major step of feature extraction needs to be performed. It is responsible for the performance of the model. While feature selection is selection of features which would contribute maximum to the output or the ones with have maximum chances of driving a particular kind of output in which we might be interested, feature extraction creates a brand-new feature set that has the essence of the most useful data. One can perform these by means of applying unsupervised methods

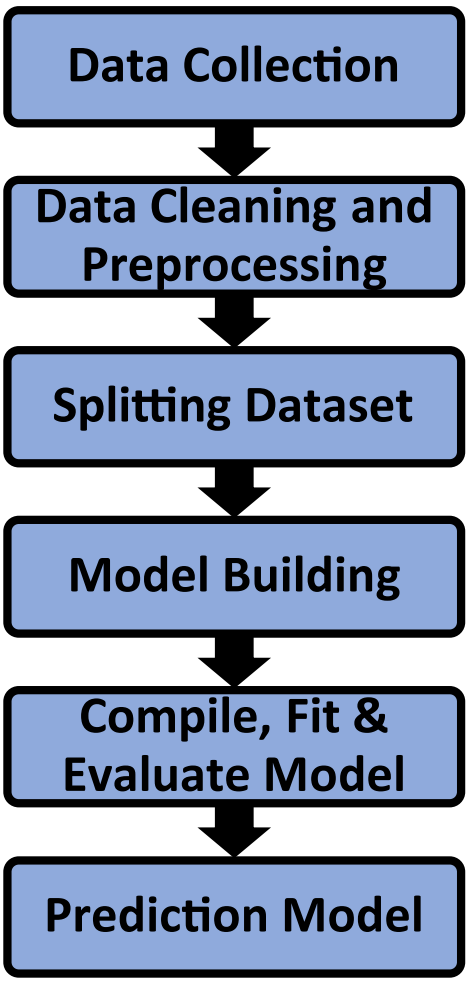


Fig. 3 Generalized program flow

like the PCA or the supervised methods like the LDA. Text-based features include the Bag-of-Words model (BoW) and N-gram Language Model extracted using the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme [36] could be used for this purpose. One could also visualize the results as well in order to get a pictorial insight into the data worked upon.

After this, the vector of words with label is formed and then the model is built. Then, the training data is passed through it. Later on, we decide a model architecture, which means deciding the number of hidden layers, the activation functions, embedding sizes etc. This is called compilation. After deciding the model architecture, we train the model to get all the parameters to the best value in order to map the inputs and outputs. This is data fitting. Lastly, the model is used for feed-forward passes for the prediction of new inputs. This being done the results is evaluated on various parameters.

5 Study relating sentiment analysis

In order to truly understand the direction of recent study revolving around sentiment analysis, careful selection of publications over the past few years is looked upon.

While going through all of the chosen papers it has been found that the study is divergent across various issues that still prevail and bother the researchers, the industry and at times, the content consumers. Vivid areas which have been a recent target of study among researchers across the globe is as shown in the Fig. 4. Broadly we happen to categorize the area of study based on the major focus performed by the authors.

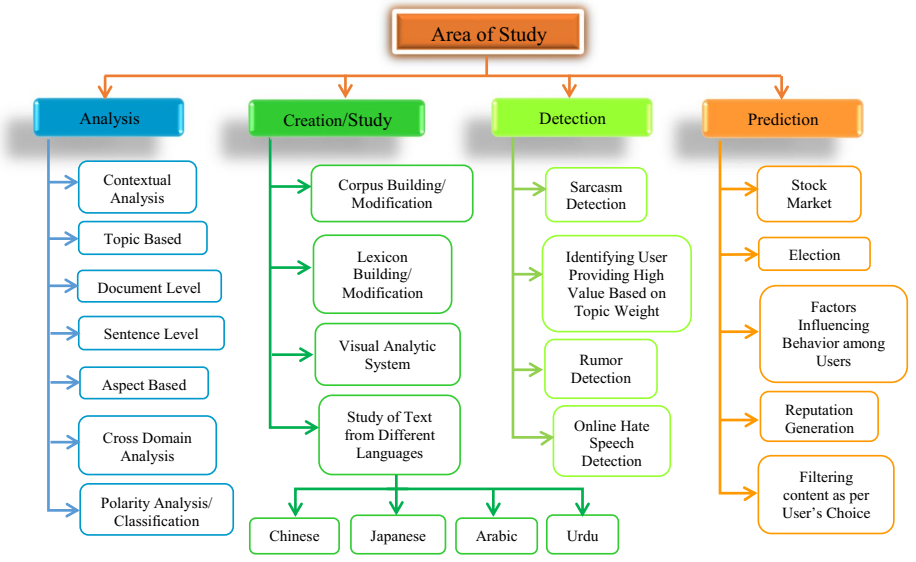


Fig. 4 Different areas of study

5.1 Areas touched by papers under this study

The categories under which the classification is done are analysis, creation/study, prediction and detection as shown in Fig. 4. The areas which have been put under the category of prediction majorly focus on forecasting the outcome of a particular task or an event; including prediction of a stock or stock market sentiment, results of an election, trying to predict the data which might interest a user and suggesting it etc. For those categorized under detection are driving to find out a particular kind of behavior or pattern. Be it sarcasm detection, rumor detection, detection of hate speech found online or recognizing a contributor based on a given topic etc. Identifying the role of a social media user to analyze and understand social media relations, understanding and forecasting user behavior and the interactions users conduct with one another [93, 100, 101].

Another group is formed comprising of methods for building or modifying the corpus, lexicon, dictionary as well as studies performed relating to understanding sentiment based on different languages like Russian, Arabic, Urdu, Chinese, Japanese etc. This study could comprise of building a dataset that could be used for analysis or using methods to detect sentiment and so on. We also have come across studies which propose CONCEPTVECTOR, an analytics aide which helps users to interactively build a lexicon for custom concepts [32]; Social media users put up many spatial and temporal-based data including texts, images and videos which had given rise to big data [36]. We came across a foresighted tweet analysis platform NeedFull, which was developed in order to help the government authorities to understand psychological needs of the people by creating this tweet analysis platform during the tough times.

Apart from these categories a collection of all the study performed with intention for analysis is grouped together. Analysis could be done considering the length of the text. By considering entire text document at once or by considering the sentences one after the other or by analyzing text based on topics or by considering a particular aspect or feature. Analysis could also be done to figure out the context of the text. Analysis of polarity is also a vividly researched topic. Polarity analysis could be binary or ternary or based on 7 or 8 emotion classification. Another hot topic is cross domain analysis which aims at facilitating a labeled source domain for an unlabeled target domain.

Different kinds of methods were proposed by researchers to tackle these issues. All of these are discussed in brief in Table 1.

5.2 Dataset used

The kind of dataset chosen has a crucial importance in the study. As the dataset is chosen for training of the model, validation and testing as well data selection heavily impacts the performance of the model and the kind of prediction it makes. Thus, choosing dataset is one essential step.

The data could be coming from sources of different kinds, namely, user generated content from the social media, reviews for services or products, news from tv or websites, mix of data from different resources. From the study of different kind of works pertaining to sentiment analysis, a few kinds of datasets dominate majority of the researches. In this section such datasets have been discussed.

There are several ways to obtain a dataset. Either by using the usual datasets built by previous studies or by web scrapping social media sites or websites or datasets from

various competitions. The datasets commonly used are the datasets by Stanford University like the twitter sentiment dataset, Stanford sentiment treebank dataset, IMDB movie review datasets, Movie review datasets etc.

Datasets for stock market prediction and analysis are obtains by scrapping financial sites like Yahoo finance and stock exchanges like Shanghai, Hong Kong etc. Dataset for analyzing sentiments associated with products, services or brands are obtained by scrapping websites. The most popular one being Amazon.com for sentiment analysis in English language. For Chinese language, Dangdang.com- e-commerce website, Ctrip—travel agency company website, Danmaku videos' live commenting etc. The app stores like Google's play store and apple store also are scraped to draw reviews. Datasets are also procured from microblogging websites like Twitter for procuring tweets in English language, Sina Weibo -Chinese microblogging website etc.

Twitter has an application programming interface (API) that allows researchers, students to collect tweets. However, the limit is up to the previous seven days for data collection. A premium account, which charges a lot can provide data older than seven days [68]. Also, Twitter has features for filtration to get the data required.

A generalized way for scaping data from websites begins by choosing the target website which needs to be crawled for scraping the data. After that it is essential to collect the URL of all the pages in that website. This helps fixating the exact location from where the data would be collected to build up the dataset. Now, to procure the data a request to these URLs is made. The request is responded by providing HTML of the page. In order to find the data in HTML locators are used and the data that is found is saved in structured formats like JSON or CSV formats.

Competitions and workshops like SemEval (Semantic Evaluation) which aim at advancing semantic analysis and help in creating annotated datasets for different challenging problems pertaining to NLP semantics. Every year's workshop consists of group of tasks worked on by different teams are presented and compared. The SemEval dataset has now become the extremely popular dataset for the analysis of sentiment [53]. Of all the papers under study, datasets from different tasks of SemEval-2014 are the most frequent ones followed by SemEval-2015 and 2016.

5.3 Features and feature processing

Certain keywords or words play an important role in driving the sentiment of the entire text. Grouping them together would ease the process of deriving the sentiment. These are identified as features. Of the variety of features which could be derived from the text, picking the ones which have the most impact in analysis of the sentiment is a big task. Reducing the number of input variables to only the most useful ones for a model in order to predict a target is feature selection. Thus, feature extraction and selection are of utmost importance. Feature selection is also related to dimensionality reduction wherein the input data is projected to a lower dimensional feature space. This could also be done by choosing a smaller set of important features [69]. Some of such methods are discussed briefly namely, chi-square, information gain, LASSO, etc.

The Chi-square test method tests independence of 2 events. It is calculated between each feature and target and the features which give the best chi-square scores are chosen.

The information gain measures information that we could get from a feature of the class. Using this we could know about the features which can meaningfully discriminate between the classes which are to be learned.

LASSO stands as Least Absolute Shrinkage and Selection Operator and is a linear L1 regularisation technique that estimates sparce coefficients. It is useful in contexts where solutions with fewer parameter values are preferred.

Correlation coefficients are also used for feature selection as good variables are well correlated with the target. The variables however must not be correlated among themselves. If they are correlated then they don't add information so one of them must be utilised for selection. Heatmaps can be used for this purpose.

The feature selection methods build focus on the existing dataset and throw away the not so important ones. The existing features are looked into and no new features are created. The Feature extraction techniques discards the original features after creating brand new features from the features present, thus, reducing the overall number of features in the dataset. The new dataset which consists of reduced features has a gist of the information which was present int the original data. In feature selection, the new feature set is formed by taking a part of the original one. The new feature space of feature extraction can be transformed from the original space by applying linear or nonlinear combinations [69]. Some vividly used feature extraction methods like PCA, LDA, bag of words, TF-IDF are discussed briefly.

PCA removes redundancy that exists between correlated features in the dataset [70]. Principle Components Analysis (PCA), a dimensionality reduction linear technique and an unsupervised learning algorithm. The higher dimensional data is mapped to data in a lower dimension but the data in the lower dimension ought to have maximum variance. New orthogonal features are linear combination of all input features. These are identified and ranked depending upon the amount of variance of the input data they can explain.

Linear Discriminant Analysis (LDA), is one of the most popular unsupervised algorithms. It is a dimensionality reduction technique which is responsible for modelling each document as a mixture of topics [71]. It tries to increase the distance in between mean of class and reduce the spreading in the class itself. Maximizing distance between means of each class after the data is projected in lower dimensional reduces overlap between different classes.

Bag of words uses words as features for training the classifier. After text processing a vocabulary of unique words is formed. A matrix of features is created; each entry in the matrix shows the absence or presence of the word, this is called as text vectorisation. Feature representation based on bag of words approach faces the problem of data sparsity [74]. Like bag-of-words we also have n-grams were a group of words are considered.

Another feature extraction technique is TF-IDF viz. term frequency-inverse document frequency which is a widely used statistical technique for measuring the importance of a term in a given text from a document corpus [72]. It raises proportionality of TF-IDF value according to the amount of times a word appears in the text and decreases with total number of documents in corpus that consists of that word. Term frequency specifies the count of the term in the text. Inverse document frequency measures if a term is infrequent of frequent across the corpus document.

5.4 Models used for sentiment classification

To perform sentiment analysis on a variety of datasets, applications it is essential to model the framework so that correct outcome is achieved. Authors and researchers have been keen on working out several models which could bring an apt solution to this problem. Over the years of research, various findings have been put out to facilitate growth in this realm of

study. It is however interesting to note that a few of these Machine learning methods have been vividly used and are found to be effective. This segment of the study aims at throwing light on these methods. The kinds of such methods which have been discussed further are jotted in the Fig. 5.

The most frequently used algorithms from the recent studies are classified in eight different groups namely, classification, regression, Bayesian based, neural networks, ensemble methods, decision trees [80], feature selection and extraction (also includes dimensionality reduction), deep learning. Of these, classification algorithms like SVM [82], Naïve Bayes [79] etc., regression algorithms like logistic regression [81], ordinary least square regression etc., decision trees [80] algorithms are a kind of supervised algorithms. Some of the Ensemble methods are random forest, boosting, adaboosting etc., whereas the Bayesian methods are naïve Bayes, Multinomial naïve Bayes, Bayesian network, Gaussian naïve Bayes etc.

The feature selection as well as extraction methods have been discussed in a brief. Neural network and deep learning algorithms are also amongst the most vividly used algorithms for sentiment analysis, namely, feedforward neural network (FFNN), Radial Basis Network (RBF), Long Short Term Memory (LSTM), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) [122], Gated Recurrent Unit (GRU) etc.

Support Vector Machines (SVM) [82], a supervised learning method focuses on finding a hyperplane in a P-dimensional space (P is the number of features) which distinctively classifies the data points. The chosen hyperplane must provide maximum distance between datapoints of different classes. The datapoints which are near the hyperplane greatly influence its location and tilt. The overall goal is to increase distance between these support vectors and hyperplane that is, to find the optimal hyperplane. This technique has been widely utilized for classification since it converts the nonlinear problem

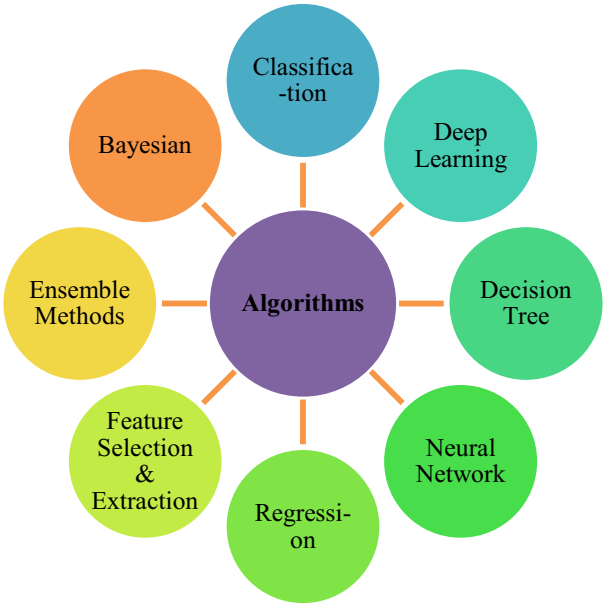


Fig. 5 Types of vividly used algorithms

to quadratic programming. Also, SVM gives out a unique and globally optimal outcome [65].

Naive Bayesian [79], a Bayesian theorem based classification method [13] is supervised learning algorithm. It is a probabilistic model which makes assumption that each and every feature makes equal and independent contribution to outcome. This would ensure conditional independence in between each pair of features which are given the value of the class variables. Different Naïve Bayes classifiers are used namely, Gaussian Naïve Bayes, Multinomial Naïve Bayes (for document classification), Bernoulli Naïve Bayes. Of these Multinomial Naive Bayes (for document classification) has feature vectors showcasing frequencies with which events are generated by multinomial distribution. Bernoulli NB has features which are independent Boolean. Gaussian NB has continuous valued features which are distributed as per Gaussian distribution.

Bayesian networks [35] model conditional independence relations among variables which learn the directed acyclic graphs and represent factors of the joint probability distribution over the relevant variables [1]. So it is basically, a probabilistic graph-based model which has each edge as conditional dependency and each node as a unique random variable. The BN uses Bayesian inference for probability.

Convolutional Neural Network (CNN) works on 2D data; however, it is also suited to work with 1D and 3D data. The method gets its name from an operation called convolution which is performed in the layers of the neural network. The basic CNN model is build using convolutional, pooling and fully connected layer [90]. The convolution, a linear operation involving element-wise multiplication of array of weights (filter) with filter-sized patch from the array of input, which is then summed up to obtain a single value. Convolutional Neural Networks (CNNs) are vividly employed for text classification and tasks which learn extracting hierarchy of crucial text elements [87].

Bi-GRU is Bidirectional Gated Recurrent Unit is two way GRU wherein output depends on the forward and backward states. It solves the problem of unidirectional GRU thus, the final output more accurate [90]. It consists of two GRU. One of the two takes input in forward direction and other in backward direction. It has input and forget gates and hence can be seen as a bidirectional Recurrent NN.

LSTM is a better recurrent neural network which overcomes the long-distance dependence problem [75]. For this reason, they are preferred for applications that require sequence prediction. LSTM reduces the effect of vanishing gradients and exploding gradients on a large scale. LSTM successfully predicts the next outcome as per information of the previous time [75]. But there are times when the current outcome relies on the past and future states.

In such times we need Bi-LSTM. Bi-directional Long Short-Term Memory is a 2-way LSTMs, that identifies and holds bidirectional semantic dependency in order to better memory [86]. Of the two, one takes input in forward direction while the other takes input from the back. This increases the total amount of data which is being made available for the network. Thereby causing improvement in context available to algorithm.

Recurrent Neural Network (RNN) has an internal memory and is recurrent over time. CNN does not have the memory to capture long-distance dependencies which exist in the texts which RNN has [90]. It does the same function for each input data while the output depends on current input and the previous computation. On generating output, it copies it and sends it back to recurrent network. RNN, a generalised feedforward NN uses its memory to process input sequences which are related to one another. Recursive NN (RNN), on the other hand are created on applying same set of weights recursively

on structured inputs, particularly directed acyclic graphs in order to obtain structured prediction. This model are both non-linear and adaptive.

Random Forest (RF) is a forest of many decision trees. The dependance is diversified at various decision trees and thereby on sub-samples of the document [77]. RF creates decision trees of data samples. These trees operate as ensemble and each of them work on input and give out result. The class which is highly voted is chosen as the outcome for the forest. This supervised learning algorithm does classification as well as regression. It reduces overfitting by averaging the result and not concentrating on the outcome obtained from a single decision tree.Logistic Regression (LR) [81] is a Supervised Classification Algorithm which builds regression model and forecasts if the data point would belong to a category. The output is modelled using a sigmoid function and takes discrete values for input which is fed to the model.

Bidirectional Encoder Representations from Transformers (BERT) [89] has encoders and decoders superimposed with layers of transformers. It is a costly in terms of expense and has a high configuration with complex training procedure and large training time [91]. It makes use of transformers, attention mechanism to learn contextual relation between words. It is used for NLP trainings as it is based on network.

Boosting aims at transforming a weak learner into a strong one [92]. This is put in practice by making a model work on correcting the errors made by the actual model and perfecting predicting the training data. Adaboost and gradient boosting are popular boosting algorithms. Adaboost is used for boosting performance machine learning algorithms like decision trees [80] and works well for binary classification. Gradient boosting has 3 parts viz., a weak learner for predicting, a loss function which is to be optimised. and an additive model. The additive model adds weak learners thereby, reducing the loss function. It is a greedy algorithm which benefits from regularisation method by penalizing parts of algorithm to enhance the performance by reducing overfitting.

Multi-layer perceptron, a supervised algorithm trains on a function using the training dataset. It is formed of input, output layers and one or one plus hidden layers. It has the ability to learn non-linear models and real time models.

5.5 Matrics for validation and testing

For validating and testing the performance, practicality and robustness of the models put forward in these study different standard parameters were chosen namely, visualization techniques, validation by a group of experts and users, effect of different model parameters like epochs, dropout, embedding size etc. The most popular ones being accuracy, F1-value, precision, recall, MSE (mean square error) as well as MAE (mean absolute error).

The parameters required to define the metrices above are true positive (positive entity correctly labelled), true negative (negative entity correctly labelled), false positive (negative entity incorrectly labelled), false negative (positive entity incorrectly labelled).

Of the metrices mentioned, the accuracy is the number of correct predictions upon the all the predictions whether correct or incorrect. It is given by the first (1) equation; signifying t as true, f as false, p as positive and n as negative.

$$accuracy = \frac{(tp + tn)}{tp + tn + fp + fn} \tag{1}$$

To understand how much of true positive values were correct from all the predicted positive values we make use of precision. It is given by (2) and shows the ability of the model's classifier to not incorrectly label a sample as positive if it is not positive.

$$precision = \frac{tp}{tp + fp} \tag{2}$$

Recall is used to know the actual number of true positives which were recognized correct and is given by Eq. (3).

$$recall = \frac{tp}{tp + fn} \tag{3}$$

F1-score which is calculated using precision and recall, as it is the weighted average of the two. It could be used if we need to strike balance among Precision and Recall or if the class distribution is uneven.

$$f1 = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

(MAE) Mean absolute error, that is, mean of absolute error measures the absolute average distance between the real and the predicted data it could at times fail to penalize large errors while predicting. The MSE gives a measure of the real and the predicted data in terms of squared average distance between them.

Apart from these measures undertaken to test the proposed model's effectiveness most papers also made comparative study of their model / dictionary or lexicon with the standard models.

6 Summarizing the study

The summarization of the study of the papers performed is put forward in the Tables 1 and 2. One may take a look at the tables to understand the key details which were the focus of performing this collective study. Consideration of the driving cause for performing a study (Area of focus / Application) along with this the train / test dataset using which the models had been trained, tested or a corpus / dictionary was built is considered. The built of the model as per our understanding is briefed as well. The parameters which were

Table 2 List of accuracies of papers which worked on SemEval-2014 Restaurant dataset

Paper	Features	Classifier	Best accuracy
[7]	Syntactic Structures	RNN	81.38%
[8]	Aspect Features	CH-HEAT	**92.40%**
[22]	Skip-gram	BERT	91.60%
[41]	Surface, Lexicon & Parsing	CNN & Bi-LSTM	80.965%
[46]	TD-BERT	BERT	82.57%
[49]	Bag of word	DRCNN	88.92%
[52]	IRAN	LSTM	81.96%
[53]	GMemN2N	ADeCNN	84.03%

Bold term signifies the maximum Accuracy and this will help readers to easily identify the maximum value

used to judge the performance, robustness and effectiveness of the model by the authors is also jot down. Table 1: summary of different aspects of literature studied, contains 5 columns namely, paper number (paper no.), year of publication (year), area of focus / application, dataset, parameters used for testing (in order of appearance).In paper [1], the authors have developed a Dynamic Bayesian network using a Gaussian Process. Along with this a sequential Monte Carlo sampler is also designed for performing Bayesian Interface.

Study [5] proposes a Naïve Bayes classifier in order to predict posterior probability for a Danmaku sample. This would categorize the sample as per he prior probability distribution and the group with maximum probability is put out as forecasted sentiment.

Utilizing a lexicon for feature enhancement, CNN and GRU for extracting major features and context features along with the attention mechanism sentiment classification is performed in [2].

From the context of the study performed in [8], the best possible accuracy of 92.4% is obtained by the features of embedding vectors which are fed to attention layers which are connected to one another in a hierarchical format. The one of the two does feature extraction using attention layer and makes utility of features to get aspect data for the sentiment using second attention layer. The second one extracts aspect features with the help of attention layer and makes use of features to get aspect data using the earlier attention layer.

This work [38] aims for sarcasm detection by proposing a multi-dimension question answer network using attention with Bi-LSTM. MQA introduces abundant semantic information and builds conversation context information using Bi-LSTM and attention mechanism. Model is based on 2 components.

The study [24] works on building a Soft attention based Bi-LSTM and CNN hybrid model to detect sarcasm in short tweets. The model built consists of the layers mentioned the embedding and input, attention with Bi-LSTM followed by convolution layer. After which is an activation ReLU layer with max pooling. The last layer is the representation one.

Paper [39] puts attention on enhancing the performance of sarcasm detection by identifying key parts of the sentence. A Bi-LSTM with multiple-head self-attention is worked on for this purpose. The NN consisting of 2 layers is built viz., multiheaded attention sentence level (understands the aspect) and word encoder (summarizes contextual information). While the word encoder represents each word by summarizing its contextual information from both directions in the comment, the sentence level multiheaded attention layer looks into different part of comment to understand the aspects of the text semantics.

In [32], there are 2 parallel paths with each path having CNN, Bi-LSTM modules combined with attention mechanism. While these are used to extract feature aspect in both local- and global domain, their fusion helps detect relevant semantic information. They are then united with semantic information received from another path through mutual operation. Thus, generalization of the model is enhanced.

For text classification, in order to filter out semantic features [45] proposes a cross layer attention model base stacked residual RNN. Various kinds of linguistic features are used to construct a stacked network. The cross layer attention mechanism refines the filtering process. It makes use of the features from high-level to supervise the low-level features. This can be utilized to select more sematic features which could be used for text classification.

The authors of [52] focuses on aspect level sentiment analysis by constructing an attention network based on interactive rules by designing a grammar-based encoder to simulate grammar functions. It does this by standardizing the output of adjacent positions and constructing an interaction attention network and learns attention information using the target and context. By implementing this model the best accuracy obtained is 81.96%.

Study [3] focuses on calculating reputation of an entity viz. movie, product, hotel, restaurant, service etc. For this review helpfulness and time scores are computed. Then BERT model is tuned to predict probability of sentiment orientation. For each review, a numerical score, reputation value is calculated by making use of the proposed formulas. The paper also proposes a way to visualize reputation.

In [22], using 3 domain related corpora authors have finely tuned the BERT model for context embedding and opinion mining at aspect level. BERT as an embedding extracts local and global context features concurrently. And for merging local and global features concurrently Multi-head attention mechanisms and convolution operations is developed. This helps in understanding the polarity of targeted aspect. To fuse the semantic text representation Multi-head Self-Attention is used whereas to model the interaction of aspect term convolutional layers are used. This model gives out the best accuracy of nearly 91.6%.

A statement could refer many targets and each of which could have its own context. A BERT model that is target dependent is discussed in [46] with three variations is proposed in this paper. The output is positioned at the target terms..

A framework for domain adaptive text classification for leveraging domain knowledge during training stage an adversarial-powered multi-domain learning-based method is put forward in [17]. It reduces the necessity for training data for each source domain. It ensures a split in between private and shared features by using orthogonality constraints and adversarial training. This ensures that shared realm does not get adulterated domain private features.

The domain specific approach of [43] focuses on at first building a manually labelled (positive, negative, neutral) football sentiment dataset which is then used to automatically build a lexicon which is football specific. A classifier then identifies the sentiment of the conversation. The algorithm performance and features for sentiment recognition on the created dataset is tested.

The study [6] at first extracts the features then tweet embeddings are created using a feature ensemble model and feature vectors. CNN model then classifies the tweets into 5 sets negative, positive, neutral, strong positive and strong negative sets.

The [11] study gets sentiment embeddings of words in and not in lexicon by utilizing a pretraining word classifier. Accuracy of the word representation increases by combining sentiment embedding and word embedding. A proposed method based on LSTM and attention suggests finding attention vector without target to improve LSTM's ability in capturing global sentiment and semantic composition.

To perform Aspect based analysis in [7] a hierarchical model built up of recurrent NN cascaded to the recursive NN. The model consists of constituency parser and dependency parser responsible for converting text to phrases, building trees of phrases and connecting words if they have binary relation respectively. Using these parsers reviews are cut into pieces including sentiment information relevant to aspect terms. This model was able to achieve the best accuracy of 79.5% and 81.38% for single and multiple aspect scenario on the restaurant dataset of SemEval- 2014 task 4 as shown in the Table 2.

For the improvement of Accuracy, overcoming shortcomings like overfitting, vanishing gradient in sentiment analysis; the study [15] puts forward a multichannel paradigm is exploited for extracting multi-grain features. Every channel extracts context information and local features using Bi-GRU and CNN respectively. VIB (Variational Information Bottleneck) for compressing extracted sentiment features. To make this work sections of model are formed for extracting context information and local features, compressing features and calculating sentiment polarity.

In [25], Backdoor attack is injected into the model this makes the model misbehave through inputs. The adversary with knowledge only regarding the few training data but not the structure or training algorithm is considered. On the backdoor injection, the model begins misclassifying text samples with trigger sentences into the target group specified by adversary. This doesn't affect the model's performance as such.

A text mining technique, NN based iterative opinion mining is used in [26]. It analyzes posts from users of social media. It makes use of automatic incremental procedure based on Feedforward NN. By exploiting a small amount of classification rules from some hashtags which are biased to specific factions, new classification rules are generated iteratively. Generated rules help determine polarization of people.

The paper [28] looks into classification of app reviews. At first textual and non-textual information is extracted, then the textual information is processed to create a digital vector. History of the reviewer is also looked into. Then, it trains CNN classifier to classify multi-class reviews.

Paper [33] deals with incorporating prior information in RNN models for classification and recognition of named entity task. Firstly, prior knowledge could be put up by extracting a group of rules consisting of syntactic dependence, knowledge and social graph and. Then a way for embedding these rules into RNN is suggested.

A sales factor neural network model that is based on multilayer back propagation perceptron is proposed [34]. It makes use of the weight matrix operation to get the relative strength which denotes the strength between i^{th} input and j^{th} output.

In [41], four different kinds of memory networks are curated for different tasks. One uses CNN for capturing local information while other uses Bi-LSTM to capture sequence information in documents. The other two are also based upon CNN, Bi-LSTM. While one of them builds hierarchical NN to combine local and sequence information together, the other captures local information and sequence information through different modules. The resultant representation is combined to classify sentiments considering the aspects and results give the best possible accuracy of 80.965% as shown in the Table 2.

Multinomial LR, SVM, DT, RF are used in the study performed in [4] for categorization purposes.

Similarly, in [9] multi-strategy sentiment analysis method based on SVM and NB is used for working with Chinese characters.

The goal of [42] is identification of hate text on twitter for which a twitter corpus is developed and distinctive features are created for ease of identification. This is used to evaluate ML algorithms based on ensemble, multi-meta-learning models of SVM, LR, RF, GB algorithms and on hyper-parameter optimization.

Four main parts of the model from [36]: Collecting data, Storing data, Analyzing data and data visualizing data. Needfull [36], consists of a module dedicated for data collection that crawls twitter to collect raw data. This data is then labelled before the data is put into the database. On entering the query from User interface, one gets all the necessary data of the database from the index system of where the data is stored. Analysis of these tweets is then put up using the data visualization module.

Operation of the framework in [44] carried out by 2 blocks namely, spam detection which operates in real time and the update module which operates in batches. The spam detection module has 4 lightweight detectors to label tweets coming from blacklisted URLs, tweets which are near duplicates of prelabelled tweets, tweets posted by trusted users without spammy words. Tweets apart from these categories are labelled by the multi-classifier-based detector. Tweets labelled in a time window are used to update the next time window's data in batch mode as required by the detection module.

Model of [49] is built comprises of RNN, CNN and has layers like input, convolutional, forward and backward LSTM, pooling, convolutional layer and fully connected and output layer. While CNN learns classification features RNN abstracts and memorizes sequence of semantic relations. Pre-trained word embeddings are used to reduce the representation dimension of text.

In [51], Sentiment information is integrated into TF-IDF to generate weighted word vectors. These are fed to Bi-LSTM to obtain context information and get a better representation of comment vector. Feedforward NN helps find out sentiment tendency.

The authors of [53] put forward a model (called ADeCNN) for analyzing sentiments using aspect. It uses an attention mechanism into deformable CNN, Bi-LSTM. The model built generates attention weights as per targets. This would also enhance correlation between target and parts in the statement. The best accuracy obtained on SemEval-2014 Restaurant dataset by this model is 84.03%.

In [10], a word refinement model is proposed that is applicable on pretrained word vectors which makes use of real valued sentiment scores from the lexicon. The model works to get the word vector closer to similar words both semantically and as per the intensity score. This improves the word embeddings and sentiment embeddings and helps in ranking each word's similar nearest neighbor. These ranked nearest ones direct the movement and distance of refinement procedure to better the word vector.

Using the add-on sentiment dictionary and the designed score rules in [13] the authors derive the sentiment. The NB classifier gives the field of text which comprise of the polysemic sentiment word by using the sentiment dictionary proposed.

Using many sentiment dictionaries and sets of semantic rules, [14] suggests a Chinese microblog sentiment classification method is proposed. The semantic rule set consists of both inter-sentence and sentence pattern rule analysis. The flow of the method begins by preprocessing the data and using sentiment dictionary and rule analysis sets to calculate the value of sentiment which is then used for sentiment classification.

A manual corpus with annotated scheme is built in [21]. It identifies sentiment in metaphorical sentiment using 3 methods. The study conducted by the authors found the third one works better in tasks that relate to metaphor and the first one works better for classification.

For study of tweets and evolution of the behavior of users as per topics they discuss over the time, a word embedding based approach is proposed in [29]. Along with this, temporal transitions of topics are tracked using 5 topic evolution events.

To begin with the model [37] functioning at first sample users are selected by random sampling. Their tweets form the corpus. Each tweet is defined with an emotional score. The emotion dictionary has an emotion score for each word and the sum is the emotion score of the tweet. After this, an emotion score average is calculated for each user to get the emotional trend of the user. Then, 2 groups viz., positive and negative are defined and sorted in descending order as per the average emotion scores. The top one forth are positive grouped and bottom one forth are negative grouped. Applying statistical test on these groups would help investigate impact of emotional behavior on user's relations over twitter.

The study [40] Comprised of 2 processes of which one works by cleaning data in regular dictionaries and the other is meant for extracting semantic relation. It is based on the interpretation relation between lexicon and corpus along with the rule-based method.

The study [12] proposes a method for contextual analysis is built, for which, a Hierarchical Knowledge Tree (HKT) using unlabeled dataset is built. This helps understand the subject or knowledge of data and build a relation between words and sources in HKT and this paves path for comparison between training and actual dataset and thereby, figure out

amount of success of any supervised ML algorithm. It also looks into detecting and understanding positive changes between both the positive words and the negative words based on the modes in the trees.

In [16], twitter hierarchical latent Dirichlet allocation, LDA is proposed. It aims at getting the hierarchy of tweet topics which are further employed for online analytical processing on tweets. The technique begins by data collection, preprocessing. After which text modelling is done to identify relation between tweeters and tweets. Hierarchical text modelling is also done to extract twitter data's topics and build the hierarchical topic dimension based on probability distribution of parts of topics. Tweeters are analyzed from different dimensions using online analytical processing.

Model proposed by authors in [19] focuses on Arabic sentiment analysis by employing and evaluating a feature ensemble model. It comprises of surface features which are manually extracted features and the deep features are generic word embeddings and word embeddings which are sentiment specific. By understanding these word embeddings from tweets, its performance with generic embeddings and manually extracted features is compared.

By reaping the benefits of MOOC-related aspects the framework proposed in [20] propagates the weakly supervised signal in order to identify aspect categories form the unlabeled students' reviews. The analysis of architecture is made up of 4 major components viz., information fed by the user, weak label propagation, aspect section learning and polarity.

The proposed framework [23] gives attention on implementation of GAWA with modified fitness function. For this purpose, Wrapper approaches (WA) are utilized for selecting features and Genetic Algorithm (GA) is utilized for reduction of those features.

The focus is to study [27] the movement of stock market. For this multiple information sources are used. Common estimated true label among hinge losses of different data is shared for modelling latent consistencies among different data sources. For the data extracted from different sources, methods involving Restricted Boltzmann Machine as well as sentence2vec to get dense vectors are used for training purposes.

In order to improve sentiment classification of product reviews cross modal hypergraph model is proposed in [30]. It captures textual information and sentiment information simultaneously. To merge multimodal feature, the hypergraph model are extended. By coupling LDA based topic model with proposed model, ambiguity of words expressing opposite polarity in different contexts is mitigated and run time is lowered.

The task of rumor detection from Chinese micro-blogs [31] uses sentiment dictionary, dynamic time series algorithm and GRU model with 2 layers. The sentiment dictionary captures the fine-grained aspects. The fuzzy clustering method, Dynamic Time Series method retains data distribution of events over the time.

Paper [35] works on: Sentiment dimensions aide to map financial news articles from source to target into the feature space. Different transfer principles are developed considering the highly correlated stocks' price time series, source and target stocks in same sector, highest prediction performance in validation dataset. Voting mechanism is designed to select most proper source stock from candidate stocks which are generated from different principles just mentioned. The final prediction is based on the prediction model which are trained on selected stocks.

In [47], Multivariate regression models (The least squares support vector regression) and time series models predict monthly total vehicle sales. 3 types of data are used to forecast monthly total vehicles viz., sentiment scores of tweets, stock market values, and hybrid data (mix of both). Multivariate regression data is dealt with least squares support vector

regression models, naïve model, the exponential smoothing model, the autoregressive integrated moving average model, the seasonal autoregressive integrated moving average model, and backpropagation neural networks and LSSVR with time series models are all the time series models.

The paper [48] works on ASP and Deep2S. Answer Set Programming (ASP) is used for implementing aspect extraction that is based on semantic structure by making use of Abstract Meaning Representation (AMR) graph. AMR represents semantic relations which prevail between the aspects and opinion words is worked on to get semantic data in reviews. Deep2S works for extracting aspects. It integrates AMR with syntactic structure. It obtains dependency relations using syntactical rules and semantic information using semantic rules.

In [50], Members of group are categorized as per their interest and the response they give to various posts. This filtering mechanism would cluster members based on their aspect-based characteristics. The posts are also clustered on various aspects. This using a 2-level clustering would help in recommending users the posts which would likely be of interest to them.

7 Challenges and gaps

There is a huge buzz among researchers and companies for building a model that fits well for data of different kinds and gives the perfect sentiment analysis. However, to build one such model, choosing the dataset and fixating parameters and hyperparameters to suit it to the model and get the correct response is a huge task, especially because there is lack of annotated datasets which would be compile, fit and train models for domains of all varieties. This is majorly as the samples from source realm which are to be collected and labelled are found to be difficult and costly [96]. There could be ways to deal with this, one of them could be to construct sharing networks at output for distinctive tasks. This could help the model to learn information of other tasks within the network. Other one, could be adding manually annotated data to the unlabeled data so that it could be used as well or by combining domain knowledge like the rules and the dictionaries. Also, by taking advantage of other models by using transfer learning the issue of poorly labeled data could be dealt with. Transfer learning has garnered interest as it utilizes prior experience and analogical learning [96]. But, it could still have to deal with issues like differences between the train and test dataset, ensuring stable outcome over source and target areas which could be affected due to various data distributions between the cultural factors, data sources, domains, noise in data and linguistic variations as well as differing contexts. Recent studies [35, 64, 95–98], however, show good results for sentiment analysis through transfer learning.

Majority of the studies pertaining to the classification of sentiments deals with binary, ternary or 7–8 emotion classes. This does not seem fully practical as there are circumstances when the amount of tilt towards a particular sentiment might be needed to give the real sentiment. Like for the case, “The ice creams from that shop are really tasty but do not seem worth the extravagant fanfare !!”- here the ice cream is expressed to be good enough but so good to be hyped about. Another takeaway from this sentence could be the issue of feature deviation, meaning, the phase—“extravagant fanfare” shifts the focus from the ice cream to the fanfare. The target data holds more importance, as a single statement could

refer to many targets, of which each has its own context [46]. It is hard to find out the sentiment for a target if no accurate aspect information is given [94].

India is a hugely populated country with approximately 1.3 billion people. Of lately, there has been a surge in the people using internet through their mobile devices. This call for a huge potential in exploring regional languages from India, namely, Hindi- the most popular language across majority of the country, Marathi, Bengali, Gujrati, Punjabi and another regional language. These languages aren't as well-resourced as the English, especially for sentiment analysis. Recently, the studies aiming creation of publicly available dataset and sentiment analysis in Hindi, widely used language across India, are been worked upon [99–102]. Further development in this realm would help understanding Indian consumers and Indian Market deeply.

It so happens that different social networks cater different audiences. This has to be taken into account. To deal with this, a few steps could be taken like collecting data from different networks. Like collecting data from Twitter, Facebook, Instagram etc. and then using all of this data to get more clarity rather than relying on dataset coming from one stream. Another step could be to look into the amount of likes, shares, saves, comments a particular post has received. As it is a good indicative of the amount of support a topic has garnered from a set of people. The study [103] explores the connection between liking behavior of a user, the user's relation with the poster and sentiment of the post. A social network connects people across the globe thus one wrong message or misinformation has the potential to trigger million minds. Also, the preaching offensive and hate over social networks has increased despite the term of service which disallows such speech [42]. Keeping a study of such activities by looking into the number of likes, shares, comments etc. and trying to form a pattern of such post creators and post spreaders could have a tremendous impact in curbing misinformation from ruining and corrupting people's mindset.

Recently, the introduction of bots has eased automation and changed social media. The use of bots isn't limited for the good, but also different malicious tasks. The content over social media may be full by the content coming from bots such as Influence Bots, Spam Bots [104]. They are put to work for increasing a kind of sentiment post receive, or inflating the number of followers or friends on social media etc. Thereby making it really crucial to identify the network of such malicious practices and desert it to maintain the sanity of social media. More study on this regard is expected to bloom in the upcoming years.

Online communication, which are relating to politics are many a times sarcastic and ironic [105]. Such texts are tough to recognize even by humans, however, of lately, many studies have been performed for detection of sarcasm. But study of identification of fake content which could influence, mislead masses causing ruckus needs to be given more attention to. Tracking how people's opinions / sentiments are shaped on social media and the users influencing and spreading misinformation must be tracked down and their patterns must be studied.

Online Social Networks have immense amount of network-centric data which can be utilized for getting the know-how of social and behavioral characters of individuals or groups. This has led the operators of social network to publish the data for use to researchers and advertisers etc. The data published on social media data is prone to a wide variety of reidentification and disclosure attacks [106], hence mechanisms to protect the privacy must be taken care of. In recent times, several cases of exploiting and breaching data safety practices have come forward. Utmost importance must be given to maintain the privacy and trust of social media users and other online platform users.

There are two limits to this review: one is related to the number of papers included, and the other is related to the quality of papers included. The filtering of papers is explained

through PRISMA [121]. PRISMA is a minimum set of elements for systematic reviews and meta-analyses that is based on evidence. PRISMA in Fig. 6 gives the justification of how the database is selected from various journals.

8 Conclusion and future work

Sentiment Analysis has been study from different perspectives in this study and it has been gaining immense popularity among NLP researchers. This systematic literature review is aimed to aid the researchers and highlight the current challenges and commonly used techniques. Papers from year 2016–2022 were considered while preparing this review from various journals. Its selection of database can be seen from Fig. 6. Table 1 shows the algorithms, features, classification method and accuracy achieved by various algorithms for sentiment analysis. This study examines the various techniques in order to discover which one is best suited for a given application. We classify and summarize common sentiment analysis approaches to make techniques like machine learning, deep learning, lexicon-based analysis, and hybrid analysis more approachable. In order to keep up with current trending research, a summary of the pros and limitations of sentiment analysis is provided. This review shows that the cutting edge schemes are directed more towards the advancement and development of deep learning schemes as compared to others. Widely used datasets and their sources are discussed briefly. Apart from this, challenges and gaps are put forward which could potentially form new research directions.

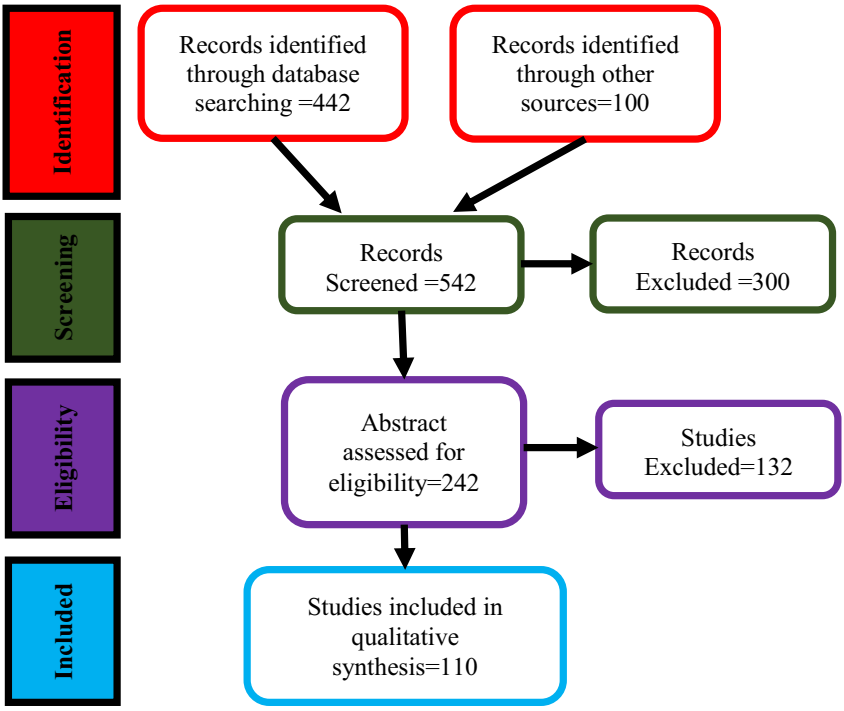


Fig. 6 Database selection through PRISMA

Deep learning neural networks are widely used recently and have achieved good success but this heavily rely on size of dataset and doesn't naturally with ease detect indirect sentiments portrayed textually. Though the existing models have showcased their utility in classification, some new text classification models just brush up some of the check parameters of most of the classification tasks. This doesn't give a clear picture whether or not the model comes to know of the statements at a real semantic level. The real test of the model takes place when classification of no or few same labelled class data is fed into it. Dependence on labelled data comes in handy while building and training the new models so that the performance of the models is not hampered when fed with less to no labelled data. Apart from this feeding the model with external knowledgebase might be of some help in improving performance and robustness of the model. There seems, yet a long journey for these approaches to aid the development in sentiment analysis. Our further studies would aim at addressing these challenges in order to help in creating a robust mechanism and also improve the performance in the way sentiment analysis is approached.

Acknowledgements We thank Mr Ruchirangad Kapoor , doing B.Tech from VIT-AP in Computer Sciences for helping in the script formatting and for making few very useful suggestions.

Funding No funding has been received from any one for carrying out this research work.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study'.

Declarations

Conflict of interests/Competing interests There is no Conflict of Interests for this research work.

References

1. Liang H, Ganeshbabu U, Thorne T (2020) A dynamic bayesian network approach for analysing topic-sentiment evolution. IEEE Access 8:54164–54174. https://doi.org/10.1109/ACCESS.2020.2979012
2. 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
3. Benlahbib A, Nfaoui EH (2020) Aggregating customer review attributes for online reputation generation. IEEE Access 8:96550–96564. https://doi.org/10.1109/ACCESS.2020.2996805
4. Saad SE, Yang J (2019) Twitter sentiment analysis based on ordinal regression. IEEE Access 7:163677–163685. https://doi.org/10.1109/ACCESS.2019.2952127
5. Li Z, Li R, Jin G (2020) Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary. IEEE Access 8:75073–75084. https://doi.org/10.1109/ACCESS.2020.2986582
6. Phan HT, Tran VC, Nguyen NT, Hwang D (2020) Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model. IEEE Access 8:14630–14641. https://doi.org/10.1109/ACCESS.2019.2963702
7. Aydin CR, Güngör T (2020) Combination of recursive and recurrent neural networks for aspect-based sentiment analysis using inter-aspect relations. IEEE Access 8:77820–77832. https://doi.org/10.1109/ACCESS.2020.2990306
8. Gao Y, Liu J, Li P, Zhou D (2019) CE-HEAT: an aspect-level sentiment classification approach with collaborative extraction hierarchical attention network. IEEE Access 7:168548–168556. https://doi.org/10.1109/ACCESS.2019.2954590
9. Fang Y, Tan H, Zhang J (2018) Multi-strategy sentiment analysis of consumer reviews based on semantic fuzziness. IEEE Access 6:20625–20631. https://doi.org/10.1109/ACCESS.2018.2820025
10. Yu L, Wang J, Lai KR, Zhang X (2018) Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Trans Audio Speech Lang Process 26(3):671–681. https://doi.org/10.1109/TASLP.2017.2788182

11. Fu X, Yang J, Li J, Fang M, Wang H (2018) Lexicon-enhanced LSTM with attention for general sentiment analysis. IEEE Access 6:71884–71891. https://doi.org/10.1109/ACCESS.2018.2878425
12. Abdul Aziz A, Starkey A (2020) Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches. IEEE Access 8:17722–17733. https://doi.org/10.1109/ACCESS.2019.2958702
13. Xu G, Yu Z, Yao H, Li F, Meng Y, Wu X (2019) Chinese text sentiment analysis based on extended sentiment dictionary. IEEE Access 7:43749–43762. https://doi.org/10.1109/ACCESS.2019.2907772
14. Wu J, Lu K, Su S, Wang S (2019) Chinese micro-blog sentiment analysis based on multiple sentiment dictionaries and semantic rule sets. IEEE Access 7:183924–183939. https://doi.org/10.1109/ACCESS.2019.2960655
15. Gu T, Xu G, Luo J (2020) Sentiment analysis via deep multichannel neural networks with variational information bottleneck. IEEE Access 8:121014–121021. https://doi.org/10.1109/ACCESS.2020.3006569
16. Yu D, Xu D, Wang D, Ni Z (2019) Hierarchical topic modeling of twitter data for online analytical processing. IEEE Access 7:12373–12385. https://doi.org/10.1109/ACCESS.2019.2891902
17. Ding X, Shi Q, Cai B, Liu T, Zhao Y, Ye Q (2019) Learning multi-domain adversarial neural networks for text classification. IEEE Access 7:40323–40332. https://doi.org/10.1109/ACCESS.2019.2904858
18. Yin F, Wang Y, Liu J, Lin L (2020) The construction of sentiment lexicon based on context-dependent part-of-speech chunks for semantic disambiguation. IEEE Access 8:63359–63367. https://doi.org/10.1109/ACCESS.2020.2984284
19. Al-Twairesh N, Al-Negheimish H (2019) Surface and deep features Ensemble for Sentiment Analysis of arabic tweets. IEEE Access 7:84122–84131. https://doi.org/10.1109/ACCESS.2019.2924314
20. Kastrati Z, Imran AS, Kurti A (2020) Weakly supervised framework for aspect-based sentiment analysis on students' reviews of MOOCs. IEEE Access 8:106799–106810. https://doi.org/10.1109/ACCESS.2020.3000739
21. Zhang D, Lin H, Zheng P, Yang L, Zhang S (2018) The identification of the emotionality of metaphorical expressions based on a manually annotated chinese corpus. IEEE Access 6:71241–71248. https://doi.org/10.1109/ACCESS.2018.2881270
22. Abas AR, El-Henawy I, Mohamed H, Abdellatif A (2020) The identification of the emotionality of metaphorical expressions based on a manually annotated chinese corpus. IEEE Access 8:128845–128855. https://doi.org/10.1109/ACCESS.2020.3008824
23. Rasool A, Tao R, Kamyab M, Hayat S (2020) GAWA–A feature selection method for hybrid sentiment classification. IEEE Access 8:191850–191861. https://doi.org/10.1109/ACCESS.2020.3030642
24. Son LH, Kumar A, Sangwan SR, Arora A, Nayyar A, Abdel-Basset M (2019) Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. IEEE Access 7:23319–23328. https://doi.org/10.1109/ACCESS.2019.2899260
25. Dai J, Chen C, Li Y (2019) A backdoor attack against LSTM-based text classification systems. IEEE Access 7:138872–138878. https://doi.org/10.1109/ACCESS.2019.2941376
26. Belcastro L, Cantini R, Marozzo F, Talia D, Trunfio P (2020) Learning political polarization on social media using neural networks. IEEE Access 8:47177–47187. https://doi.org/10.1109/ACCESS.2020.2978950
27. Zhang X, Qu S, Huang J, Fang B, Yu P (2018) Stock market prediction via multi-source multiple instance learning. IEEE Access 6:50720–50728. https://doi.org/10.1109/ACCESS.2018.2869735
28 De Oliveira Góes AS, De Oliveira RCL (2020) A process for human resource performance evaluation using computational intelligence: an approach using a combination of rule-based classifiers and supervised learning algorithms. IEEE Access 8:39403–39419. https://doi.org/10.1109/ACCESS.2020.2975485
29. Abulaish M, Fazil M (2018) Modeling topic evolution in twitter: an embedding-based approach. IEEE Access 6:64847–64857. https://doi.org/10.1109/ACCESS.2018.2878494
30. Chen Z, Lu F, Yuan X, Zhong F (2018) TCMHG: topic-based cross-modal hypergraph learning for online service recommendations. IEEE Access 6:24856–24865. https://doi.org/10.1109/ACCESS.2017.2782668
31. Wang Z, Guo Y, Wang J, Li Z, Tang M (2019) Rumor events detection from chinese microblogs via sentiments enhancement. IEEE Access 7:103000–103018. https://doi.org/10.1109/ACCESS.2019.2928044
32. Park D, Kim S, Lee J, Choo J, Diakopoulos N, Elmqvist N (2018) ConceptVector: text visual analytics via interactive lexicon building using word embedding. IEEE Trans Visual Comput Graphics 24(1):361–370. https://doi.org/10.1109/TVCG.2017.2744478

33. Chen B et al (2019) Embedding logic rules into recurrent neural networks. IEEE Access 7:14938–14946. https://doi.org/10.1109/ACCESS.2019.2892140
34. Zhao Z, Wang J, Sun H, Liu Y, Fan Z, Xuan F (2020) What factors influence online product sales? online reviews, review system curation, online promotional marketing and seller guarantees analysis. IEEE Access 8:3920–3931. https://doi.org/10.1109/ACCESS.2019.2963047
35. Li X, Xie H, Lau RYK, Wong T, Wang F (2018) Stock prediction via sentimental transfer learning. IEEE Access 6:73110–73118. https://doi.org/10.1109/ACCESS.2018.2881689
36. Long Z, Alharthi R, Saddik AE (2020) NeedFull – a tweet analysis platform to study human needs during the COVID-19 pandemic in New York state. IEEE Access 8:136046–136055. https://doi.org/10.1109/ACCESS.2020.3011123
37. Tago K, Jin Q (2018) Influence analysis of emotional behaviors and user relationships based on twitter data. Tsinghua Sci Technol 23(1):104–113. https://doi.org/10.26599/TST.2018.9010012
38. Diao Y et al (2020) A multi-dimension question answering network for sarcasm detection. IEEE Access 8:135152–135161. https://doi.org/10.1109/ACCESS.2020.2967095
39. Kumar A, Narapareddy VT, Aditya Srikanth V, Malapati A, Neti LBM (2020) Sarcasm detection using multi-head attention based bidirectional LSTM. IEEE Access 8:6388–6397. https://doi.org/10.1109/ACCESS.2019.2963630
40 Liang L, Tian F (2020) Using normal dictionaries to extract multiple semantic relationships. J Eng 2020(13):595–600. https://doi.org/10.1049/joe.2019.1212
41. Han H, Liu J, Liu G (2018) Attention-based memory network for text sentiment classification. IEEE Access 6:68302–68310. https://doi.org/10.1109/ACCESS.2018.2879481
42. Oriola O, Kotzé E (2020) Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. IEEE Access 8:21496–21509. https://doi.org/10.1109/ACCESS.2020.2968173
43. Aloufi S, Saddik AE (2018) Sentiment identification in football-specific tweets. IEEE Access 6:78609–78621. https://doi.org/10.1109/ACCESS.2018.2885117
44. Sedhai S, Sun A (2018) Semi-supervised spam detection in twitter stream. IEEE Trans Comput Soc Syst 5(1):169–175. https://doi.org/10.1109/TCSS.2017.2773581
45. Lan Y, Hao Y, Xia K, Qian B, Li C (2020) Stacked residual recurrent neural networks with cross-layer attention for text classification. IEEE Access 8:70401–70410. https://doi.org/10.1109/ACCESS.2020.2987101
46. Gao Z, Feng A, Song X, Wu X (2019) Target-dependent sentiment classification with BERT. IEEE Access 7:154290–154299. https://doi.org/10.1109/ACCESS.2019.2946594
47. Pai P, Liu C (2018) Predicting vehicle sales by sentiment analysis of twitter data and stock market values. IEEE Access 6:57655–57662. https://doi.org/10.1109/ACCESS.2018.2873730
48. Li X et al (2020) Deep2s: improving aspect extraction in opinion mining with deep semantic representation. IEEE Access 8:104026–104038. https://doi.org/10.1109/ACCESS.2020.2999673
49. Zhang X, Huang S, Zhao J, Du X, He F (2019) Exploring deep recurrent convolution neural networks for subjectivity classification. IEEE Access 7:347–357. https://doi.org/10.1109/ACCESS.2018.2885362
50. Athira U, Thampi SM (2018) Linguistic feature based filtering mechanism for recommending posts in a social networking group. IEEE Access 6:4470–4484. https://doi.org/10.1109/ACCESS.2017.2789200
51. Xu G, Meng Y, Qiu X, Yu Z, Wu X (2019) Sentiment analysis of comment texts based on BiLSTM. IEEE Access 7:51522–51532. https://doi.org/10.1109/ACCESS.2019.2909919
52. Lu Q, Zhu Z, Zhang D, Wu W, Guo Q (2020) Interactive rule attention network for aspect-level sentiment analysis. IEEE Access 8:52505–52516. https://doi.org/10.1109/ACCESS.2020.2981139
53. Zhou J, Jin S, Huang X (2020) ADeCNN: an improved model for aspect-level sentiment analysis based on deformable CNN and attention. IEEE Access 8:132970–132979. https://doi.org/10.1109/ACCESS.2020.3010802
54. Ghaddar B, Naoum-Sawaya J (2018) High dimensional data classification and feature selection using support vector machines. Eur J Oper Res 265(3):993–1004. https://doi.org/10.1016/j.ejor.2017.08.040
55. Ding X, Cai B, Liu T, Qiankun S (2018) Domain adaptation via tree kernel based maximum mean discrepancy for user consumption intention identification. Proc IJCAI 4026–4032
56. Blitzer J, Blitzer M, Pereira F (2007) Biographies, Bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. Proc ACL 7:440–447
57. Maas AL, Daly RE, Pham PT (2011) Learning word vectors for sentiment analysis. Proc ACL-HLT 142–150
58. Pang B, Lee L (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. Proc ACL 115–124

59. Al-Twairesh N, Al-Khalifa H, Al-Salman A, Al-Ohali Y (2017) 'AraSenTi-tweet: a corpus for arabic sentiment analysis of saudi tweets.' Proc Comput Sci 117:63–72
60. Nabil M, Aly M, Atiya A (2015) ASTD: Arabic sentiment tweets dataset. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), pp 2515–2519
61 Beladev M, Rokach L, Shapira B (2016) Recommender systems for product bundling. Knowl-Based Syst 111:193–206
62. Oliveira N, Cortez P, Areal N (2017) The impact of microblogging data for stock market prediction: using twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Syst Appl 73:125–144
63. Chung W, Mustaine E, Zeng D (2017) Criminal intelligence surveillance and monitoring on social media: Cases of cyber-trafficking. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing. pp 191–193. https://doi.org/10.1109/ISI.2017.8004908
64. Castillo SRM, Chen Z (2016) Using transfer learning to identify privacy leaks in tweets. In: 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), Pittsburgh, PA. pp 506–513, doi: https://doi.org/10.1109/CIC.2016.078
65. Ren R, Wu DD, Liu T (2019) Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Syst J 13(1):760–770. https://doi.org/10.1109/JSYST.2018.2794462
66. Vo A, Nguyen Q, Ock C (2018) Opinion–Aspect relations in cognizing customer feelings via reviews. IEEE Access 6:5415–5426. https://doi.org/10.1109/ACCESS.2018.2797224
67 Wang L, Niu J, Yu S (2020) SentiDiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis. IEEE Trans Knowl Data Eng 32(10):2026–2039. https://doi.org/10.1109/TKDE.2019.2913641
68. Feizollah A, Ainin S, Anuar NB, Abdullah NAB, Hazim M (2019) Halal products on twitter: data extraction and sentiment analysis using stack of deep learning algorithms. IEEE Access 7:83354–83362. https://doi.org/10.1109/ACCESS.2019.2923275
69. Wang Z, Xiao X, Rajasekaran S (2020) Novel and efficient randomized algorithms for feature selection. Big Data Mining Anal 3(3):208–224. https://doi.org/10.26599/BDMA.2020.9020005
70. Garg I, Panda P, Roy K (2020) A low effort approach to structured CNN design using PCA. IEEE Access 8:1347–1360. https://doi.org/10.1109/ACCESS.2019.2961960
71. Xu K, Wang F, Wang H, Yang B (2020) Detecting fake news over online social media via domain reputations and content understanding. Tsinghua Sci Technol 25(1):20–27. https://doi.org/10.26599/TST.2018.9010139
72. Hu T, She B, Duan L, Yue H, Clunis J (2020) A systematic spatial and temporal sentiment analysis on geo-tweets. IEEE Access 8:8658–8667. https://doi.org/10.1109/ACCESS.2019.2961100
73. Zhou J, Lu Y, Dai H, Wang H, Xiao H (2019) Sentiment analysis of chinese microblog based on stacked bidirectional LSTM. IEEE Access 7:38856–38866. https://doi.org/10.1109/ACCESS.2019.2905048
74. Mehmood F, Ghani MU, Ibrahim MA, Shahzadi R, Mahmood W, Asim MN (2020) A precisely xtreme-Multi Channel hybrid approach for Roman urdu sentiment analysis. IEEE Access 8:192740–192759. https://doi.org/10.1109/ACCESS.2020.3030885
75. Jiahao Y, Jiang X, Wang S, Jiang K, Yu X (2020) SVM-BiLSTM: a fault detection method for the gas station IoT system based on deep learning. IEEE Access 8:203712–203723. https://doi.org/10.1109/ACCESS.2020.3034939
76. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP). pp 1–15
77. Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22
78. Smetanin S (2020) The applications of sentiment analysis for russian language texts: current challenges and future perspectives. IEEE Access 8:110693–110719. https://doi.org/10.1109/ACCESS.2020.3002215
79. Schütze H, Manning CD, Raghavan P (2008) Introduction to information retrieval, vol 39. Cambridge Univ. Press, Cambridge
80. Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106
81. Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M (2002) Logistic regression. Springer, Atlanta
82. Gunn SR (1998) Support vector machines for classification and regression. ISIS Tech Rep 14(1):5–16
83. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NIPS), vol. 2. Curran Associates, Red Hook, pp 3111–3119

84. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP). Association for Computational Linguistics, Doha, pp 1532–1543
85. Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of tricks for efficient text classification. In: Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, vol. 2. Association for Computational Linguistics, Valencia, pp 427–431
86. Graves A, Jaitly N, Mohamed A-R (2013) Hybrid speech recognition with deep bidirectional LSTM. In: Proc. IEEE Workshop Autom. Speech Recognit. Understand, pp 273–278
87. Rao G, Zhang Y, Zhang L, Cong Q, Feng Z (2020) MGL-CNN: a hierarchical posts representations model for identifying depressed individuals in online forums. IEEE Access 8:32395–32403. https://doi.org/10.1109/ACCESS.2020.2973737
88. Duppada V, Jain R, Hiray S (2018) SeerNet at SemEval-2018 task 1: Domain adaptation for affect in tweets. In: Proc. 12th Int. Workshop Semantic Eval. Association for Computational Linguistics, New Orleans, pp 18–23
89. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pretraining of deep bidirectional transformers for language understanding. In: Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., vol. 1. Association for Computational Linguistics, Minneapolis, pp 4171–4186
90. Cheng Y, Yao L, Xiang G, Zhang G, Tang T, Zhong L (2020) Text sentiment orientation analysis based on Multi-Channel CNN and bidirectional GRU with attention mechanism. IEEE Access 8:134964–134975. https://doi.org/10.1109/ACCESS.2020.3005823
91. Cai R et al (2020) Sentiment analysis about investors and consumers in energy market based on BERT-BiLSTM. IEEE Access 8:171408–171415. https://doi.org/10.1109/ACCESS.2020.3024750
92. Tiantian C, Hongwei L, Shuisheng Z (2009) Large scale classification with local diversity AdaBoost SVM algorithm. J Syst Eng Electron 20(6):1344–1350
93. Zhao Q, Yang J, Wang S, Li M, Zhang W (2019) High-value user identification based on topic weight. IEEE Access 7:175917–175928. https://doi.org/10.1109/ACCESS.2019.2953243
94. Jiang L, Yu M, Zhou M, Liu X, Zhao T (2011) Target-dependent Twitter sentiment classification. In: Proc. 49th Annu. Meeting Assoc. Comput. Linguistics Hum. Lang. Technol., vol. 1. pp 151–160
95 Sun C, Qiu X, Xu Y, Huang X (2019) How to fine-tune BERT for text classification? In: Sun M, Huang X, Ji H, Liu Z, Liu Y (eds) Chinese computational linguistics. Springer, Cham, pp 194–206
96. Zhong X, Guo S, Shan H, Gao L, Xue D, Zhao N (2018) Feature-based transfer learning based on distribution similarity. IEEE Access 6:35551–35557. https://doi.org/10.1109/ACCESS.2018.2843773
97. Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, Abrego GH, Yuan S, Tar C, Sung Y-H, Strope B, Kurzweil R. Multilingual universal sentence encoder for semantic retrieval, arXiv:1907.04307. Available: http://arxiv.org/abs/1907.04307. Accessed 7 June 2022
98. Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. In: Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, vol. 1, pp 328–339
99. Malviya S, Mishra R, Tiwary US (2016) Structural analysis of Hindi phonetics and a method for extraction of phonetically rich sentences from a very large Hindi text corpus. In: 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), Bali. pp 188-193. https://doi.org/10.1109/ICSDA.2016.7919009
100. Magdum D, Dubey MS, Patil T, Shah R, Belhe S, Kulkarni M (2015) Methodology for designing and creating Hindi speech corpus. In: 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur. pp 336–339. https://doi.org/10.1109/SPACES.2015.7058279
101. Jain G, Lobiyal DK (2019) Word sense disambiguation of Hindi text using fuzzified semantic relations and fuzzy Hindi WordNet. In: 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India. pp 494–497. https://doi.org/10.1109/CONFLUENCE.2019.8776967.
102. Kumar P, Raunak, Yadav S (2015) Semantic generator for disambiguation of Hindi sentences. In: 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata. pp 450–454. https://doi.org/10.1109/ICRCICN.2015.7434281
103. Levordashka A, Utz S, Ambros R (2016) What's in a like? Motivations for pressing the like button. In: Proc. 10th Int. AAAI Conf. Web Social Media. pp 623–626
104. Velayutham T, Tiwari PK (2017) Bot identification: Helping analysts for right data in twitter. In: 2017 3rd International Conference on Advances in Computing,Communication & Automation (ICACCA) (Fall), Dehradun. pp 1–5. https://doi.org/10.1109/ICACCAF.2017.8344722

105 Liu B (2012) Sentiment analysis and opinion mining. Morgan & Claypool, San Rafael
106. Abawajy JH, Ninggal MIH, Herawan T (2016) Privacy preserving social network data publication. In: IEEE Communications Surveys & Tutorials, vol. 18, no. 3. pp 1974–1997, third quarter. https://doi.org/10.1109/COMST.2016.2533668
107. Acheampong FA, Wenyu C, Nunoo-Mensah H (2020) Text-based emotion detection: advances, challenges, and opportunities. Eng Rep 2(7):e12189
108. Short J, Williams E, Christie B (1976) The social psychology of telecommunications. Wiley, Toronto
109. Zad S, Heidari M, James Jr, H, Uzuner O (2021) Emotion detection of textual data: An interdisciplinary survey. In: 2021 IEEE World AI IoT Congress (AIIoT). IEEE, pp 0255–0261
110. Kumar A, Garg G (2020) Systematic literature review on context-based sentiment analysis in social multimedia. Multimed Tools Appl 79(21):15349–15380
111. Soleymani M, Garcia D, Jou B, Schuller B, Chang SF, Pantic M (2017) A survey of multimodal sentiment analysis. Image Vis Comput 65:3–14
112. Zhang L, Wang S, Liu B (2018) Deep learning for sentiment analysis: a survey. Wiley Interdiscip Rev Data Min Knowl Discov 8(4):e1253
113. Do HH, Prasad PWC, Maag A, Alsadoon A (2019) Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl 118:272–299
114. Ligthart A, Catal C, Tekinerdogan B (2021) Systematic reviews in sentiment analysis: a tertiary study. Artif Intell Rev 54(7):4997–5053
115. Peng H, Cambria E, Hussain A (2017) A review of sentiment analysis research in chinese language. Cogn Comput 9(4):423–435
116. Mäntylä MV, Graziotin D, Kuutila M (2018) The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Comput Sci Rev 27:16–32
117. Boudad N, Faizi R, Thami ROH, Chiheb R (2018) Sentiment analysis in arabic: a review of the literature. Ain Shams Eng J 9(4):2479–2490
118. Chan JYL, Bea KT, Leow SMH, Phoong SW, Cheng WK (2022) State of the art: a review of sentiment analysis based on sequential transfer learning. Artif Intell Rev 1–32
119. Verma S (2022) Sentiment analysis of public services for smart society: Literature review and future research directions. Gov Inf Q 101708
120. Mukherjee P, Badr Y, Doppalapudi S, Srinivasan SM, Sangwan RS, Sharma R (2021) Effect of negation in sentences on sentiment analysis and polarity detection. Proc Comput Sci 185:370–379
121. Moher D, Altman DG, Liberati A, Tetzlaff J (2011) PRISMA statement. Epidemiology 22(1):128
122. Chen K, Franko K, Sang R (2021) Structured model pruning of convolutional networks on tensor processing units. arXiv preprint arXiv:2107.04191
123. Singh N, Kapoor R (2023) Multi-modal expression detection (MED): a cutting-edge review of current trends, challenges and solutions. Eng Appl Artif Intell 125:106661

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.

e-Prime - Advances in Electrical Engineering, Electronics and Energy 6 (2023) 100374

Contents lists available at ScienceDirect

e-Prime - Advances in Electrical Engineering, Electronics and Energy

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

Reliability analysis of a CNT-TF-FinFET for hostile temperature

Praween Kumar Srivastava, Atul Kumar *, Ajay Kumar *

Department of Electronics & Communication Engineering, Jaypee Institute of Information Technology, Noida 201309, India

ARTICLE INFO

Keywords:
Hostile temperature
CNT-TF-FinFET
TF-FinFET
Analog
Linearity

ABSTRACT

In the concerned research work, an extensive study has been carried out for a newly developed carbon nanotube truncated fin (CNT-TF) FinFET from an analog as well as a linearity point of view. All the simulated work has been gleaned for n-type FET on Silvaco software for TCAD. Firstly, CNT-TF-FinFET has been combatted against TF-FinFET which gave out the enhanced results as 750 times and 2.13 times for switching ratio and V_{th} respectively. After perceiving, CNT-TF-FinFET was studied at different temperatures, out of which the device at 200 K achieved zenith with 67.7 %- and 1000-times enhancement in efficiency and IMD3 respectively. The device also shows the linear behavior at higher temperatures in terms of g_{m3} and VIP3. All these accomplishments made the CNT-TF-FinFET a much better competent to be operated at extreme geographic poles of earth or in outer space with extremely low temperatures.

Introduction

Today's scenario, of going beyond the conventional imagination, triggered ever more intense research with the best possible outcomes at the nanoscale. The inclusion of beyond 10 nm scale sets novel outcomes that have never been grasped so far. Since the day, the chip technology of 10 nm was introduced by commercial giants, new phenomenal nano-distortion explanations came into existence thus urging the developers to incline more towards the senile principles of material science and quantum technology. The overall prognostication has become so essential that commercial distortion could be possible if not analyzed sedulously [1–3].

The chip performance has not only to ensure the circuitry or segregated device aspects but also the device's corrigibility to surrounding atmospheric deterioration as well. So, paving the same characteristic view proved the robustness of the device over hostile temperature conditions. By keeping in mind its suitability for harsh conditions, the CNT-TF-FinFET has been designed. The device has been tested across a wide variety of temperatures in order to predict the best outcomes before its actual implementation. By considering the TF-FinFET as the base, the inclusion of the channel of carbon nanotube as well as the spacer at source-channel, drain-channel, and substrate-channel have been proved productive with the application of high k-dielectric as gate oxide [4–6]. All the alterations cause performance enhancement from the perspective of analog and linearity. The kind of improvements are highly desirable from mixed signal circuitry in CMOS technology and that's what has been investigated throughout this work [7–9].

The aforementioned amalgamation has not only provided better performance at active state but also gives better durability over catholic hostile surroundings. The appearance of investigated device has been refined in a way to take advantage of conventional MOSFET, FinFET, and even GAA (Gate All Around) FET and the same could be understood by having a discrete look at its appearance and performance [11–13].

Simulation description

Device structure

By considering Fig. 1, it is shown that the carbon nanotube (CNT) is placed in the channel, one of the most essential constraints in the design of this device. There are some significant aspects of the CNT channel like its diameter, chirality, and bandgap as well. The main building block behind the conductive property of the carbon nanotube is its chirality no. Apart from this, the diameter of the tube decides the level of conduction and power handling capacity. Carbon nanotubes with smaller diameters could bear the smaller, conduction and can possess high power handling capacity, than that with larger diameters. A shallow trench isolation (STI) is used as box isolation technique in the proposed device. Below is the overall representation of the proposed device from the front and lateral side respectively.

Various dimensional parameters were accepted into consideration to give a semiconducting characteristic to our CNT. The parameters have

* Corresponding authors.
E-mail addresses: atul.nit304@gmail.com (A. Kumar), ajaykumar@dtu.ac.in (A. Kumar).

https://doi.org/10.1016/j.prime.2023.100374
Received 23 October 2023; Received in revised form 18 November 2023; Accepted 21 November 2023
Available online 23 November 2023
2772-6711/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

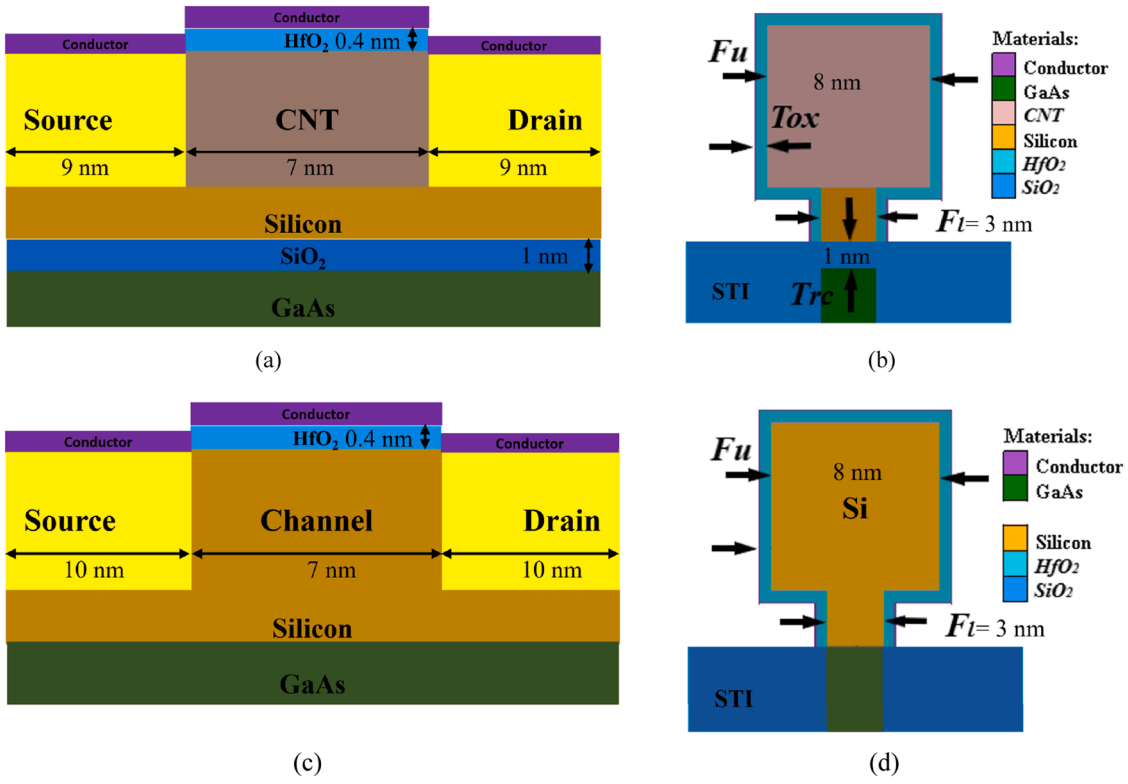


Fig. 1. (a) 2D side view of CNT-TF-FinFET. (b) 2D cross-sectional view of CNT-TF-FinFET at 7 nm node. (c) 2D side view of TF-FinFET. (d) 2D cross-sectional view of TF-FinFET at 7 nm node.

Table 1
Dimensional parameters.

Serial No.	Dimensional Parameters		
	Dimensions	TF-FinFET (nm)	CNT-TF-FinFET (nm)
1	Thickness of gate oxide (T_{OX})	0.4	0.4
2	Channel Length (L_c)	7	7
3	Lower fin width (F_l)	3	3
4	Upper fin width (F_u)	8	8
5	Fin height	8	8
6	Spacer width (T_{rc})	–	1
7	CNT Chirality	–	(6,2)
8	CNT diameter	–	1.41
9	Souce/Drain Length	10	9

been gleaned implicitly from [10] and subsequently verified as well. Fig. 1(a) shows the 2D side view of CNT-TF-FinFET and Fig. 1(b) shows the 2D cross-sectional view of CNT-TF-FinFET. The proposed device is compared with the TF-FinFET which 2D side view shown in Fig. 1(c) while Fig. 1(d) reflects the 2D cross-sectional view of TF-FinFET at 7 nm node. The remaining dimensional elements are mentioned below in Table 1. The doping concentration of all the simulated n-type devices kept as same which are 1×10^{20} (cm^{-3}) for the source, 1×10^{19} (cm^{-3}) for the drain, and 1.0×10^{15} (cm^{-3}) for the substrate. There is no usage of high-profile Gaussian style of doping. This is also useful if the device will further be considered as the idea of fabrication.

Methodology of simulation

Several physical simulation models were employed to anticipate & analyze the electrical behavior of the concerned devices. The Bohm Quantum Potential (BQP) model evaluates the effects of quantum confinement on electron and hole concentrations. Hence, the position-dependent quantum potential was investigated using the BQP model,

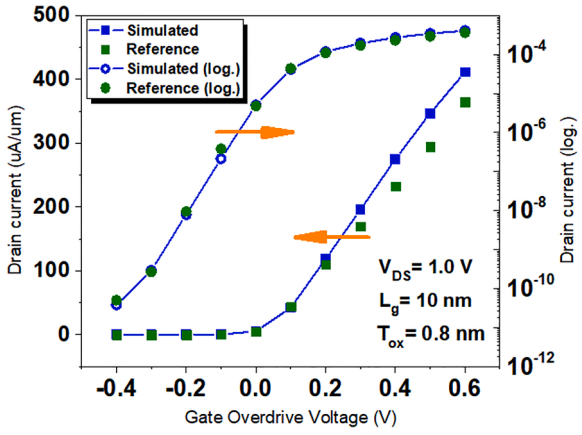


Fig. 2. Results calibrations of the experimental work [18] against our simulated work at a gate length of 10 nm.

which ultimately determines the charge carrier potential energy at the concerned region. The f-N tunneling model has also introduced the distortion caused by the undesirable entry of the electrons towards the substrate and gate oxide of the simulated device. *Fermi statistical model* is used, as shrinkage of bandgap occurs when the impurity concentration is particularly high. Doping-induced shifting of the conduction band minimum and valence band maximum may greatly influence the device's behavior. To address the distortion of the simulated device properties by the combination of the electrons and ionized atoms created by the development of a sufficiently large electric field at the drain-channel interface, the doping relying model of SRH (Shockley-Read-Hall) recombination was employed. Crowell and Sze's impact ionization model is also considered to describe the correlation between the generation of the electric field and the ionization rate [16,17].

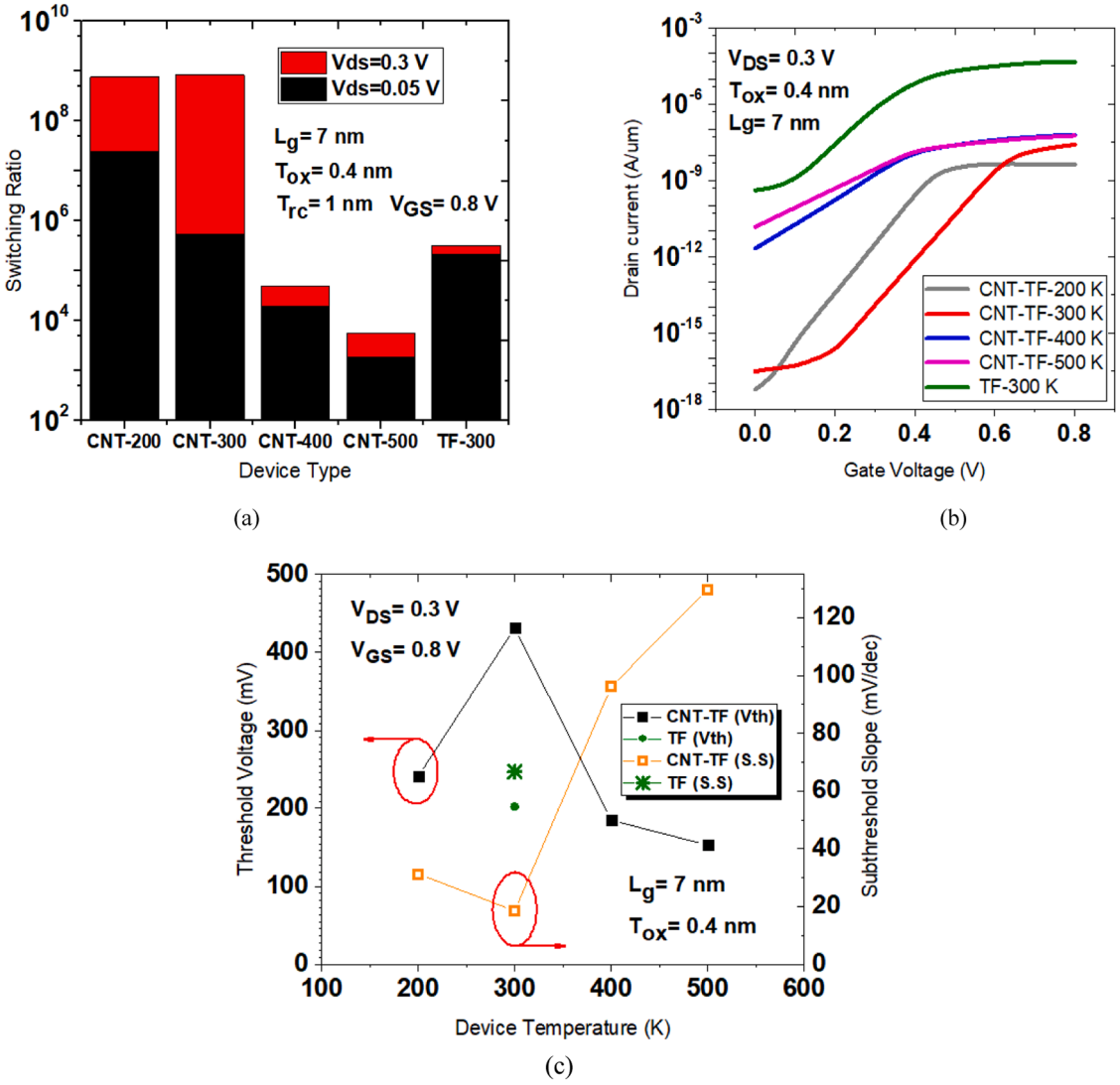


Fig. 3. (a) Switching ratio w.r.t different device temperatures, (b) Variation of drain current w.r.t gate voltage at different constant temperatures, and (c) threshold voltage and subthreshold slope (SS) w.r.t different device temperatures.

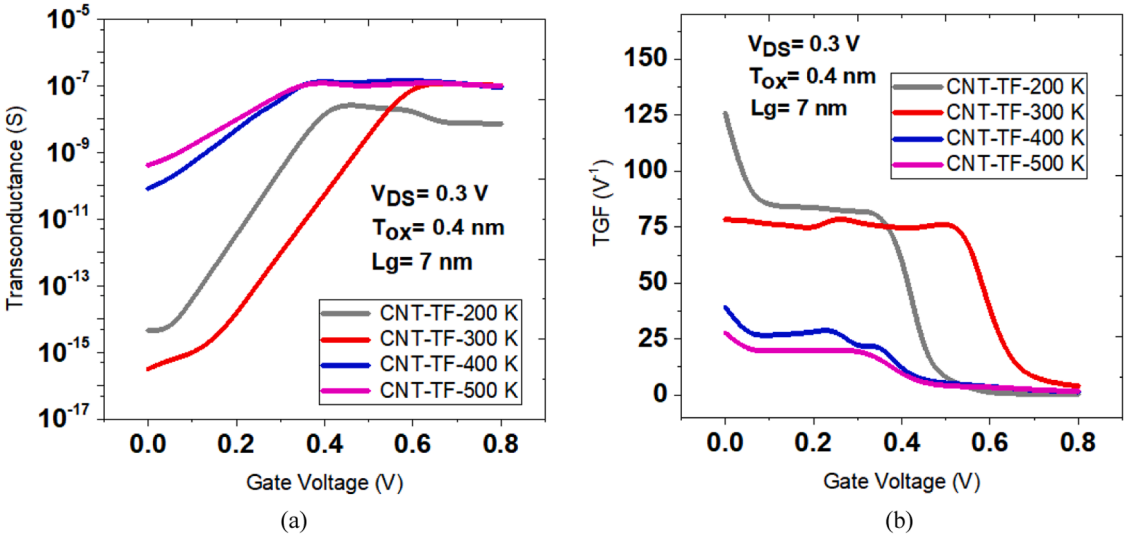


Fig. 4. Dependency of (a) Transconductance (g_m), (b) Transconductance generating factor (TGF) on input applied voltage at different constant temperatures.

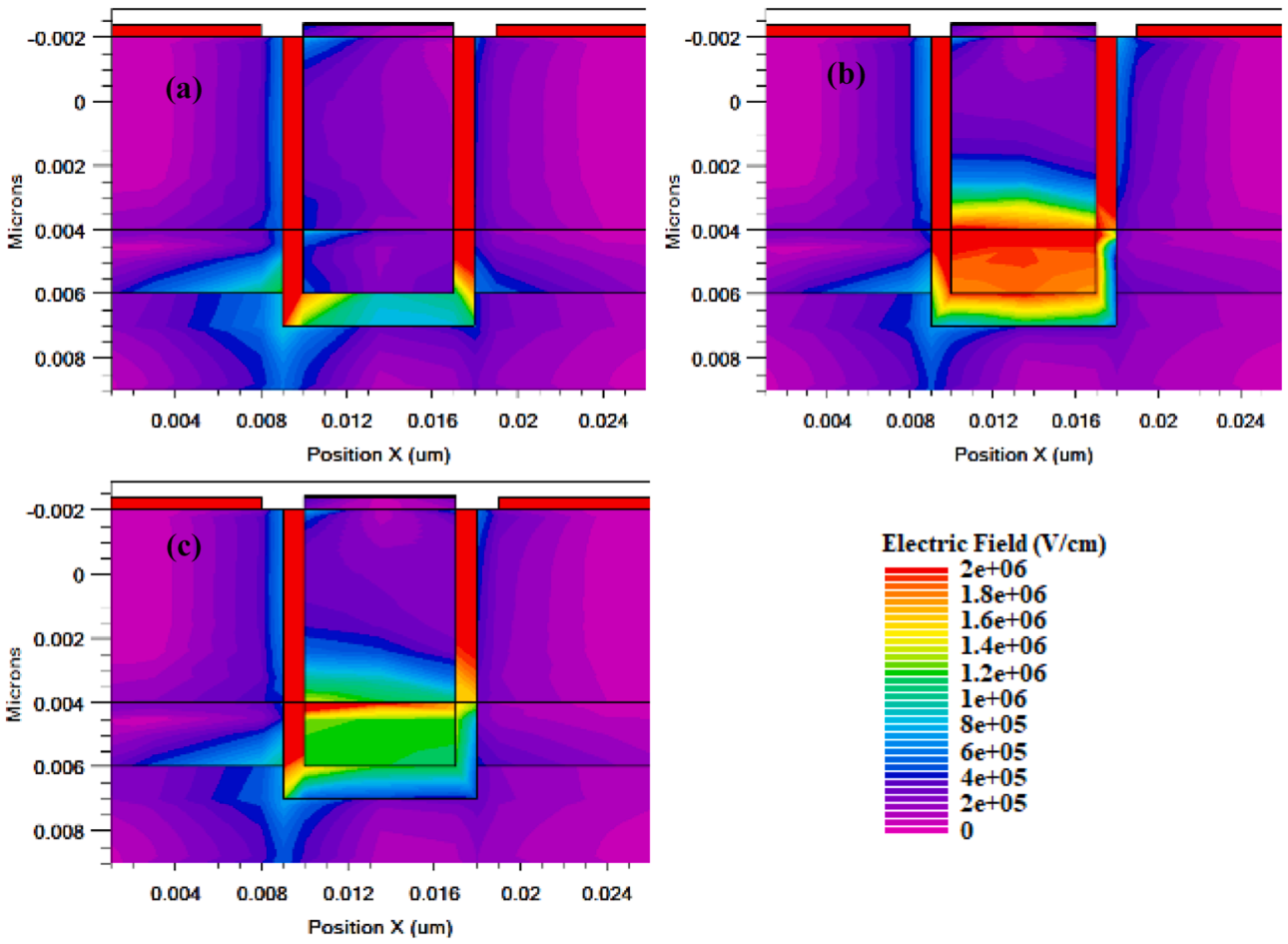


Fig. 5. Variation of electric field spread at V_{GS} = 0.8 V and V_{DS} = 0.3 V for (a) 200 K, (b) for 300 K, and (c) for 500 K.

Calibration & Validation

As mentioned below, Fig. 2 depicts the calibration of the simulated results of this work with the results of previously published experimental research [18], which is annotated by respective keywords for our simulated results (i.e., "Simulated") and for the results of previously published research (i.e., "Reference") which is being referenced by us. The minor deviations can be grasped from this figure. By taking the results of previously published research as a reference [18], approximately 4.65 % and 1 % deviations have been deduced, if seen from the point of view of I_{ON} and I_{OFF} respectively.

Discussion on results

This section is solely dedicated to assimilating the robustness of CNT-TF-FinFET for n-type device configuration. The results are well arranged in chronological order to be understood seamlessly.

The current characteristics of CNT-TF-FinFET and the unaltered TF-FinFET are shown, in Fig. 3 (a & b), to analyze the compatibility of both at the same temperature conditions. The I_{ON} achieved by TF-FinFET at 300 K was found as 1780 times higher than CNT-TF-FinFET (45.766 uA and 0.0255 uA for TF-FinFET and CNT-TF-FinFET respectively) at the same power delivery instances. The more domineering factor that can be seen easily, is the amount of OFF-state current occurring in the sub-threshold region. As I_{OFF} for CNT-TF-FinFET was found in the range of 10^{-18} to 10^{-17} uA and 10^{-10} to 10^{-9} uA (shown in Fig. 3(b)) for TF-FinFET thus making the CNT-TF-FinFET a better choice with an improvement of switching ratio around 10^{7} times as shown in Fig. 3(a). Further, for higher temperature, switching ratio reduces in the CNT-TF-FinFET as compared to TF-FinFET owing to reduction of conductivity at higher temperatures and it has been observed from Fig. 3(a).

Now, by accounting for Fig. 3(c), one could depict the variation of V_{th} and SS at different designated temperatures. The device is at 300 K, being most productive over its counterparts. Higher and lower the values of V_{th} and SS respectively, higher would be the durability of the device at ON-state as well as OFF-state. CNT-TF-FinFET was found better in these terms, even its 200 K variant performed well better (i.e., with 241.08 mV and 31.25 mV/dec as V_{th} and SS respectively) than the TF-FinFET at 300 K (i.e., with 202.29 mV and 66.70 mV/dec as V_{th} and SS respectively) as shown. So, CNT-TF-FinFET could be assumed as the better choice between 200 K to 300 K over the TF-FinFET under the same circumstances.

In evaluating the device performance, the input orientation has a big influence which is determined by analyzing the transconductance [19, 20], in Fig. 4(a), of the device under test. The higher the transconductance, the lower would be the device's inclination towards the output side. CNT-TF-FinFET at 300 K has the highest orientation whereas, its counterpart at 200 K has the lowest with around 70 times decrement, making the device at 200 K with the least significance. However, it remains in the good lead if seen from the lower applied voltage (i.e., V_{GS} = 0 to 0.55 V) point of view.

As far as the efficiency is concerned, one can analyze it from Fig. 4 (b), the plot values exemplify the robustness of CNT-TF-FinFET at lower voltages in comparison to its counterpart at 200 K. The highest value attained by the device was 125.84 V^{-1} and 78.38 V^{-1} at 200 k and 300 K respectively thus enabling the device to draw the 63.8 % enhancement and to work efficiently at 200 K.

As shown in Fig. 5, the electric field spread in CNT-TF-FinFET

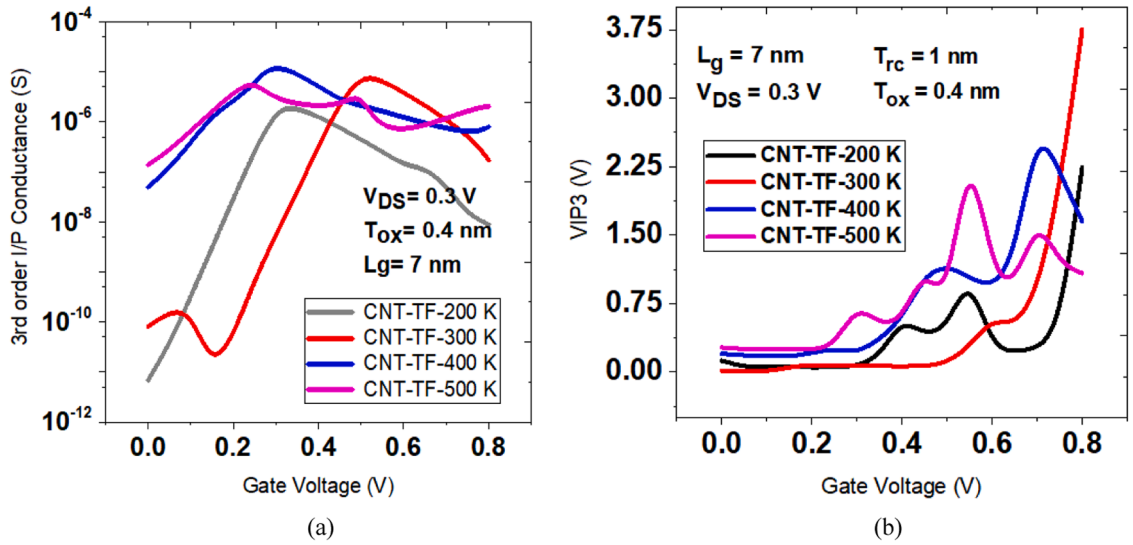


Fig. 6. (a) Dependence of g_{m3} and (b) Dependence of VIP3 on input applied signal at different particular temperatures.

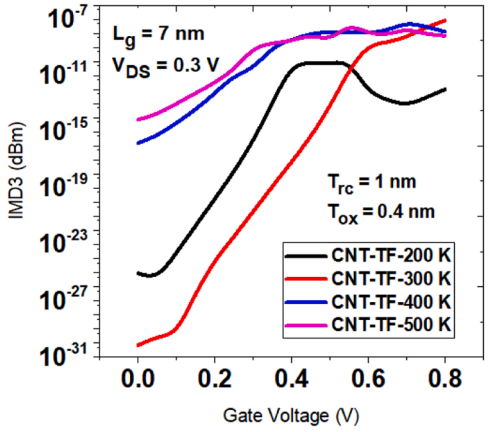


Fig. 7. Variation of 3rd order inter-modulation distortion with respect to gate voltage.

explains and confirms the reason behind its improvement of current characteristics at lower temperatures. The observation has been gleaned out b/w the constant steps of increment as shown. The variation of electric field spread is shown in Fig. 5(a) for 200 K, 5(b) for 300 K, and 5 (c) for 500 K. It comprises the least electric field inside the lower channel region and hence has maximum control over the channel region. In the case of 500 K, the electric spread was found maximum which ultimately indicates the undesirable conduction of charges in the channel as well as in the substrate region, thus, making the device more fragile than ever before [14,15].

By considering the linearity [21,22] aspects of CNT-TF-FinFET, an overall scenario has been predilected before its actual implementation into the CMOS technology. Being the fundamental 3rd-order characteristic determining factor, the g_{m3} has been extracted from Eq. (1) [23, 24], in Fig. 6(a), and proved its feasibility at a much lower temperature (i.e., 200 K). Since it stood minimum at 200 K for higher operable input voltage, it asserts the least variation of distortion parameters ahead. The impact of 3rd-order conductance can be seen in Fig. 6(b) in terms of VIP3 (third order voltage intercept point) which is calculated from Eq. (2). As clearly illustrated, the device with temperature 300 K bagged the highest slot and followed by the device with 200 K. Device with 300 K has been found with 3.75 V and 2.23 V at $V_{GS} = 0.8$ V, for the device at 200 K, which ultimately made them above all in terms of voltage interrupt point where 1st and 3rd harmonics of the device get equal.

$$g_{m3} = \frac{\partial^3 I_d}{\partial V^3_{gs}} \tag{1}$$

$$VIP3 = \sqrt{24 \times g_{m1}/g_{m1}} \tag{2}$$

where $g_{m1} = \frac{\partial I_d}{\partial V_{gs}}$ and$g_{m3} = \frac{\partial^3 I_d}{\partial V^3_{gs}}$

Coming up for the analysis is influenced by the 3rd order input conductance at higher temperature, the third order intermodulation distortion (IMD3) calculated from Eq. (3) [25,26] where $R_S = 50\ \Omega$ for analog and RF applications. Fig. 7 reflects that the 1st and 3rd harmonics of the device are equal. From the amplification point of view, this distortion plays an important role to predict the device behavior and sot is for CNT-TF-FinFET for 200 K. The overall variation of IMD3 for 200 K has meant more productive (i.e., 10^3 times improved) than that from 300 K, hence developing better stability. This is due to the appropriate variation of 1st and 3rd-order harmonics particularly for devices at 200 K.

$$IMD3 = \left(\frac{9}{2} \times (VIP3)^3 \times g_{m3}\right)^2 \times R_S \tag{3}$$

Conclusion

A comprehensive study has been performed to test the temperature adaptability of a newly developed CNT-TF-FinFET over its base i.e., TF-FinFET. CNT-TF-FinFET bagged the more sophisticated results in contrast to TF-FinFET like 10^7 times improvement and around 20 % enhancement in switching ratio and threshold voltage respectively, making it a better choice. After knowing the dominance of CNT-TF-FinFET over its counterpart, a regress analysis has been performed to predict its compatibility over a wide range variety of temperatures. By doing intense observation, CNT-TF-FinFET was found robust at 200 K and 300 K majorly. The observation has been drawn with results like 67.7 %, 1000 times and about 10 times improvement in efficiency, IMD3, and g_{m3} correspondingly. All these achievements made CNT-TF-FinFET an ideal choice to be used in operations in outer space and at extreme geographical poles of our planet.

CRediT authorship contribution statement

Praween Kumar Srivastava: Methodology, Software, Validation,

Writing – review & editing, Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. **Atul Kumar:** . **Ajay Kumar:** .

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.

Acknowledgment

The authors are dedicated to thanking the authorities of the ECE department of Jaypee Institute of Information Technology, Noida for all the support provided while conducting this research work.

References

[1] S. Saini, M.P. Kashyap, R. Chaujar, Recessed channel carbon nanotube truncated fin finfet for high performance ULSI applications, in: 2021 Devices for Integrated Circuit (DevIC), 2021, pp. 124–128, https://doi.org/10.1109/DevIC50843.2021.9455799.

[2] A. Kumar, N. Gupta, R. Chaujar, Analysis of novel transparent gate recessed channel (TGRC) MOSFET for improved analog behaviour, Microsyst. Technol. 22 (11) (2016) 2665–2671, https://doi.org/10.1007/s00542-015-2554-z.

[3] W.T. Huang, Y. Li, Electrical characteristic fluctuation of 16-nm-gate trapezoidal bulk FinFET devices with fixed top-fin width induced by random discrete dopants, Nanoscale Res. Lett. 10 (1) (2015) 1–8, https://doi.org/10.1186/s11671-015-0739-0.

[4] A.B. Sachid, M.C. Chen, C. Hu, FinFET with high-κ spacers for improved drive current, IEEE Electron Device Lett 37 (7) (2016) 835–838, https://doi.org/10.1109/LED.2016.2572664.

[5] A. Kranti, G.A. Armstrong, Design and optimization of FinFETs for ultra-low-voltage analog applications, IEEE Trans. Electron Devices 54 (12) (2007) 3308–3316, https://doi.org/10.1109/TED.2007.908596.

[6] M. Singh, S.R. Mishra, Soumya Mohanty, Guru Mishra, Performance analysis of SOI MOSFET with rectangular recessed channel, Adv. Natural Sci.: Nanosci. Nanotechnol. 7 (2016), 015010, https://doi.org/10.1088/2043-6262/7/1/015010.

[7] S.K. Mohapatra, K.P. Pradhan, Prasanna Sahu, Manikant Kumar, The performance measure of GS-DG MOSFET: an impact of metal gate work function, Adv. Natural Sci.: Nanosci. Nanotechnol. 5 (6) (2014), https://doi.org/10.1088/2043-6262/5/2/025002.

[8] A. Kumar, Effect of trench depth and gate length shrinking assessment on the analog and linearity performance of TGRC-MOSFET, Superlattices Microstruct. 109 (2017) 626–640.

[9] T. Kang, Y. Yang, F. Chien, Poly Si nanowire thin film transistors with vacuum gap design, IEEE Trans. Electron Devices 61 (6) (2014) 2113–2118, https://doi.org/10.1109/TED.2014.2318706.

[10] Bhoop Singh, Prasad B, D. Kumar, DFT based estimation of CNT parameters and simulation-study of GAA CNTFET for nano scale applications, Mater. Res. Express. Vol. 7 (2020), https://doi.org/10.1088/2053-1591/ab6924.

[11] E.R.Kim Jung, M. Choi, Learning nanotechnology through crossbar-based architecture and carbon nanotube (CNT) FETs, in: 2011 IEEE International Conference on Microelectronic Systems Education, San Diego, CA, 2011, pp. 60–63, https://doi.org/10.1109/MSE.2011.5937093.

[12] G. Navin, T.N. Basavaraj, Factors affecting performance of CNT FET as a switch in memory cell, in: International Conference on Advanced Nanomaterials & Emerging Engineering Technologies, Chennai, 2013, pp. 655–657, https://doi.org/10.1109/ICANMEET.2013.6609376.

[13] D. Tan, M. Cometto, P. Roux-Levy, P. Coquet, T.B. Kang, D. Baillargeat, Carbon nanotubes based RF packaging solutions, in: 2018 IEEE 18th International Conference on Nanotechnology (IEEE-NANO), 2018, pp. 1–4, https://doi.org/10.1109/NANO.2018.8626279.

[14] A. Loumi, A. Ghammaz, H. Belahrach, Crosstalk and delay analysis in multi-walled carbon nanotube and single-walled carbon nanotube interconnects, in: 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 2019, pp. 1–4, https://doi.org/10.1109/WITS.2019.8723814.

[15] H. Yin, Y. Zhang, P. Lu, Y. Zhang, J. Liu, Analysis of heat dissipation characteristics of three-dimensional graphene-carbon nanotube composite structures, in: 2020 21st International Conference on Electronic Packaging Technology (ICEPT), 2020, pp. 1–5, https://doi.org/10.1109/ICEPT50128.2020.9202463.

[16] ITRS 2.0 Publication, "International technology roadmap for semiconductors, 2015," Available online at http://www.itrs2.net/.

[17] P.H. Drive, S. Clara, Atlas user's manual, device structure editing and simulation software; 2017. Pp. 1–1715.

[18] A.B. Sachid, M.C. Chen, C. Hu, Bulk FinFET with low-κ spacers for continued scaling, IEEE Trans. Electron Devices 64 (4) (2017) 1861–1864, https://doi.org/10.1109/TED.2017.2664798.

[19] K. Kumar, A. Kumar, V. Kumar, S.C. Sharma, Comparative investigation of band gap and gate metal engineered novel Si0. 2Ge0. 8/GaAs charge plasma-based JLTFET for improved electrical performance, Silicon (2023) 1–14.

[20] A. Kumar, N. Gupta, S. Singh, B. Tiwari, M.M. Tripathi, R. Chaujar, Carbon nanotube recessed channel (CNT-RC) MOSFET for high linearity/ULSI applications, in: TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 2564–2567, https://doi.org/10.1109/TENCON.2019.8929540.

[21] A. Kumar, N. Gupta, S.K. Tripathi, M. Tripathi, R. Chaujar, Performance evaluation of linearity and intermodulation distortion of nanoscale GaN-SOI FinFET for RFIC design, AEU-Int. J. Electron. Commun. 115 (2019), 153052.

[22] S. Sahu, M.M. Tripathi, A. Kumar, Performance investigation of nanoscaled GaN-BTG MOSFET for analog/linearity and low power applications, in: 2021 IEEE International Conference on Nanoelectronics, Nanophotonics, Nanomaterials, Nanobioscience & Nanotechnology (5NANO), Kottayam, Kerala, India, 2021, pp. 1–5, https://doi.org/10.1109/5NANO51638.2021.9491128.

[23] N. Gupta, A. Jain, A. Kumar, 20 nm GAA-GaN/Al2O3 nanowire MOSFET for improved analog/linearity performance metrics and suppressed distortion, Appl. Physics A 127 (7) (2021) 1–9.

[24] N. Gupta, A. Kumar, R. Chaujar, Effect of dielectric engineering on analog and linearity performance of gate electrode workfunction engineered (GEWE) silicon nanowire MOSFET, in: 2015 IEEE 15th International Conference on Nanotechnology (IEEE-NANO), Rome, Italy, 2015, pp. 928–931, https://doi.org/10.1109/NANO.2015.7388768.

[25] A. Kumar, M.M. Tripathi, R. Chaujar, Linearity and distortion assessment of black phosphorus-based junctionless RC MOSFET, in: 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2018, pp. 1649–1652, https://doi.org/10.1109/RTEICT42901.2018.9012398.

[26] K. Kumar, A. Kumar, S.C. Sharma, Analog/RF performance analysis of a novel Si0.9Ge0.1/InAs charge plasma-based junctionless TFET, in: 2022 IEEE International Conference of Electron Devices Society Kolkata Chapter (EDKCON), Kolkata, India, 2022, pp. 375–379, https://doi.org/10.1109/EDKCON56221.2022.10032951.

Praween Kumar Srivastava received B.*Sc* and M.*Sc* degree from Gorakhpur University Gorakhpur. He is pursuing his Ph. D. degree from the Electronics and Communication Engineering Department in Jaypee Institute of Information Technology (JIIT), Noida. He has worked as Associate professor in ECE department from 2006 to 2022 in UPTU.

Atul Kumar received B.Tech degree in Electronics & Communication Engineering from Uttar Pradesh Technical University, M.Tech degree in Electronics & Communication Engineering from National Institute of Technology, Kurukshetra and PhD in Electronics & Communication Engineering from Jaypee Institute of Information Technology, Noida, India. He is presently employed as Assistant Professor (Senior Grade) in Electronics and Communication Engineering Department at Jaypee Institute of Information Technology, Noida, India. His main research areas comprise analog signal processing and current-mode circuits. He has around 50 research publications published in reputed International Journals and International Conferences.

Ajay Kumar received the B.Tech. degree in Electronics and Communication from UPTU, Lucknow, India in 2009 and M. Tech. degree in Nanoscience and Technology from Delhi Technological University, Delhi, India in 2014. He has completed his Ph.D. degree in Electrical Engineering Department, Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India. He is currently Assistant Professor in Electronics and Communication Engineering Department in Jaypee Institute of Information Technology (JIIT), Ajay Kumar is the senior member of IEEE, OSA and other professional societies.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17418-y

Secure chaotic image encryption method using random graph traversal and three step diffusion

Varun Agarwal[1] · **Dhirendra Kumar**[1]

Received: 8 August 2021 / Revised: 28 August 2023 / Accepted: 1 October 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
In this paper, we have presented an image encryption technique using random graph traversal, chaotic maps, and a secure hashing algorithm. Due to the innate features of images, such as a high correlation between pixels, high redundancy, and the bulk capacity of images, traditional cryptography methods cannot be used for image encryption. Several theories can be exploited for image cryptography, but we choose random graph traversal because of its immense potential in this field. Chaotic maps are highly sensitive to initial conditions, which makes them suitable for encryption tasks. It is difficult to predict the dynamic properties of high-dimensional maps since they are more complex in nature. We use a three-dimensional Lorenz map to make our algorithm more robust. The initial parameters of the Lorenz map are modified by using the SHA-3 algorithm in the plain image. All the pixels of the image are assumed to be nodes of a graph, and a novel method is used to find a random path of pixel nodes in this graph network using a Lorenz map sequence such that each pixel node is visited exactly once. Finally, the chaotic sequences are used to diffuse the image into three sub-parts to obtain the final cipher image. The experiment has been performed on several standard images to compare the proposed approach with related methods in terms of qualitative and quantitative measures. The proposed approach fails to decrypt the cipher image even with a key deviation as low as 10^{-17} compared to related methods, which show excellent key sensitivity compared to the proposed method. Experimental results show that the proposed algorithm is highly secure and can resist several attacks.

Keywords Image encryption · Lorenz map · Random graph traversal · Chaotic sequence

1 Introduction

With the rapid development of internet services, vast amounts of data in multimedia forms like images, text, sound, etc. are shared through various unsecure networks. Some of this

✉ Dhirendra Kumar
dhirendrakumar@dtu.ac.in

Varun Agarwal
agarwalvarun2000@gmail.com

[1] Department of Applied Mathematics, Delhi Technological University, Delhi, India

Published online: 26 October 2023

shared data is highly confidential, and it is essential to protect it from unauthorized access. The three major goals for information security are data confidentiality, data integrity, and data authentication. Mathematical theories have extensively been used with programming languages to develop efficient data encryption methods and achieve goals of information security. Some of the popular encryption algorithms for text data are IDEA, DES, and RSA. However, these traditional algorithms cannot be used for encrypting images because of some inherent features of images, such as the high correlation between pixels, high redundancy, and bulk capacity of images [1]. In recent years, encryption techniques based on chaotic systems have gained the attention of researchers. Because chaotic key streams contain a wide key space and good unpredictability, they are suitable for the encryption of image data with high correlation [2, 3].

Chaos theory is a field of mathematics that focuses on the study of chaos and deals with nonlinear dynamical systems [4]. The theory of chaos has shown wide applications in several fields such as geology, microbiology, economics, finance, algorithmic trading, computer science, physics, etc. [5]. It is also used extensively in the field of cryptography because of the highly complex properties of these systems, such as their sensitivity to initial conditions, pseudo-random property, and non-periodicity of their states [6]. Moreover, chaos systems are highly complex in nature, which makes the encryption algorithm more robust and reliable. Some of the chaotic functions used in past image encryption studies are the Logistic function, 3-D Arnold Cat map, Baker map, Lorenz map, Chen map, etc. Higher-dimensional chaotic maps are more complex and robust in nature, making them a better choice for image encryption. The image encryption algorithm is divided into two parts, namely, the permutation stage and the diffusion stage [7]. The permutation stage deals with randomly permuting the pixels of an image without altering the pixel values. In the diffusion stage, the pixel values of the permuted image are altered. In the research work [8], the authors used a chaotic logistic function and a linear feedback shift operator to produce a pseudo-random sequence and used it as an encryption key to encrypt the image. However, these encryption methods were not secure enough to resist differential attacks on images. To overcome this, the authors of the research paper [9] proposed a novel encryption algorithm using a 1-D Logistic map, 3-D chaotic maps, and the secure hash algorithm 3 (SHA-3). Because of the high sensitivity of SHA-3 to message inputs, even a small change in the pixel value of the input image gives a new cipher image. In their research paper, [10], the authors proposed a $2D - e\pi$ hyperchaotic non-linear discrete map using $Euler$ and π numbers. The map is used to generate two sequences that are utilized in the confusion and diffusion stages. In the research work [11], the authors suggested an encryption algorithm that utilizes SHA-256 for introducing sensitivity to plain images. The authors utilize the two-dimensional Logistic-Sine-Coupling Map, the Logistic-Sine-Cosine Map, and cellular automata for confusion and diffusion. In the research work [12], the authors proposed a two-dimensional cross-mode hyperchaotic map based on logistic and sine maps (2D-CLSS). Further, an image encryption system is developed using the simultaneous combination of permutation and diffusion to alter the location and the value of the pixels via the sequences generated using the proposed 2D-CLSS map. In the research work [13], the authors utilize a fractal sorting matrix for the confusion process along with a global pixel diffusion mechanism using two chaotic sequences to develop a secure image cryptography method. In the research work [14], authors proposed a mixed one and two dimensional chaotic map with more complex non linear dynamics to develop image cryptography system. Further, authors suggested simultaneous shuffling and diffusion mechanism for robust image encryption using the chaotic sequences obtained from the proposed map. The authors of [15] used the Arnold Cat map in the permutation stage and multi-chaotic maps in the diffusion stage to develop an efficient encryption algorithm.

Many interdisciplinary fields and chaos theory have been combined to develop new image encryption techniques in the past decade. The authors of [16–18] combined neural networks to chaos theory for developing image encryption techniques. Chai et al. [19, 20] combined DNA coding techniques and chaos theory to build novel encryption algorithms. Various studies also focused on combining image compression and image encryption techniques. In the research work [21], the authors used chaotic compressing sensing using discrete cosine transform and nonlinear exponential function to encrypt images. The authors of [22, 23] used wavelet transformation image compression techniques with chaos theory. These compression-based algorithms ensure that the transmission cost and viability are as low as possible. Other disciplinary fields that are employed in image encryption are quantum mechanics, genetic recombination, and matrix tensor products. It is essential to constantly develop new encryption algorithms because the adversary may crack the algorithm by referring to a similar technique from the database [24].

Graph theory has wide-ranging applications in several fields such as computer networks, biology, computer science, etc. It is worth mentioning that graph theory also has the potential to show excellent applications in the field of data security. The research work [25] used graph theory with chaotic logistic function to make the coding algorithm more convoluted. The authors of [26] used random walks and two hyper-chaotic sequences to propose a novel image encryption algorithm. The image was divided into several blocks, and random walk matrices were used to permute these blocks individually. To improve security, an algorithm was used to permute the pixels between the permuted blocks, followed by diffusion using Chen's hyper-chaotic system. The random Hamiltonian path of a graph is also extensively used in text cryptography, and it also has wide-ranging applications in image cryptography. The research work [3] used a random Hamiltonian path to permute pixels of the image.

The encryption of data never guarantees confidentiality, as it is dependant on the key space of the encryption process which may be susceptible to several type of attacks. So, it is necessary to create strong cryptography systems with large secret key space. Image cryptography systems utilize permutation and diffusion of the pixel locations and pixel intensity values respectively for encryption process. Therefore, it is essential to develop a methodology which can produce a highly random permuted image in the permutation stage. The graph theory can be leveraged to produce such an algorithm because of its immense potential to introduce psuedo-randomness using the concept of Hamiltonian paths.

Motivated by the above discussion, we attempt to propose a secure image encryption algorithm that utilizes random graph traversal based on Hamiltonian path, 3-dimensional Lorenz chaotic map, and SHA-3 for developing a robust image cryptography system. The contributions of the research work are summarised as follows:

- A technique to modify the initial values of chaotic map using the SHA-3 value obtained from input image is introduced. In this way, the proposed encryption technique becomes resistant to differential attacks.
- Random graph traversal method is proposed for generating a highly random permuted image in the permutation step.
- Three step diffusion method is suggested to make the encryption method robust.
- An extensive comparison with state-of-the-art methods is performed to demonstrate the effectiveness of the proposed technique.

The input image is passed through SHA-3, and the hashed values are used to modify the initial parameters of the Lorenz map. We introduce the random graph traversal method to permute the pixels of the image in the permutation stage. All the pixels in an image are assumed as nodes of a fully connected graph and pseudo-random sequences generated by the Lorenz map

are used to obtain a random path traversal of this graph in which every node is visited only once. The permutation stage terminates after reshaping this random sequence of pixel values into M×N dimensions. Thereafter, this permuted image is diffused using Lorenz sequences to obtain the final encrypted image. The proposed encryption technique shows promising results in comparison with existing techniques on qualitative and quantitative analysis. The rest of the paper is organized as follows. In Section 2, some preliminary concepts are discussed in detail. In Section 3, we focus on the proposed encryption algorithm. Section 4 focuses on quantitative and qualitative experimental results, and we finally conclude in Section 5.

2 Preliminary concepts

The description of notations used in this work and related definitions are described in this section.

2.1 Lorenz map

The Lorenz map [27] is a high dimensional continuous chaotic map. The prediction of its dynamic properties is difficult because high-dimensional maps are more complex in nature. The mathematical equations for the Lorenz map are:

$$\begin{aligned} \dot{x} &= a(y - x) \\ \dot{y} &= cx - y - xz \\ \dot{z} &= xy - bz \end{aligned} \tag{1}$$

Here $\mathbf{X} = [x\ y\ z]^T$ refers to three state variables and a, b and c represent system parameters for the Lorenz map. $[\dot{x}\ \dot{y}\ \dot{z}]$ represent derivative of variables with respect to time. For $a = 10$, $b = 8/3$, and $c = 28$ Lorenz map is said to be in a chaotic state.

2.2 Graph theory

A graph is a data structure, made up of a collection of vertices connected via a set of edges. A graph can either be directed or undirected depending on whether the edges link the vertices symmetrically or asymmetrically. Furthermore, we can classify a graph as a weighted graph or an unweighted graph depending on whether the edges have weights or not. A graph is said to be a fully connected graph if each pair of vertices is connected by an edge i.e. if n is the number of vertices in a graph, then the number of edges will be $(n * (n - 1))/2$. One of the applications of graph theory is link prediction in social networks, in which all the individuals in a social network are assumed to be vertices of a graph and any two vertices are connected by an edge if the two individuals are connected on the social media platform. We can predict future links that will be formed in this social network according to its topological properties and the properties of individual [28]. Another application can be DNA sequencing in which the theory is used to ascertain the sequence of nucleotides in a DNA strand [29]. We utilize the graph theory concept to develop the image encryption algorithm in this work. A graph can be denoted as $G = (V, E)$ where V represents the set of vertices and E represents the set of edges.

3 Proposed methodology

The proposed methodology for image encryption is divided into two phases, namely permutation and diffusion of pixels. Firstly, the image is passed through 256 bits SHA-3 to obtain a resultant sequence of 32 elements each of size one byte. This array is used to modify the initial parameters of the Lorenz map so that we can get three unique pseudo-random sequences for the input image. After that, one of the pseudo-random sequences is utilized for random Graph traversal to permute the image which ends the first phase of the algorithm. In the diffusion phase, pixel values of the permuted image are altered by the proposed three-step diffusion process.

3.1 Generating initial values

The 256-bit SHA-3 of the input image gives us 32 integers (8 bits each) and we utilize these numbers in the form of an integer array of length 32 represented using K. These integers are used to obtain parameters p_1, p_2, p_3, and p_4 using (2), (3), (4), and (5) respectively.

$$p_1 = (K[0] \oplus K[1]) * (K[2] \oplus K[3]) * (K[4] \oplus K[5]) * (K[6] \oplus K[7]) \tag{2}$$

$$p_2 = (K[8] \oplus K[9]) * (K[10] \oplus K[11]) * (K[12] \oplus K[13]) * (K[14] \oplus K[15]) \tag{3}$$

$$p_3 = (K[16] \oplus K[17]) * (K[18] \oplus K[19]) * (K[20] \oplus K[21]) * (K[22] \oplus K[23]) \tag{4}$$

$$p_4 = (K[24] * K[25]) * (K[26] * K[27]) * (K[28] * K[29]) * (K[30] * K[31]) \tag{5}$$

Where $\oplus$ is the XOR operation and $*$ is the multiplication operation. Each of the values p_1, p_2, p_3 is divided by 10 until their corresponding values become less than 1. Let these modified values be p_1', p_2' and p_3'. Initial values of the Lorenz map are given which are x_0, y_0, z_0. These values are modified by (6), (7) and (8) :

$$x_0' = x_0 p_1' \tag{6}$$

$$y_0' = y_0 p_2' \tag{7}$$

$$z_0' = z_0 p_3' \tag{8}$$

3.2 Random graph traversal for permutation phase

We iterate through the Lorenz map using the modified initial values $x_0^{'}$, $y_0^{'}$, $z_0^{'}$ to obtain three pseudo-random sequences (each of length $M \times N$ where $M > 1$ is the length and $N > 1$ is the width of the original input image). From these three pseudo-random sequences, the first sequence is used for random graph traversal in the permutation phase. As some of the values in this random sequence may be negative, we modify this sequence by adding the lowest value of the sequence to each element to produce a non-negative random sequence. This scaled sequence is reshaped to form a $M \times N$ dimensional chaotic matrix $E^{(0)}$. The sum of all values inside the generated chaotic matrix is computed and used to modify each pixel value of the matrix using (9)

$$E_{i,j}^{(0)} = \frac{E_{i,j}^{(0)}}{(sum - E_{i,j}^{(0)})} \tag{9}$$

where $sum = \sum_i \sum_j E^{(0)}_{i,j}$ is the initial sum of all values of chaotic matrix $E^{(0)}$ and $E^{(0)}_{i,j}$ is the value in i^{th} row and j^{th} column of the chaotic matrix $E^{(0)}$.

All the pixel positions of the initial input image (size M×N) are assumed as nodes of a fully connected graph. Since the graph is fully connected, we can assume that a connecting edge is present for each pair of pixels. The objective is to find a starting point and acquire a random traversal of this graph in which each node is visited exactly once. The starting node for graph traversal is obtained using (10) and (11).

$$r = floor((p_4\%(M * N))/M) + 1 \tag{10}$$

$$c = ((p_4\%(M * N))\%M) + 1 \tag{11}$$

where % is the modular operation, $*$ is the multiplication operation, / is the division operation. The cell corresponding to r^{th} row and to c^{th} column of the matrix is considered as the starting node for traversal. The next node that we visit will be an unvisited node corresponding to i^{th} row and j^{th} column such that the absolute value of $E^{(0)}_{i,j} - param$ is minimum where param is calculated by (12).

$$param = ((p_4\%(M * N))/(M * N)) * (maxE - minE), \quad (maxE - minE) \neq 0 \tag{12}$$

where % is the modular operation, $*$ is the multiplication operation, / is the division operation, $maxE$ is maximum value in chaotic matrix $E^{(0)}$ and $minE$ is minimum value in chaotic matrix $E^{(0)}$.

The random traversal is done until all nodes are visited. The order of nodes visited in this traversal gives a random sequence of pixels. Finally, we reshape this sequence to a matrix of size $M \times N$ to obtain a permuted cipher matrix.

3.3 Three step diffusion algorithm

To enhance the security of the pixel value level in the image encryption process, we have proposed a three-step diffusion process on the permuted image obtained from the previous phase. For this, the pixel values of the permuted image C are now diffused using the other two chaotic sequences for obtaining the encrypted image C'. To diffuse the image C, two chaotic matrices $E^{(1)}$ and $E^{(2)}$ of size $M \times N$ are obtained using the initial $M * N$ elements from the second and third sequence of Lorenz map respectively (Note: each sequence is modified every value in the series is non-negative as done for the first sequence). Each value of both matrices $E^{(1)}$ and $E^{(2)}$ is converted to integer using greatest integer function. Using these two matrices, i.e., $E^{(1)}$ and $E^{(2)}$, we perform element-wise diffusion operations on permuted image C in three steps (i) diagonal elements (ii) upper triangular elements and (iii) lower triangular elements separately. The involved operations are described as follows:

Step 1: The encrypted grayscale values C' for the diagonal elements are obtained using the following operation:

$$C'_{i,i} = C_{i,i} \oplus E^{(1)}_{i,i} \oplus E^{(2)}_{i,i} \tag{13}$$

Step 2: The encrypted grayscale values C' for the upper triangular elements are obtained using the following operation:

$$C'_{i,j} = C_{i,j} \oplus E^{(1)}_{i,j} \oplus E^{(2)}_{i,j} \oplus C_{i,j-1} \oplus C_{i+1,j} \tag{14}$$

Step 3: The encrypted grayscale values C' for the lower triangular elements are obtained using the following operation:

$$C'_{i,j} = C_{i,j} \oplus E^{(1)}_{i,j} \oplus E^{(2)}_{i,j} \oplus C_{i-1,j} \oplus C_{i,j+1} \tag{15}$$

Where $\oplus$ is XOR operation, $C_{i,j}$ is the value at i^{th} row and j^{th} column of permuted matrix, $E^{(1)}_{i,j}$ and $E^{(2)}_{i,j}$ are the value of chaotic matrix $E^{(1)}$ and $E^{(2)}$ at i^{th} row and j^{th} column respectively.

The diffusion stage is complete and the final encrypted image C' is obtained. Further, the decryption process of the cipher image can be performed in a similar way.

The flow diagram of the proposed methodology is given in Fig. 1.

3.4 Decryption process of the proposed methodology

The secret keys x'_0, y'_0 and z'_0 are used to generate the Lorenz sequences, and these sequences are used to generate chaotic matrices $E^{(1)}$ and $E^{(2)}$. We use (13), (14), and (15) in diagonal, upper triangular, and lower triangular elements in the cipher image along with $E^{(1)}$ and $E^{(2)}$ to invert the diffusion process. Further, we use secret keys p_4, M, and N to calculate $param$ using (12) and use the first sequence of the Lorenz map to compute chaotic matrix $E^{(0)}$ as done earlier in the encryption process. To invert the permutation process, we assume each pixel as a node of a fully connected graph and start the inverse random traversal from the pixel present in row r and column c computed from (10) and (11) respectively. The next node that we visit will be an unvisited node for which the value of $E^{(0)}_{i,j} - param$ is maximum. Similarly, we visit all the nodes to invert the permutation process and finally obtain the plain image. The flow diagram of the proposed decryption methodology is given in Fig. 2.

3.5 Time complexity analysis

There are two important steps in the proposed technique: random graph traversal based permutation and 3-step diffusion. In the permutation step, we consider each pixel as the node of a fully connected graph. The random graph traversal algorithm is implemented using a custom priority queue with the chaotic sequences. Therefore, the time complexity of the permutation algorithm is $O(MNlogMN)$ where M and N are the dimensions of the plain image. The

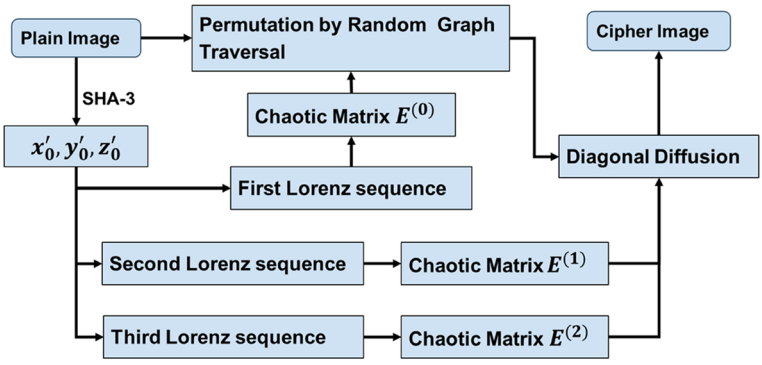


Fig. 1 Flow diagram of the proposed methodology

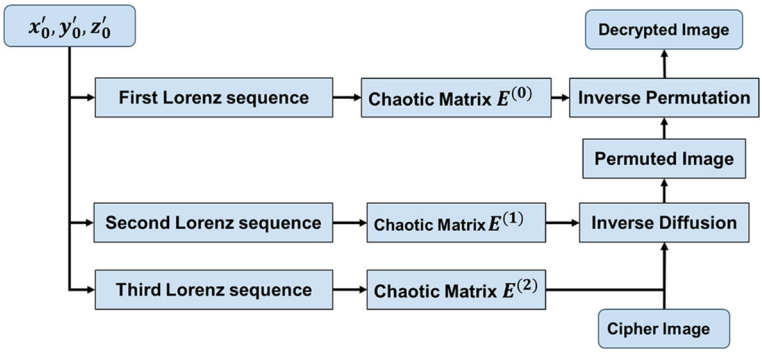


Fig. 2 Flow diagram of the proposed decryption methodology

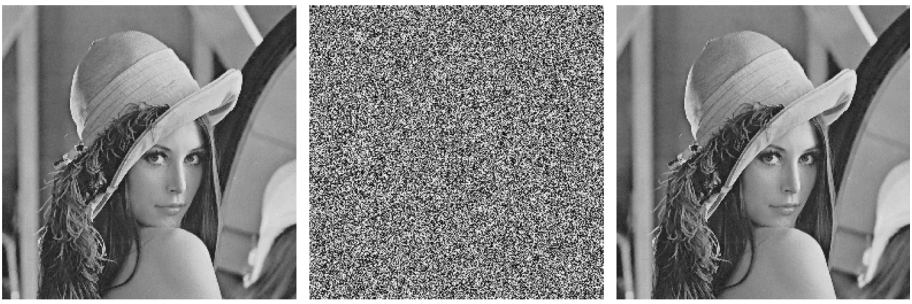
time complexity of the proposed 3-step diffusion algorithm is $O(MN)$. Hence, the final time complexity of the encryption algorithm is $O(MNlogMN) + O(MN) \approx O(MNlogMN)$. Although there are some past research works with time complexity of $O(MN)$, they provide poor security in comparison to our proposed image encryption approach in terms of many security metrics. The detailed description of comparison on these security metrics with the-state-of-the-art methods are given in Section 4.

4 Simulation results

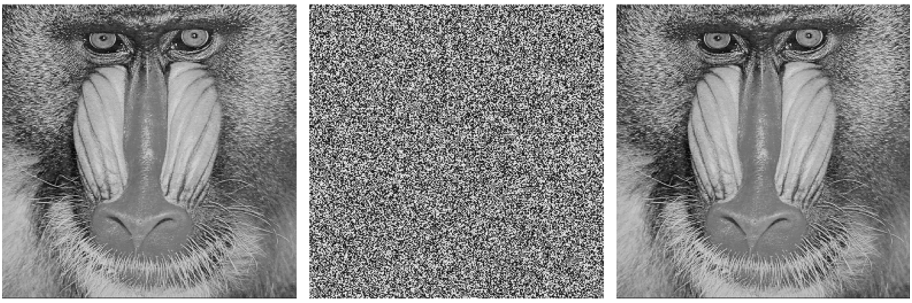
To evaluate the performance of the proposed image encryption algorithm, we have utilized standard images of different sizes for the simulation. Further, to check the effectiveness of the proposed method, we use several quantitative and qualitative evaluation metrics such as key space and perform a detailed comparison of the proposed method with other state-of-the-art methods. We use Python 3 to implement the proposed method and the-state-of-the-art methods and and compare the performance on a computer with macOS Mojave, a 3.1 GHz Intel Core i5 processor, and 8GB RAM. The initial values of Lorenz map are set to $\{x_0, y_0, z_0\} = \{0.023154673, 0.0175686547, 0.056147382\}$ for all the images for encryption. Images of different sizes, i.e., 256×256 (Lena, Baboon, Bird, Flower, and two medical images) and 512×512 (Lena, Baboon, and Tiffany) are utilized to make extensive evaluation and comparison of the proposed approach with other recent methods. The experimental results show that the method has good security and can resist various attacks. Figure 3 represents the encrypted and decrypted images using the proposed approach.

4.1 Key space analysis

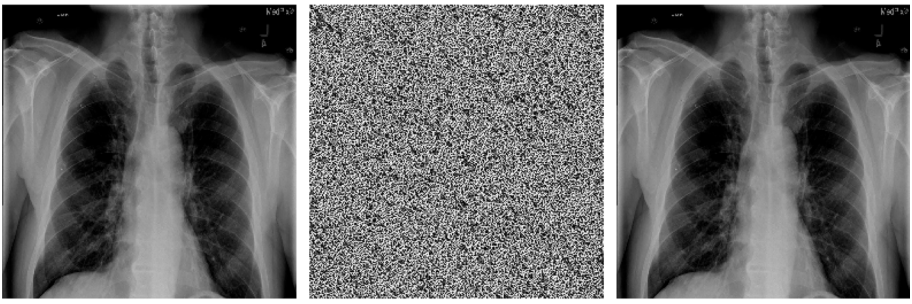
The secret keys required to decrypt the encrypted image are x'_0, y'_0, z'_0, p_4 , M, N where x'_0, y'_0, z'_0 are modified initial parameters of Lorenz map, M is the length and N is width of the original plain image. It has been reported that the key space of any image encryption decryption technique should be greater than 2^{100} for resisting brute force attacks [30]. We use a 64-bit computer with a calculation precision of 10^{-15}, which makes our key space 10^{90}. This value is higher than 2^{100} so we can conclude that the algorithm can effectively resist brute-force attacks.



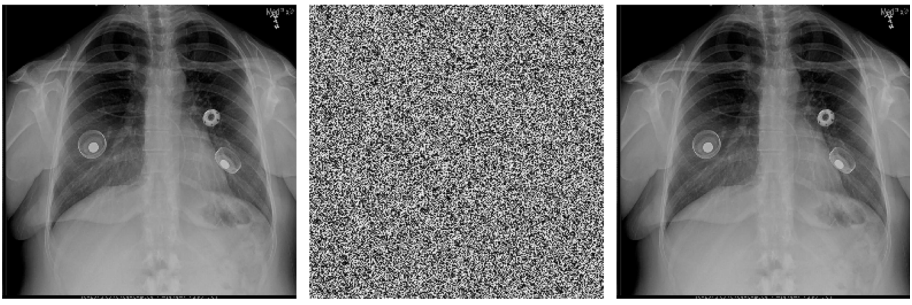
(a) Plain Lena Image (b) Encrypted Lena Image (c) Decrypted Lena Image



(d) Plain Baboon Image (e) Encrypted Baboon Image (f) Decrypted Baboon Image



(g) Plain Medical Image 1 (h) Encrypted Medical Image 1 (i) Decrypted Medical Image 1



(j) Plain Medical Image 2 (k) Encrypted Medical Image 2 (l) Decrypted Medical Image 2

Fig. 3 Encryption and decryption simulation result

4.2 Key sensitivity analysis

Key sensitivity means that even if there is a slight change in secret keys, the algorithm will not be able to decrypt the encrypted image. A good algorithm has high key sensitivity [31]. We have performed the experiment to check the key sensitivity of the proposed method and several related methods [3, 8–13, 20], and [24]. The minimum possible deviation in the values of secret key for decrypting the image successfully is lower in comparison to the related methods indicating higher key sensitivity. Further, we have shown a case for decrypting the encrypted Lena image in Fig. 4 using the proposed approach with the key deviations 10^{-18} and 10^{-17}. It can be noted from Fig. 4 the image is successfully decrypted using a key deviation of 10^{-18} (see Fig. 4 (c)) whereas a key deviation of 10^{-17} fails to do so (see Fig. 4 (d)).

4.3 Histogram analysis

The graphical distribution of image pixels can be represented by a histogram. We use histogram analysis to evaluate the performance of the proposed encryption technique along

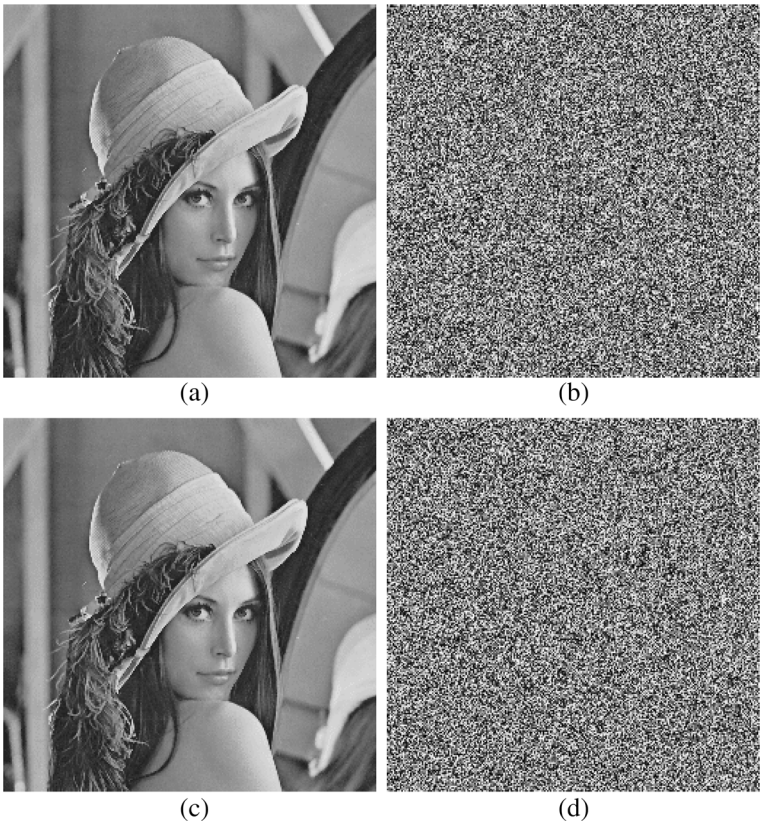


Fig. 4 Key sensitivity analysis results (**a**) Plain Lena image (**b**) Encrypted Lena Image (**c**) Decrypted Lena image with key deviation of 10^{-18} (**d**) Decrypted Lena image with key deviation of 10^{-17}

with the related methods [3, 8–13, 20], and [24]. It has been reported in the literature that the histogram distribution of cipher images formed by a secure encryption technique is always uniform [32]. Figure 5 represents the results of histogram analysis of cipher images obtained using the proposed method. Similar results have been obtained for other related methods. It can be noted from Fig. 5 that the histogram plot of cipher images is uniform for all four images. From Fig. 5, it can be concluded that the encryption image is well diffused.

4.4 Information entropy

Information entropy describes the degree of randomness and unpredictability of an image source [33]. We define information entropy for a source c as:

$$H(c) = \sum_{i=0}^{2^N-1} p(c_i) \log \frac{1}{p(c_i)} \tag{16}$$

where $p(c_i)$ is probability of sample c_i and N is the number of samples. It signifies the distribution of grayscale pixel values. The entropy of the cipher image is higher than the plain image because the distribution of pixels is uniform for the cipher image. The highest theoretical value of Information entropy is 8. The higher the entropy information, the more secure the encryption algorithm. To evaluate and compare the performance of the proposed method with the related method, the information entropy index is obtained on several images. Table 1 shows the values of information entropy for the different images using different methods along with the proposed approach. It can be noted from Table 1 that the proposed

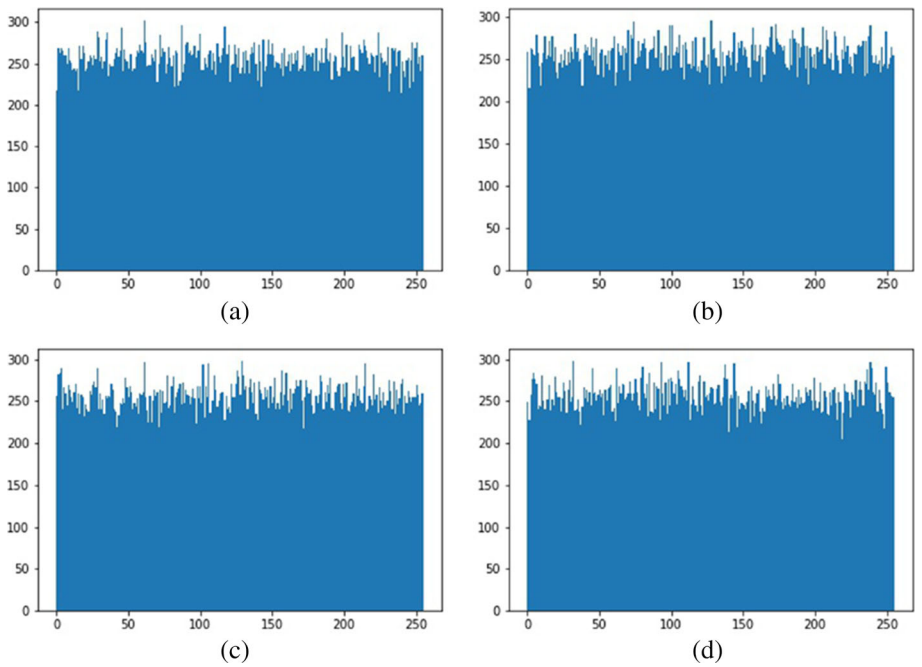


Fig. 5 Simulation result for histogram analysis (**a**) histogram of encrypted Lena image (**b**) histogram of encrypted baboon image (**c**) histogram of encrypted bird image (**d**) histogram of encrypted flower image

Table 1 Information Entropy

Image (size)	Proposed Method	Ref. [8]	Ref. [24]	Ref. [9]	Ref. [3]	Ref. [20]	Ref. [10]	Ref. [11]	Ref. [12]	Ref. [13]
Lena (256 × 256)	7.9971	7.9968	7.9973	7.9144	7.9970	7.9969	7.9973	7.9750	7.9971	7.9974
Baboon (256 × 256)	7.9967	7.9971	7.9971	7.9302	7.9973	7.9970	7.9973	7.9787	7.9967	7.9966
Bird (256 × 256)	7.9975	7.9971	7.9968	7.8871	7.9973	7.9968	7.9967	7.9780	7.9972	7.9973
Flower (256 × 256)	7.9971	7.9971	7.9964	7.8525	7.9973	7.9974	7.9974	7.9769	7.9974	7.9968
Medical 1 (256 × 256)	7.9969	7.9973	7.9971	7.9746	7.9971	7.9974	7.9968	7.9768	7.9974	7.9972
Medical 2 (256 × 256)	7.9973	7.9972	7.9972	7.9048	7.9968	7.9970	7.9974	7.9765	7.9974	7.9975
Lena (512 × 512)	7.9993	7.9993	7.8248	7.9729	7.9986	7.9992	7.9992	7.9788	7.9993	7.9992
Baboon (512 × 512)	7.9993	7.9992	7.8472	7.9866	7.9985	7.9992	7.9993	7.9791	7.9991	7.9993
Tiffany (512 × 512)	7.9985	7.9993	7.8476	7.8438	7.9984	7.9992	7.9993	7.9789	7.9992	7.9992
Average Values	7.9977	7.9978	7.9446	7.9185	7.9975	7.9977	7.9978	7.9776	7.9979	7.9978

method is secure as the information entropy is close to 8 for each sample image. It can be observed that the information entropy of the proposed technique is highest for Bird image as compared to the related methods. Additionally, the information entropy for other images are comparable to the past works. This observation can be attributed to the randomness properties of the 3-step diffusion process of the proposed encryption method.

4.5 Correlation analysis

4.5.1 Adjacent correlation

It is a statistical analysis technique that is used to analyze the relationship between adjacent pixels of an image. We can calculate the correlation coefficient between adjacent pixels of an image in horizontal, vertical, and diagonal directions. It is reported in the literature that an image encryption algorithm is secure if the cipher image produced has a low correlation between adjacent pixels [34]. To validate the security of the proposed method, we have sampled a total of 2048 pixel values. For all these sampled pixel values, we obtained the adjacent pixel values in vertical, horizontal, and diagonal directions. In this way, we obtain three sets of ordered pairs of size 2048 in vertical, horizontal, and diagonal directions. The plot of all three sets for the plain image and the encrypted image is shown in Fig. 6. It can be observed that the plots for the plain image are scattered around a straight line which signifies a high correlation between adjacent pixel values for all three sets (See Fig. 6 (a), (b), and (c)).

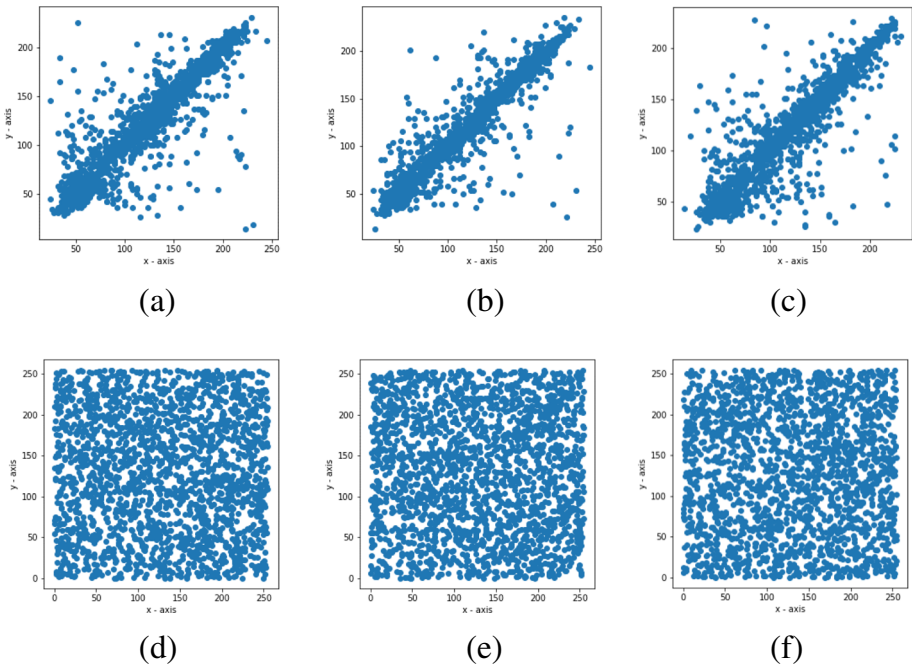


Fig. 6 The plot of 2048 random adjacent pixel pairs for Lena plain image (**a**)horizontal direction (**b**) vertical direction (**c**) diagonal direction and Lena cipher image (**d**) horizontal direction (**e**) vertical direction (**f**) diagonal direction

Table 2 Correlation coefficients of plain images

Plain Image (size)	Horizontal	Vertical	Diagonal
Lena (256 × 256)	0.909427	0.942526	0.908766
Baboon (256 × 256)	0.785339	0.694908	0.662920
Bird (256 × 256)	0.980655	0.989480	0.981102
Flower (256 × 256)	0.927709	0.951748	0.893393
Medical Image 1 (256 × 256)	0.982717	0.988801	0.978201
Medical Image 2 (256 × 256)	0.978026	0.959856	0.955277
Lena (512 × 512)	0.963332	0.975227	0.961956
Baboon (512 × 512)	0.850006	0.727110	0.666820
Tiffany (512 × 512)	0.926766	0.955208	0.944259

However, the plots for the cipher image are scattered uniformly on the plane which shows that after encryption the data has no significant relationships among them.

Further, to evaluate the quantitative measure for estimating the relationship of pixel values pair from all three sets, the correlation coefficient has been evaluated for the proposed method and compared with the other related methods available in literature. The correlation coefficient value r can be obtained using the following formula:

$$r = \frac{Cov(x, y)}{\sqrt{D(x)}\sqrt{D(y)}} \tag{17}$$

where $Cov(x, y)$ is co-variance of x and y. $D(x)$ and $D(y)$ are variances of x and y respectively.

Tables 2 and 3 represent the correlation coefficient of 2048 randomly selected pixel values from the plain images and cipher images for all three sets of pair of pixel values, i.e., horizontal, vertical, and diagonal directions. From Tables 2 and 3, it can be observed that the correlation between the pixel value are high in the case of plain images whereas the correlation coefficients are around zero showing the low correlation between the pixel values in the cipher images. Further, Table 4 shows the correlation coefficients of the Lena cipher image obtained using the proposed and other related methods [3, 8–13, 20], and [24]. From Table 4, it can be concluded that the correlation between adjacent pixels tend to vanish due

Table 3 Correlation coefficients of cipher images

Cipher Image (size)	Horizontal	Vertical	Diagonal
Lena (256 × 256)	-0.001885	-0.012793	0.007396
Baboon (256 × 256)	0.016159	-0.006297	0.003388
Bird (256 × 256)	-0.022028	-0.023611	-0.006860
Flower (256 × 256)	0.004765	0.035096	0.013687
Medical Image 1 (256 × 256)	0.004327	-0.005094	-0.009582
Medical Image 2 (256 × 256)	-0.005368	-0.010886	0.011465
Lena (512 × 512)	-0.023191	0.014196	0.009235
Baboon (512 × 512)	-0.015672	-0.017938	-0.003838
Tiffany (512 × 512)	0.047668	0.019182	0.012051

Table 4 Comparison of correlation coefficients obtained for the encrypted Lena image

Methods\Correlation direction	Horizontal	Vertical	Diagonal
Proposed algorithm	-0.001885	-0.012793	0.007396
Ref. [8]	-0.022122	0.037363	-0.010696
Ref. [24]	-0.001748	-0.025503	0.012221
Ref. [9]	-0.015595	0.028436	-0.007102
Ref. [3]	0.016720	0.006996	0.011698
Ref. [20]	-0.040583	-0.027371	-0.014449
Ref. [10]	0.004555	-0.015623	-0.005112
Ref. [11]	0.009998	0.001372	-0.006567
Ref. [12]	0.009901	-0.019352	0.000631
Ref. [13]	-0.007931	0.002863	0.001745

to effective permutation and 3 step diffusion procedure utilized in the proposed encryption approach.

4.5.2 Cross correlation

A strong image cryptography system reduces the correlation between the plain and cipher images significantly such that the resistance to differential attack is increased. A low cross-correlation (CC) makes it possible to prevent a differential attack by reducing similarities between pixels of plain image and corresponding cipher image. The CC between the two is calculated using (17). The range of the CC coefficient is [-1, 1]. A correlation value of zero or near one implies no CC and a high correlation, respectively. The CC obtained corresponding to several images for the proposed approach is shown in Table 5. The Table 5 indicates that our proposed approach achieves lower values of CC for all the images showing its effectiveness.

4.6 Differential attack

A secure algorithm should be highly sensitive to minor changes in the plain image for the encryption process. For a secure encryption algorithm, a small modification in the plain image

Table 5 Cross Correlation Analysis of the proposed algorithm

Images (size)	Cross Correlation
Lena (256 × 256)	-0.017084
Baboon (256 × 256)	-0.009154
Bird (256 × 256)	0.010393
Flower (256 × 256)	0.002887
Medical 1 (256 × 256)	0.038354
Medical 2 (256 × 256)	0.004908
Lena (512 × 512)	0.020611
Baboon (512 × 512)	0.005036
Tiffany (512 × 512)	0.037579

results in a totally different encrypted image [35]. Number of Pixels Change Rate (NPCR) and Unified Averaged Changed Intensity (UACI) are evaluated to check the capability of resisting differential attack. The $NPCR$ and $UACI$ can be calculated using mathematical formula defined in (18) and (19), respectively.

$$NPCR = \frac{\sum_{i,j} D(i,j)}{M \times N} \times 100\% \tag{18}$$

$$Where,\ D(i,j) = \begin{cases} 1, & \text{if } R_1(i,j) \neq R_2(i,j); \\ 0, & \text{if } R_1(i,j) = R_2(i,j); \end{cases}$$

$$UACI = \frac{1}{M \times N}\left[\sum_{i,j} \frac{|R_1(i,j) - R_2(i,j)|}{255}\right] \times 100\% \tag{19}$$

Where M is the width, N is the height of the input image, R_1 is the encrypted image obtained from plain image, and R_2 is the modified encrypted image obtained after modifying plain image. For an image, the $NPCR$ has an ideal value of 99.6094%, and the $UACI$ has an ideal value of 33.4635% [36]. The values of $NPCR$ and $UACI$ obtained for the proposed method along with the other related research works is shown in Tables 6 and 7. From the Tables 6 and 7, it can be concluded that the algorithm can resist differential attacks since the values of NPCR and UACI are very close to the ideal values.

4.7 Peak signal to noise ratio (PSNR)

The $PSNR$ is used to assess the resilience of the proposed image cryptosystem in the presence of channel noise. The difference between the pixel intensity values of the plain and encrypted images are used to calculated the PSNR values [37]. The mathematical formula to compute $PSNR$ is represented using (20). To evaluate and compare the performance of the proposed method with the related method, PSNR value is obtained on several images. Table 8 exhibits the obtained $PSNR$ values for ciphered images coresponding to the proposed approach along with other related methods. From Table 8, it is observed that average $PSNR$ values over all the images for the proposed method is smaller and comparable to most of the related works.

$$PSNR = 10\log_{10}\left(\frac{(2^{n-1})^2}{MSE}\right) \tag{20}$$

Where n is the number of bits per pixel. 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 P and E are the original and the encrypted images respectively.

4.8 Histogram deviation

Histogram deviation is the maximum amount of deviation between the histograms of the plain image and cipher image. It is used to evaluate the ciphering quality of the cryptosystems. It can be calculated as follows:

$$D_H = \frac{\frac{d_0+d_{255}}{2}\sum_{i=1}^{254} d_i}{M \times N} \tag{21}$$

Table 6 Comparative analysis of NPCR values obtained using the proposed approach and other methods

Images (size)	Proposed Method	Ref. [8]	Ref. [24]	Ref. [9]	Ref. [3]	Ref. [20]	Ref. [10]	Ref. [11]	Ref. [12]	Ref. [13]
Lena (256 × 256)	99.6506	-	99.6200	98.9807	99.6346	99.6323	99.6154	99.6307	99.5925	99.6215
Baboon (256 × 256)	99.6307	-	99.6139	99.1073	99.6216	99.5911	99.6231	99.5941	99.5925	99.5895
Bird (256 × 256)	99.6719	-	99.6017	98.6694	99.6557	99.5773	99.6276	99.5895	99.5956	99.5986
Flower (256 × 256)	99.6154	-	99.5945	98.6343	99.6293	99.5636	99.5788	99.6154	99.6383	99.6276
Medical Image 1 (256 × 256)	99.5956	-	99.5971	98.4619	99.6319	99.6185	99.6017	99.6017	99.6124	99.6261
Medical Image 2 (256 × 256)	99.6078	-	99.5681	97.9110	99.5743	99.5925	99.6307	99.6200	99.6353	99.6429
Lena (512 × 512)	99.6059	-	99.6261	99.5033	99.6436	99.6093	99.5845	99.6082	99.6028	99.6101
Baboon (512 × 512)	99.6147	-	99.6398	99.4785	99.6268	99.6627	99.6299	99.5979	99.6040	99.6265
Tiffany (512 × 512)	99.6082	-	99.6047	99.2465	99.5985	99.6261	99.6051	99.5811	99.6150	99.6109
Average values	99.6223	-	99.6073	98.8881	99.6240	99.6082	99.6108	99.6043	99.6098	99.6171

Table 7 Comparative analysis of UACI values obtained using the proposed approach and other methods

Images (size)	Proposed Method	Ref. [8]	Ref. [24]	Ref. [9]	Ref. [3]	Ref. [20]	Ref. [10]	Ref. [11]	Ref. [12]	Ref. [13]
Lena (256 × 256)	33.4858	-	33.4156	33.0815	33.1984	33.2378	33.3377	33.0881	33.4441	33.5825
Baboon (256 × 256)	33.1598	-	33.5338	33.7694	33.2025	33.5648	33.4941	33.1112	33.5791	33.3826
Bird (256 × 256)	33.3796	-	33.2533	33.0078	33.2356	33.6060	33.4548	33.9914	33.5209	33.5529
Flower (256 × 256)	33.2517	-	33.0846	33.7483	33.0225	33.4592	33.4698	33.9385	33.4095	33.5184
Medical Image 1 (256 × 256)	33.3584	-	33.8256	33.6967	33.4945	33.5705	33.5443	33.1394	33.5875	33.5445
Medical Image 2 (256 × 256)	33.3831	-	33.4152	33.0045	33.4796	33.5141	33.4022	33.9003	33.6223	33.4464
Lena (512 × 512)	33.4209	-	33.8043	33.2334	33.6028	33.4873	33.5262	33.0228	33.3860	33.5257
Baboon (512 × 512)	33.4921	-	33.4556	33.7609	33.1173	33.4588	33.4316	33.0019	33.4130	33.4896
Tiffany (512 × 512)	33.1192	-	33.6935	33.7927	33.6096	33.3917	33.4016	33.0618	33.4744	33.4856
Average values	33.3390	-	33.4979	33.4550	33.3292	33.4767	33.4514	33.3617	33.4930	33.5031

Table 8 PSNR values obtained using the proposed approach

Images (size)	Proposed Method	Ref. [8]	Ref. [24]	Ref. [9]	Ref. [3]	Ref. [20]	Ref. [10]	Ref. [11]	Ref. [12]	Ref. [13]
Lena (256 × 256)	8.5964	8.5445	8.6313	9.0750	8.6103	8.5790	8.5668	8.8081	8.5499	8.5597
Baboon (256 × 256)	9.5309	9.5087	9.5364	9.8480	9.5357	9.5368	9.5131	9.5767	9.5267	9.5108
Bird (256 × 256)	9.3448	9.3303	9.3415	9.5290	9.3420	9.3391	9.3302	9.4079	9.3288	9.2895
Flower (256 × 256)	8.9050	8.9013	8.9210	9.2159	9.3420	9.3391	9.3302	9.4079	9.3288	9.2895
Medical Image 1 (256 × 256)	8.3988	8.3734	8.4709	9.1299	8.7457	8.3634	8.4491	8.5808	8.3899	8.4434
Medical Image 2 (256 × 256)	8.5748	8.5572	8.8612	9.4544	8.9172	8.5755	8.5428	8.8118	8.5700	8.5697
Lena (512 × 512)	9.2377	9.2248	8.8952	9.2164	8.8457	9.2229	9.2227	9.3249	9.2408	9.2791
Baboon (512 × 512)	9.5281	9.5040	9.3417	9.6125	9.0145	9.5316	9.5261	9.5682	9.5324	9.5014
Tiffany (512 × 512)	7.0106	6.8849	8.8386	9.4058	8.8372	6.8804	6.8810	6.5662	6.8951	6.4947
Average values	8.7919	9.3875	8.9758	8.9820	8.7708	8.7588	8.8533	8.7304	8.7768	8.7704

where d_i is the absolute difference between the histograms for the gray level i. The histogram deviation is obtained for several images to evaluate and compare the performance of the proposed method with other related method. Table 9 shows the histogram deviation values obtained for the proposed cryptosystem and other related methods. The Table 9 indicates that average histogram deviation values over all the images for the proposed method is higher in comparison to most of the related works which shows its superior performance.

4.9 Structural similarity index measure (SSIM)

This measure calculates the degree of similarity between plain and encrypted images [37]. The mathematical formula to calculate $SSIM$ is described in (22).

$$SSIM = \frac{(2\mu_p\mu_e + C_1)(2\sigma_{pe} + C_2)}{(\mu_p^2 + \mu_e^2 + C_1)(\sigma_p^2 + \sigma_e^2 + C_2)} \tag{22}$$

where μ_p represents the mean of the plain image and μ_e represents the mean of the encrypted image. σ_p^2 and σ_e^2 represents variance of plain and encrypted images respectively and σ_{pe} represents covariance of plain and encrypted image. Further, $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$, where $k_1 \ll 1$, $k_2 \ll 1$, and L represents the dynamic range of the pixel values for 8-bit grayscale images i.e. $L = 255$. The $SSIM$ should lie in the range $[-1, 1]$.

To evaluate and compare the performance of the proposed method with the related method, SSIM value is obtained on several images. Table 10 exhibits the $SSIM$ results for the plain and corresponding encrypted images for the proposed approach along with other related methods. From Table 10, it can be concluded that average histogram deviation values over all the images for the proposed method is small in comparison to most of the related works which shows its superior performance.

4.10 Occlusion and noise attack

There is always the possibility of data loss due to noise or other adversaries in every communication channel. The encryption and decryption algorithms should be built in such a way that even if they do not recover the complete encrypted image, they can recover a portion of it. An occlusion attack causes a large portion of the encrypted image to be lost during transmission. To show the robustness of the proposed image encryption algorithm toward the occlusion attack, we have performed experiments on Lena image of size 256×256. In our experiment, the pixels in the top-left corner of lena cipher image are changed to the value of 0 and the resultant cipher is decrypted to analyse the visual information of the decrypted image. Figure 7 shows the decrypted image after 12.5%, 25%, and 50% of the Lena cipher was removed from top-left corner. It can be observed from the figure that sufficient visual information was retained. Therefore, the proposed algorithm is averse to the occlusion attack.

Similarly, we introduced salt and pepper noise in the Lena cipher image and decrypted it to analyse the amount of visual information left. From Fig. 8, it can be observed that the decryption algorithm is able to recover some of the image's visual information, which is sufficient to understand its visual content. Hence, the proposed algorithm is also resistant to the noise attack.

Table 9 Histogram Deviation values obtained using the proposed approach

Images (size)	Proposed Method	Ref. [8]	Ref. [24]	Ref. [9]	Ref. [3]	Ref. [20]	Ref. [10]	Ref. [11]	Ref. [12]	Ref. [13]
Lena (256 × 256)	0.5792	0.5847	0.5891	0.4705	0.5821	0.5828	0.5831	0.5496	0.5886	0.5913
Baboon (256 × 256)	0.7530	0.7628	0.7556	0.6623	0.7563	0.7661	0.7557	0.7517	0.7527	0.7617
Bird (256 × 256)	0.9900	0.9884	0.9185	0.9878	0.9854	0.9960	0.9706	0.9931	0.9960	0.9968
Flower (256 × 256)	1.3853	1.3915	1.3955	1.0896	1.3839	1.3865	1.3862	1.3918	1.3893	1.3887
Medical Image 1 (256 × 256)	0.5122	0.5117	1.4929	0.3936	0.7518	0.5173	0.5056	0.5058	0.5123	0.5131
Medical Image 2 (256 × 256)	0.8485	0.7597	0.7514	0.6512	0.5136	0.7524	0.7586	0.7365	0.7593	0.7539
Lena (512 × 512)	0.6451	0.6443	0.7330	0.7269	0.7607	0.6463	0.6486	0.6486	0.6437	0.6556
Baboon (512 × 512)	0.7525	0.7546	0.8502	0.7415	0.8088	0.7557	0.7538	0.7505	0.7513	0.7612
Tiffany (512 × 512)	1.1941	1.2195	1.5676	0.9641	1.0673	1.2186	1.2189	1.2819	1.2182	1.2963
Average values	0.8511	0.7354	0.8459	1.0143	0.8457	0.8464	0.8430	0.8577	0.8454	0.8452

Table 10 SSIM values obtained using the proposed approach

Images (size)	Proposed Method	Ref. [8]	Ref. [24]	Ref. [9]	Ref. [3]	Ref. [20]	Ref. [10]	Ref. [11]	Ref. [12]	Ref. [13]
Lena (256 × 256)	0.0085	0.0078	0.0007	0.0142	0.0086	0.0084	0.0087	0.0078	0.0075	0.0072
Baboon (256 × 256)	0.0075	0.0064	0.0127	0.0112	0.0099	0.0096	0.0098	0.0091	0.0025	0.0089
Bird (256 × 256)	0.0089	0.0085	0.0005	0.0334	0.0098	0.0087	0.0097	0.0102	0.0103	0.0081
Flower (256 × 256)	0.0081	0.0061	0.0091	0.0107	0.0085	0.0069	0.0077	0.0115	0.0133	0.0079
Medical Image 1 (256 × 256)	0.0082	0.0081	0.0311	0.0163	0.0083	0.0088	0.0087	0.0083	0.0080	0.0083
Medical Image 2 (256 × 256)	0.0076	0.0082	0.0076	0.0149	0.0093	0.0093	0.0086	0.0078	0.0096	0.0091
Lena (512 × 512)	0.0093	0.0090	0.0154	0.0195	0.0017	0.0087	0.0090	0.0101	0.0096	0.0098
Baboon (512 × 512)	0.0091	0.0089	0.0168	0.0128	0.0016	0.0102	0.0095	0.0079	0.0096	0.0072
Tiffany (512 × 512)	0.0093	0.0095	0.1427	0.0172	0.0027	0.0093	0.0088	0.0102	0.0097	0.0099
Average values	0.0085	0.0167	0.0068	0.0263	0.0089	0.0081	0.0093	0.0085	0.0089	0.0090

(a) (b) (c)

Fig. 7 (**a**)Decrypted Lena cipher after darkening 12.5% cipher image from top-left corner (**b**) Decrypted Lena cipher after darkening 25% cipher image from top-left corner (**c**) Decrypted Lena cipher after darkening 50% cipher image from top-left corner

4.11 NIST SP 800-22 test

To validate the pseudo-random properties of the encryption algorithm, the NIST SP 800-22 test is performed on cipher Lena image (256×256). NIST SP 800-22 includes fifteen tests and is widely utilized to validate the pseudo-random properties [38]. The test uses a p-value at a level of significance β. The sequence passes the test whenever $p - value \geq \beta = 0.01$, and it is considered as random [38]. Table 11 shows the NIST test results for the obtained cipher Lena image. All the tests have been successfully passed since the $p - value \geq 0.01$ but we have excluded the linear complexity test and universal statistical test as both tests require a larger binary stream. It can be concluded that the cipher image is random.

4.12 Resistance to classic attacks

As stated in Kerckhoff's principle, the procedure for the encryption and decryption processes is usually known. Due to this, the overall focus should be on the complexity of the cryptanalysis of these methods. The attacker always tries to obtain the secret keys used for the encryption and decryption processes. If the privacy of the key is breached, then the con-

(a) (b) (c)

Fig. 8 (**a**) Decrypted Lena cipher after introducing salt and pepper noise to 12.5% cipher image (**b**) Decrypted Lena cipher after introducing salt and pepper noise to 25% cipher image (**c**) Decrypted Lena cipher after introducing salt and pepper noise to 50% cipher image

Table 11 NIST test results on cipher Lena image using the proposed approach

Test name	p-value	Result
Frequency Test	0.953748	Pass
Block Frequency Test	0.211071	Pass
Run Test	0.561917	Pass
Run Test (Longest Run of Ones)	0.718945	Pass
Binary Matrix Rank Test	0.306156	Pass
Discrete Fourier Transform (Spectral) Test	0.847187	Pass
Non-overlapping Template Matching Test	0.078790	Pass
Overlapping Template Matching Test	0.110434	Pass
Serial Test	0.766182	Pass
Approximate Entropy Test	0.700073	Pass
Cumulative Sums	0.669886	Pass
Random Excursion Test	0.786868	Pass
Random Excursion Variant Test	0.137861	Pass

fidentiality, integrity, and authenticity of the data being compromised. Typically, there are four types of attacks: known-plaintext attacks (The attacker has plaintext and corresponding ciphertext information), chosen-plaintext attack (the attacker can construct the corresponding ciphertext according to a certain plaintext), ciphertext-only attack (the attacker has only ciphertext information), and chosen-ciphertext attack (the attacker can construct the corresponding plaintext according to a ciphertext).

Initially, we use the SHA-3 algorithm on the input plain image to obtain a set of values. These values are used to modify the initial parameters of the Lorenz map. In this way, we get a distinct group of initial parameters for every distinct image. SHA-3 is highly sensitive to input and hence affects the Lorenz sequences, which are used in the permutation and diffusion processes. This implies that even a subtle change in the initial parameters will lead to a completely different cipher image. Therefore, the algorithm can resist the chosen-plaintext attack. It also resists all classical attacks as it resists chosen-plaintext attacks, which is the most powerful classical attack.

5 Conclusion

This paper proposes a novel chaotic image cryptography technique that uses Secure Hashing Algorithm 3, random graph traversal and the Lorenz map. The process of encryption involves mainly two phases, i.e., permutation and diffusion operations. Initially, the initial parameters are derived by applying the SHA-3 algorithm to image data. These initial parameters are used for generating the Lorenz sequences. Further, these random sequences are utilized for the permutation process in conjunction with random graph traversal and diffusion processes for image encryption. To validate the performance of the proposed method, standard images are used for the encryption and decryption processes. We have compared the proposed method with state-of-the-art methods using several qualitative and quantitative measures. The proposed algorithm fails to decrypt the cipher image even with the key deviation as low as 10^{-17} compared to related methods, which shows excellent key sensitivity of the proposed method.

Furthermore, it is highly sensitive to input images, such that even a slight change in the input image will result in a completely different cipher image. The cipher images have a low correlation between adjacent pixels. The algorithm is highly secure and can resist several common attacks, which makes it a suitable choice for image encryption.

Acknowledgements The authors would like to thanks Mr. Puneet Kumar Pal, Research Scholar, Department of Applied Mathematics, Delhi Technological University, Delhi, India for his constructive idea and help on addressing the reviewers' comments.

Author Contributions There are equal contributions in this research from all the authors of this article.

Funding details We have received no funding for this research.

Data availability statements The dataset used for validation and comparison of the proposed approach are 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. Chen G, Mao Y, Chui CK (2004) A symmetric image encryption scheme based on 3d chaotic cat maps. Chaos, Solitons Fractals 21(3):749–761
2. Wang X, Zhang X, Gao M, Tian Y, Wang C, Iu HHC (2023) A color image encryption algorithm based on hash table. Hilbert curve and hyper-chaotic synchronization. Mathematics 11(3):567
3. Zhang W, Wang S, Han W, Yu H, Zhu Z (2020) An image encryption algorithm based on random hamiltonian path. Entropy 22(1):73
4. Mao Y, Chen G (2005) Chaos-based image encryption. In: Handbook of geometric computing, Springer, pp 231–265
5. Biswas HR, Hasan MM, Bala SK (2018) Chaos theory and its applications in our real life.
6. Askar SS, Karawia AA, Alshamrani A (2015) Image encryption algorithm based on chaotic economic model. Math Probl Eng 2015
7. Mondal B, Kumar P, Singh S (2018) A chaotic permutation and diffusion based image encryption algorithm for secure communications. Multimed Tools Appl 77(23):31177–31198
8. Rohith S, Hari Bhat KN, 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, IEEE, pp 1–6
9. Ye G, Huang X (2016) A secure image encryption algorithm based on chaotic maps and sha-3. Secur Commun Netw 9(13):2015–2023
10. Erkan U, Toktas A, Toktas F, Alenezi F (2022) 2d eπ-map for image encryption. Inf Sci 589:770–789
11. Li L, Luo Y, Qiu S, Ouyang X, Cao L, Tang S (2022) Image encryption using chaotic map and cellular automata. Multimed Tools Appl 1–19
12. 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
13. Xian Y, Wang X (2021) Fractal sorting matrix and its application on chaotic image encryption. Inf Sci 547:1154–1169
14. Gui X, Huang J, Li L, Li S, Cao J (2022) A novel hyperchaotic image encryption algorithm with simultaneous shuffling and diffusion. Multimed Tools Appl 81(15):21975–21994
15. Zhu W, Shen Y (2010) Encryption algorithms using chaos and cat methodology. In: 2010 International conference on anti-counterfeiting, security and identification, IEEE, pp 20–23

16. Chen J, Li X-W, Wang Q-H (2019) Deep learning for improving the robustness of image encryption. IEEE Access 7:181083–181091
17. Li X, Jiang Y, Chen M (2018) Li F (2018) Research on iris image encryption based on deep learning. EURASIP J Image Video Process 1:1–10
18. Sharma K, Aggarwal A, Singhania T, Gupta D, Khanna A (2019) Hiding data in images using cryptography and deep neural network. arXiv preprint arXiv:1912.10413
19. Chai X, Chen Y, Broyde L (2017) A novel chaos-based image encryption algorithm using dna sequence operations. Opt Lasers Eng 88:197–213
20. 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
21. Musanna F, Kumar S (2020) A novel image encryption algorithm using chaotic compressive sensing and nonlinear exponential function. J Inf Secur Appl 54:102560
22. He X, Zhang Q (2008) Image encryption based on chaotic modulation of wavelet coefficients. In: 2008 Congress on image and signal processing, vol 1, pp 622–626. IEEE
23. Vaish A, Gautam S, Kumar M (2019) A wavelet based approach for simultaneous compression and encryption of fused images. J King Saud Univ-Comput Inf Sci 31(2):208–217
24. 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
25. Faridnia S, Fae'z K (2010) Image encryption through using chaotic function and graph. In: International conference on computer vision and graphics, Springer, pp 352–359
26. Xu C, Sun J, Wang C (2020) An image encryption algorithm based on random walk and hyperchaotic systems. Int J Bifurc Chaos 30(04):2050060
27. Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20(2):130–141
28. Daud NN, Hamid SHA, Saadoon M, Sahran F, Anuar NB (2020) Applications of link prediction in social networks: a review. J Netw Comput Appl 166:102716
29. Jafarzadeh N, Iranmanesh A (2013) A new graph theoretical approach to dna sequencing with nanopores. Match-Commun Math Comp Chem 70(1):401–415
30. Alvarez G, Li S (2006) Some basic cryptographic requirements for chaos-based cryptosystems. Int J Bifurc Chaos 16(08):2129–2151
31. Zhang Y (2021) Statistical test criteria for sensitivity indexes of image cryptosystems. Inf Sci 550:313–328
32. Zhang X, Wang L, Niu Y, Cui G, Geng S (2019) Image encryption algorithm based on the h-fractal and dynamic self-invertible matrix. Comput Intell Neurosci 2019
33. Tsai D-Y, Lee Y, Matsuyama E (2008) Information entropy measure for evaluation of image quality. J Digit Imaging 21(3):338–347
34. Roohi M, Zhang C, Chen Y (2020) Adaptive model-free synchronization of different fractional-order neural networks with an application in cryptography. Nonlinear Dyn 100(4):3979–4001
35. Yang Y-G, Wang B-P, Yang Y-L, Zhou Y-H, Shi W-M, Liao X (2021) Visually meaningful image encryption based on universal embedding model. Inf Sci 562:304–324
36. Wu Y, Noonan JP, Agaian S et al (2011) Npcr and uaci randomness tests for image encryption. Cyber J Multidiscip J Sci Technol J Sel Areas Telecommun (JSAT) 1(2):31–38
37. El-Shafai W, Khallaf F, El-Rabaie E-SM, 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
38. Cao L-C, Luo Y-L, Qiu S-H, Liu J-X (2015) A perturbation method to the tent map based on lyapunov exponent and its application. Chin Physics B 24(10):100501

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.

Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-023-30596-4

REVIEW ARTICLE

Study of vapor compression refrigeration system with suspended nanoparticles in the low GWP refrigerant

Uma Shankar Prasad[1,2] · **Radhey Shyam Mishra**[3] · **Ranadip Kumar Das**[1]

Received: 2 February 2023 / Accepted: 18 October 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
The rising global temperatures, attributed to the high global warming potential (GWP) of conventional refrigerants, necessitate the adoption of low-GWP alternatives in HVAC systems. However, these low-GWP refrigerants often exhibit high toxicity and flammability, limiting their usage. To address these challenges, compact heat exchangers incorporating blended refrigerants have been introduced to enhance HVAC system performance. Researchers have also made significant strides in improving HVAC system efficiency by introducing the concept of suspending nanolubricants and nanorefrigerants within the system. This review paper seeks to comprehensively assess the potential of alternative refrigerants containing suspended nanoparticles, commonly referred to as nanorefrigerants. The paper reviews various mechanisms and potential combinations of different nanorefrigerants employed to enhance refrigeration system effectiveness and efficiency. A detailed examination of key heat transfer parameters and the performance predictions of low-GWP refrigerants, including those from the hydrofluoroolefin (HFO) and hydrocarbon (HC) classes, is conducted through energy and exergy analyses. Commercial refrigerants like R-134a, R-290, R-600, R-600a, R-123, R-125, R-22, R-141b, R-152, R-11, R-113, R-404a, R-407c, R-502, R-600a, R-507a, R-1234yf, R-1234ze, 1336mzz(Z), and R-410a are evaluated in conjunction with suspended nanoparticles, considering their specific properties. The findings indicate that the utilization of nanorefrigerants leads to notable improvements in overall system performance, characterized by reduced compressor workloads and increased heat transfer rates. Consequently, the integration of blended nanoparticles into refrigerants holds significant promise for advancing the HVAC field.

Keywords VCRS · Alternative refrigerants · Eco-friendly refrigerants · Nanorefrigerant · Energy exergy analysis · Irreversibility analysis

Introduction

Refrigeration is the practice of actively regulating and attaining temperatures lower than the surrounding environment. Its primary objective is to reduce the temperature of a given item or area to reach a specific desired level. Food preservation and air-conditioning are the most important applications of refrigeration systems. Several researchers investigated and found that the efficiency can be enhanced by suspending nanoparticles and nanolubricants in pure conventional refrigerants. Nanofluids are prepared by suspending nanoparticles in a base fluid to form a colloidal mixture, generally, carbides, metals, and oxides used as nanoparticles as base fluids like water, oil, ethylene, and glycols are used. Numerous investigations have highlighted the remarkable properties of nanofluids, rendering them highly valuable across a wide spectrum of applications. These include fuel cells, heat sinks, heat exchangers, heat transfer processes, hybrid engines, pharmaceutical production, microelectronics, grinding, machining, cooling systems for diverse components, chillers, household refrigerators, solar technologies,

Responsible Editor: Philippe Garrigues

✉ Uma Shankar Prasad
dp16dp.16dp000005@mech.iitism.ac.in

Radhey Shyam Mishra
professor_rsmishra@yahoo.co.in

Ranadip Kumar Das
ranadipdas69@gmail.com

[1] IIT(ISM), Dhanbad, India

[2] PIEMR, Indore, India

[3] Delhi Technological University, New Delhi, India

Published online: 28 November 2023

and greenhouse systems. Notably, nanofluids exhibit the capacity to enhance the thermal conductivity of the underlying fluid, while also demonstrating a substantial improvement in convective heat transfer. Refrigerants are divided into two groups, i.e., primary and secondary. Primary refrigerants are chlorofluorocarbons (CFC) and hydro-chlorofluorocarbons (HCFCs), and secondary refrigerants are classified by different parameters like ozone depletion potential (ODP), global warming potential (GWP), toxicity, flammability, corrosion, and odor researchers keep on exploring. In this paper, behavior of suitable nanorefrigerants and nanolubricants is reviewed to show the new possibilities of utilization of different blended refrigerants in the field of refrigeration and HVAC systems.

VCRS setup

In a basic vapor compression refrigeration system, the essential components that work together to handle the cooling requirements are compressors, condensers, evaporators, and expansion valves. However, certain applications demand varying temperature zones within the system, such as in food processing facilities, food storage units, and hotels. The demand for different temperatures at different sections will be fulfilled by multi evaporator vapor compression refrigeration system, but multi evaporator system consumes high electricity, and it can be only reduced by enhancing the system's efficiency.

Preparation of nanorefrigerants

Nanorefrigerants are prepared by suspending small quantity of nanoparticles in a base fluid like water and ethylene glycol with or without stabilization techniques. There are two most common methods to produce nanorefrigerants:

- *Two-step method*: the two-step method is the simplest and most common method for the development of nanofluids. This process evolves with the means of physical or chemical processes nanosized dry powders which are produced of the desired material. Then, with the help of different mixing processes like homogenizing, ball milling, ultrasonic agitation, high-shear mixing, and intensive magnetic force agitation, these dry powders are diffused in a base fluid. It is the most economical method to produce nanofluids as different industries are producing nanofluids with the same synthesis technique. Aggregation of nanoparticles is the major issue in this method due to surface activity and high surface area, and due to aggregation, the nanofluids are unstable. To overcome the problem of stability in the two-step method, new and advanced techniques are introduced like one-step method. In the next section, we will discuss on one-step method with the use of surfactants in the fluids. Surfactants are active agents; when added to any fluid, they increase the spreading and wetting properties, but in high-temperature applications, the behavior of surfactants is still an issue.
- *One-step method*: Eastman et al. (1996) introduced a one-step method named physical vapor condensation process to eliminate the problem of aggregation of nanoparticles in the fluid and prepared nanofluid of copper nanoparticles suspended in the solution of ethylene glycol. The process of manufacturing and suspension of nanoparticles is done simultaneously at the same time in this method. In this method, the fluid is very stable, and agglomeration is very low due to the absence of processes like depositary, conveyance, drying, and suspension of nanoparticles in the fluid. Abdel-Hadi et al. (2011) discussed the formation of environment-friendly nanorefrigerants for vapor compression refrigeration cycles.

Nanofluids with a uniform dispersion of nanoparticles and a high level of stability can be achieved by the one-step method. For different dielectric liquids, a vacuum-based process named submerged arc nanoparticle synthesis system (SANSS) is used to prepare stable nanofluids. The thermal conductivity of the dielectric liquid can predict the behavior of the nanofluid at different operating conditions. Several investigations are done and found that due to undesired aggregation, nanoparticles create different undesirable structures of the square, polygon, circular, and needle-like structures. The one-step method is the most suitable method to eliminate the problem of aggregation. The one-step chemical method is the most useful method due to its cheap operation and large-scale production as compared to the one-step physical method.

Proposed system explanation

The schematic circuit diagram in Fig. 1 of the proposed model for investigation is shown, and it is clearly visible that in the evaporator, the input variables are inlet brine

Fig. 1 Photograph of single-stage VCR setup

temperature, the water temperature of condenser inlet, compressor speed in RPM, type of refrigerant, the mass flow rate of fluid, and water and structure parameters of condenser, and evaporator is used to calculate the efficiency (COP) of the VCRS system without using nanorefrigerants. After the introduction of the nanorefrigerant in the system, we need to consider a few additional parameters like nanoparticle diameter, type of nanoparticle, and thermo-physical properties of nanoparticles. The model we proposed is a chiller machine based on a simple VCRS system in which the evaporation and condensation process is performed under two copper (concentric) tubes. The refrigerant will flow under the inner tube of the evaporator, and the brine will flow in the annuli side; similarly, the water is allowed to flow on the inner side, and the refrigerant will flow in the annuli side of the condenser as shown in Fig. 2.

Literature review

Many researchers have investigated the effects of nanoparticles suspended in low GWP refrigerants to improve efficiency, and in Table 1, their concept, techniques, and observation are discussed.

Maxwell (1873) discussed the history of refrigeration with the basic concept of modern refrigeration system. Hamilton and Crosser (1962) discussed on pure material, shape of particle, and mixture composition affects the thermal conductivity of the nanorefrigerant. Spaushus (1988) discussed about the commercial applications of new refrigerant CF-134a with high stability, solubility, and lubricity as a replacement of CFC 12. Liu et al. (1994) performed test result of hydro carbon mixtures shows the energy saving upto17.3% in domestic refrigerators. The potential benefit of Cu and Al_2O_3 is estimated and observed by Choi (1995) with reduced pumping power. Richardson and Butterworth et al. (1995) used R134a mixture of propane/isobutene (1:1 ratio) is used and found better COP than R12 but highly flammable. Nanoparticle-infused fluids exhibit a threefold increase in thermal conductivity when compared to conventional fluids in identical heat transfer equipment, resulting in double the rate of heat transfer. An experimental investigation was performed and found by Eastman (1996) that the thermal conductivity of the fluid (HE-200) is increased by up to 60% with the addition of 5% of the volume of Cu (35nm) nanoparticles resulting in improvement in the performance of VEROS-type evaporation system. Eastman et al. (1996) Thermal conductivity of fluid (HE-200) is increased up to 60% with addition 5% of volume of Cu (35nm). Havelsky et al. (2000) Results show that the COP is increased by using R-134a, R-401A and R-409A refrigerants instead of R-12 it also reduces the value of TEWI. The effect of the structural character of a gold nanoparticle of size 24 nm suspended in the nanofluid of tannic acid; Tri-sodium citrate over the heat pipe was performed by Ding (2007) and found that thermal resistance is reduced from 7 to 20%. Choi (1995) prepared an iron nanofluid with ethylene glycol and passed the solution through a chemical vapor condensation process and sonication with high-powdered pulses to improve the dispersion of iron particles in the fluid.

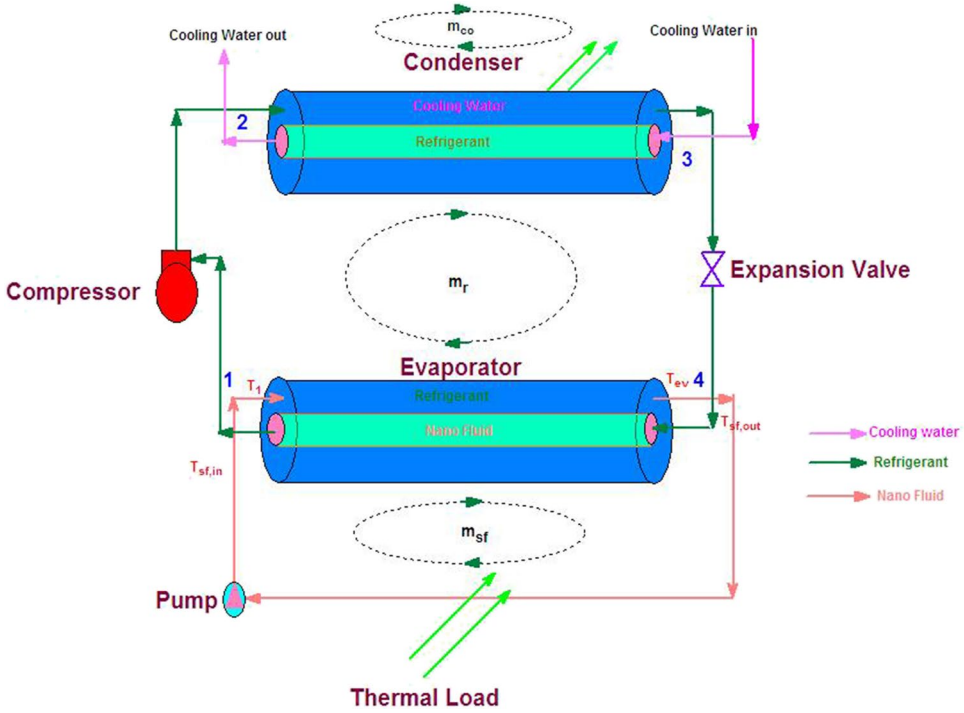


Fig. 2 Schematic structure of proposed model (adapted from Mishra 2014)

Table 1 Comparison table for concept and technique with observation

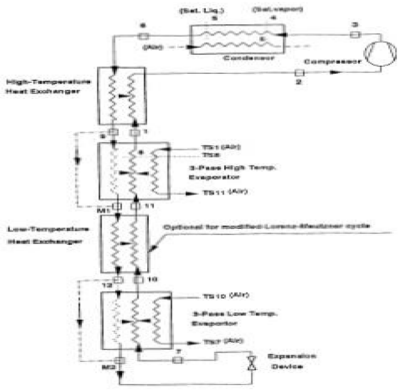
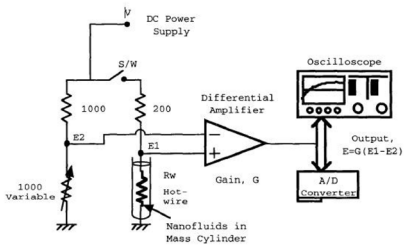
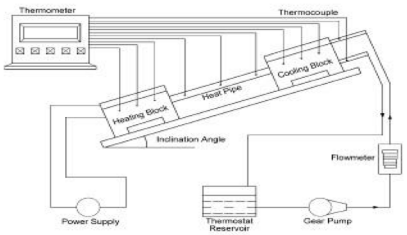
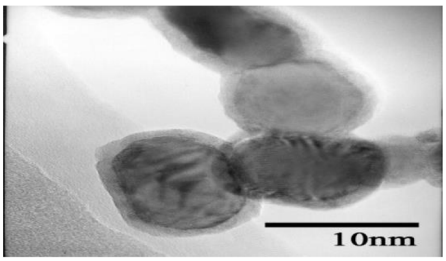
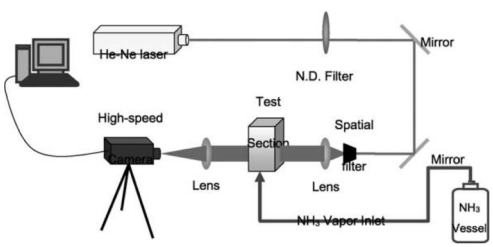
Year	Investigators	Refrigerant; Nano-particle; Avg. size	Concepts and techniques	Observation
1994	[99]	R290/R600/n-c5		Hydrocarbon mixture can save energy up to17.3% in domestic refrigerators.
1996	[18]	HE-200 Cu 35 nm		Thermal conductivity of fluid (HE-200) is increased upto 60% with addition 5% of volume of Cu (35nm)
2004	[93]	Tannic acid; Trisodium citrate; Gold nanoparticles; Au 24 nm		Thermal resistance is reduced from 7 to 20%
2005	[26]	Ethylene glycol; Fe 10 nm		Results shows 0.55 volume fraction of iron in nanofluid exhibits 18% increment in thermal conductivity.
2006	[41]	Cu; CuO; Al_2O_3		Results shows that adding surfactants and nanoparticles together can improve the absorption performance by 5.32 times.

Table 1 (continued)

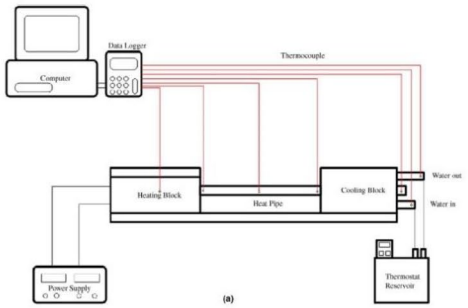
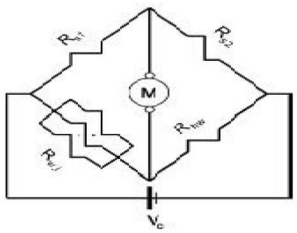
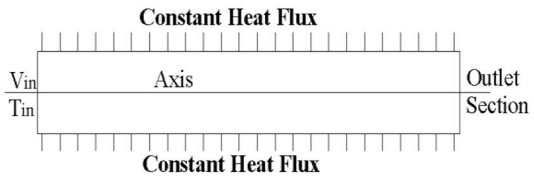
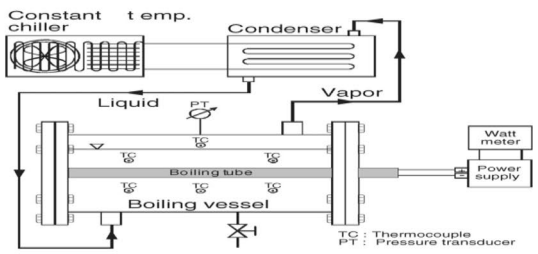
2006	[38]	Ag (Silver) 35 nm		Silver nanoparticles suspended in DI-water can reduce thermal resistance by 50% and 80% respectively.
2006	[43]	CuO; SiO_2;		Higher surface charge state can affect the suspension stability and improve the desirable thermal conductivity of nanofluids.
2006	[45]	CuO/ethylene glycol; 30–50 nm diameter		The hot wire method is employed to measure thermal conductivity. At a volume fraction of 0.05 (equivalent to 5%), there is a notable improvement of 22.4% in thermal conductivity.
2007	[42]	Al_2O_3;		Maximum of 2.25 times heat transfer enhancement and 1.42 times of friction is obtained by using nanofluid as working medium.
2007	[73],[74]	R123; R134; CNT R22; H_2O; CNT		Heat transfer coefficients of R134 & R123 can be improved up to 36.6% by use of CNTs also can improve the telmperature of water rise up to 28.7%
2008	[9]	R134a; TiO_2		Compared to the HFC134a and POE oil system, it reduces energy consumption by a significant margin of 26.1%.

Table 1 (continued)

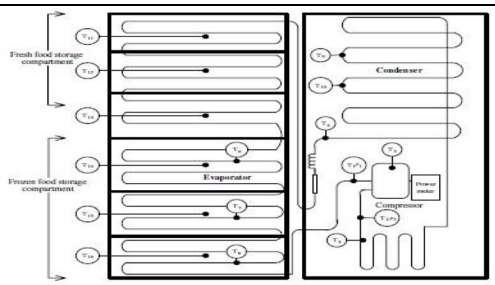
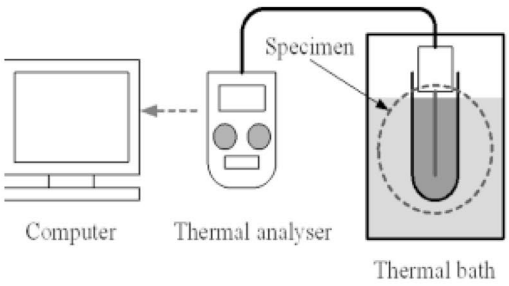
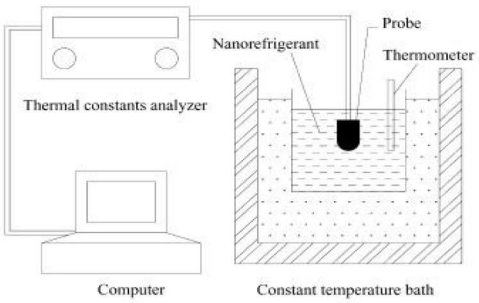
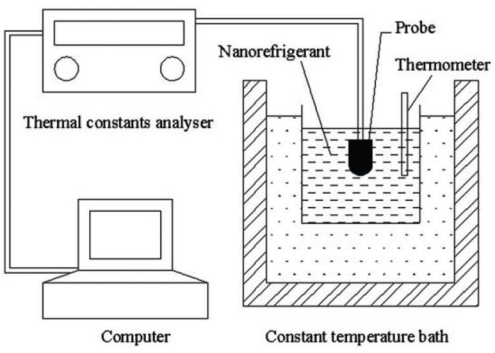
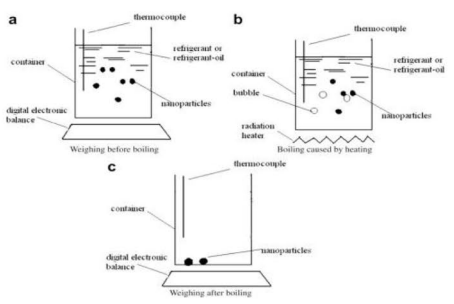
				
2008	[13]	R134a; Al_2O_3		Thermal conductivity is improved by 2%, 4.6%, and 2.5% by use of Al_2O_3 nanoparticles by 1, 1.5, and 2 wt.% at 40 °C and reported better thermal conductivity at higher temperatures
2009	[95]	R113; CNT		Thermal conductivity can be influenced by both the diameter and aspect ratio of carbon nanotubes (CNTs). Specifically, when the CNT diameter is smaller, it results in a more significant enhancement in thermal conductivity.
2009	[34]	R113; Cu; Al; Ni; CuO; Al_2O_3		The rise in thermal conductivity is pronounced as the nanoparticle volume fraction increases. Additionally, the thermal conductivity is notably influenced by the aggregation and formation of nanoparticle clusters.
2009a	[16]	R-113; CuO		Thermal conductivity of the nanofluid can maximum enhance up to 29.7% by a disturbance in nanoparticles due to reduction in boundary layer height.

Table 1 (continued)

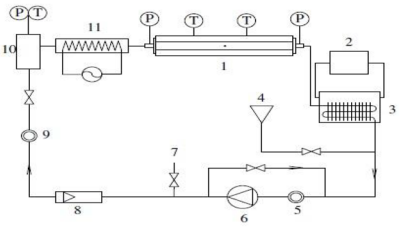
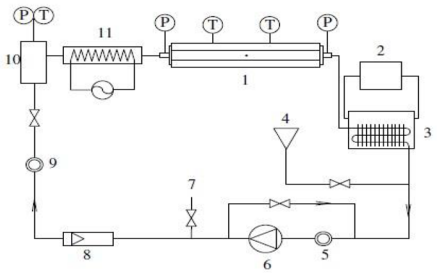
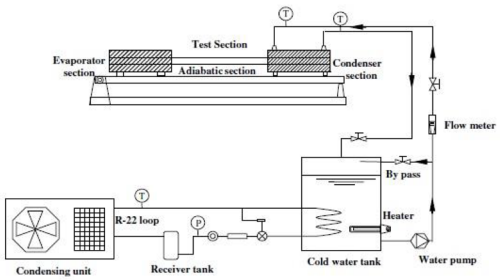
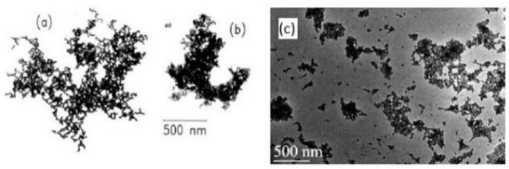
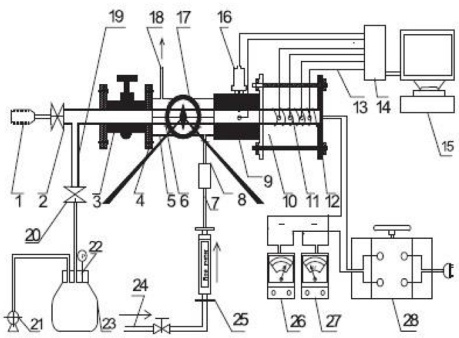
2009	[78]	R-113; CuO 40 nm diameter		Suspending nanoparticles can lead to a substantial increase in the heat transfer coefficient, with a maximum enhancement of up to 29.7%.
2009	[36]	HR-12 POE; HR-12 MO; Al_2O_3(0.05– 0.2 wt %)		COP has improved by 4.4 % also power consumption is reduced by 2.4%. They performed TEM and XRD spectroscopy for the analysis of Al_2O_3 nanoparticles.
2009	[72]	R11; Titanium (Ti) 21 nm diameter		The inclusion of nanoparticles at a concentration of 0.1% in the refrigerant can enhance performance by up to 1.40 times when compared to using the pure refrigerant.
2010	[97]	SiO_2; Al_2O_3 H_2O		Clustering of suspended nano-particles of Al_2O_3 gives best result for improvement in thermal conductivity.
2010	[48]	CuO nanoparticles 50 nm diameter		The highest heat flux is observed when the mass concentration reaches 1.0 wt%. Furthermore, the heat transfer coefficients are influenced by the operating pressure of the system; specifically, at lower pressures, there is an increase in the heat transfer coefficient.
2010	[10]	R-134a TiO_2 21 nm diameter		Nano-oil properties can change the tribological behaviour of base lubricants at different temperature and depends upon the nanomaterial.

Table 1 (continued)

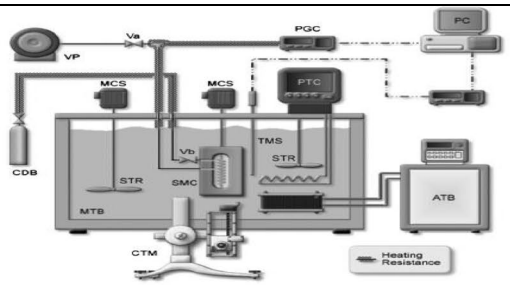
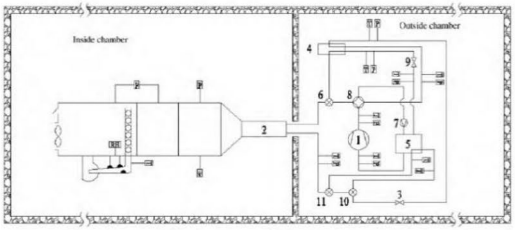
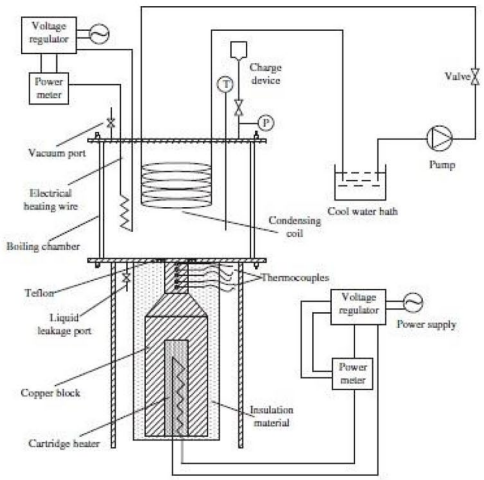
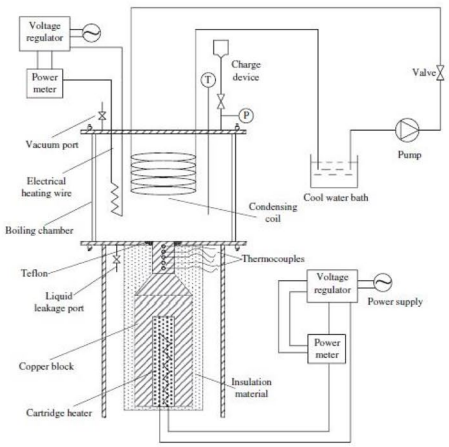
				
2010	[94]	R410a $NiFe_2O_4$	 1---- compressor; 2---- inside heat-exchanger; 3---- main expansion valves; 4---- outside heat-exchanger; 5---- vapor/liquid separator; 6---- three-way valve; 7---- check valve; 8---- four-way valve; 9---- auxiliary expansion valves; 10----three-way valve; 11----three-way valve; RH----relative humidity indicator (Wet/Dry-bulb thermometer); P---- pressure transducer; $\bar{v}$ ---- the air velocity transducer, T---- temperature transducer.	Naphthene-based mineral nano-refrigeration oil can improve the energy efficiency ratio by upto 6%.
2010	[75]	R113 CNTs nanolubricant15 nm to 80 nm		The addition of CNT to the R-113 oil mixture can enhance the nucleate pool boiling heat transfer coefficient by as much as 61%.
2010	[77]	R113 Diamond nanoparticles		Elevating the concentration of nanoparticles leads to an increase in the nucleate pool boiling heat transfer coefficient. Moreover, when diamond nanoparticles are introduced into the oil and R113 mixture, the maximum enhancement is observed at approximately 63.4%.

Table 1 (continued)

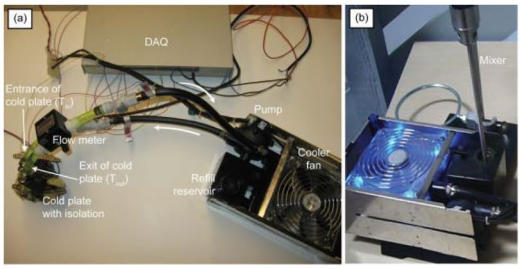
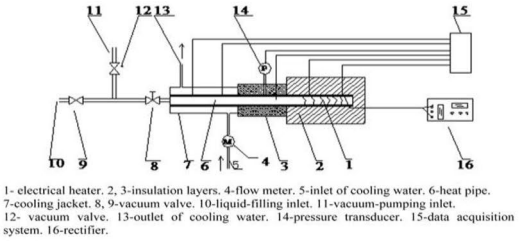
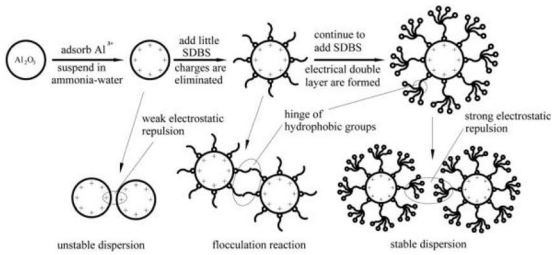
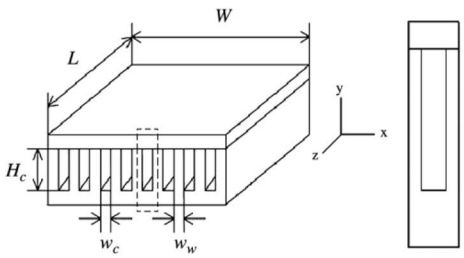
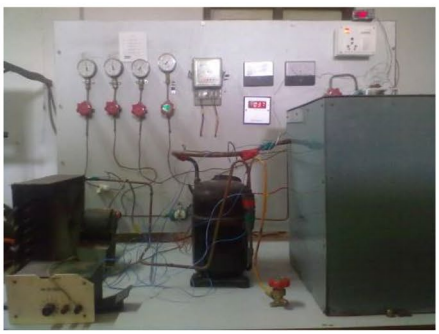
2011	[85]	Oleic acid and CTAB surfactants Cu nano powder of 1 µm		CTAB surfactant based 1.0% Cu nano powder mixed in the fluid can increase 48% of thermal conductivity.
2011	[47]	Cu; 40 nm and 20 nm CuO; 50 nm and 20 nm SiO; 30 nm	 1- electrical heater. 2, 3-insulation layers. 4-flow meter. 5-inlet of cooling water. 6-heat pipe. 7-cooling jacket. 8, 9-vacuum valve. 10-liquid-filling inlet. 11-vacuum-pumping inlet. 12- vacuum valve. 13-outlet of cooling water. 14-pressure transducer. 15-data acquisition system. 16-rectifier.	Experiment shows that heat transfer in Cu nanoparticles are better than CuO nanoparticles, thermal performance was decreased by SiO nanofluid.
2011	[46]	Carbon black nanoparticles ;OP -10 emulsifier; Al_2O_3 20 nm with SDBS		Carbon nano fluids are stable with increase in mass fraction of the surfactant, Al_2O_3 nanofluid is disturbed initially then stable.
2011	[22]	amorphous carbonic nanofluids;20 nm Al_2O_3; 20–50 nm		Results shows that nano fluid as a coolant can enhance the pressure drop and heat transfer coefficient significantly.
2011	[70]	R-134a Al_2O_3 ; diameter<50nm		Addition of nanoparticles in the refrigerants can reduce the power consumption up to 25 % , and COP is improved by 33%.
2011	[7]	R600a TiO_2; 50 nm		TiO_2 mixed with R600a can save energy consumption up to 5.94% and works normal, safe and efficient.

Table 1 (continued)

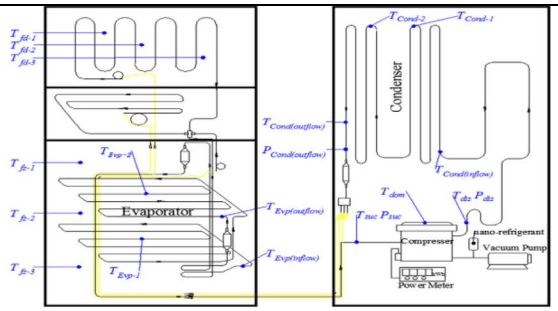
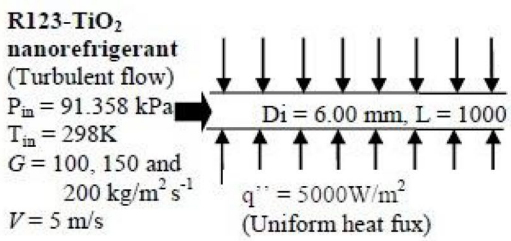
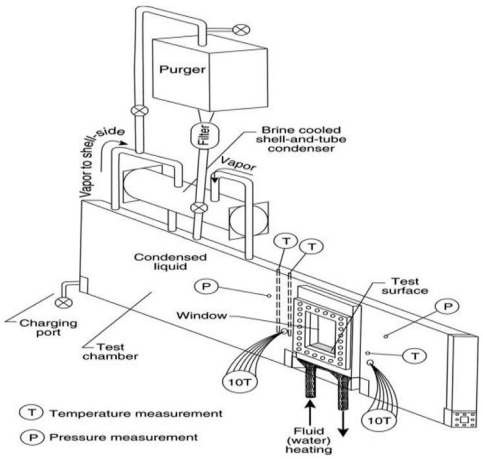
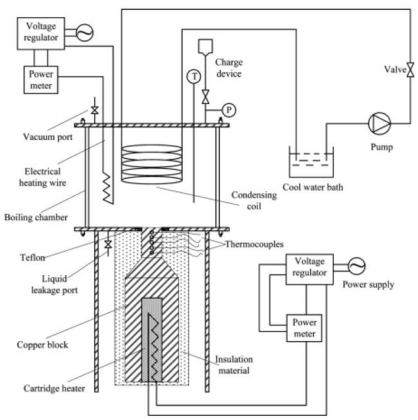
				
2011	[28]	R123 TiO_2; 21 nm		They found that volume fractions have effect on pressure drop, pressure gradient increases with increase in particle volume fraction.
2011	[39]	R134a/polyolester mixtures; Al_2O_3; 50 nm		Using various mass fractions of lubricants, such as 0.5%, 1%, and 2%, results in average heat flux improvements of 105%, 49%, and 155%, respectively.
2011	[76]	R113 & Ester oil VG68 Cu;20, 50 and 80 nm		Reducing the particle size from 80 nm to 20 nm maximizes the enhancement of the nucleate pool boiling heat transfer coefficient, resulting in a 23.8% increase.
2012	[49]	R-123 TiO_2		As the particle volume fraction rises, viscosity increases, but it decreases with an increase in

Table 1 (continued)

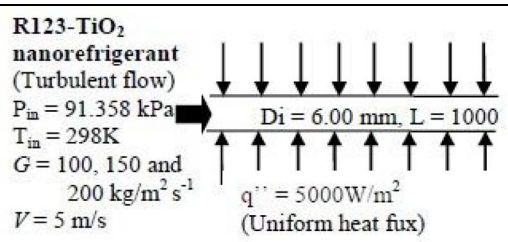
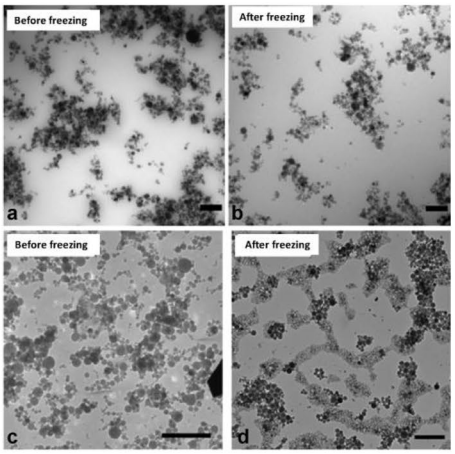
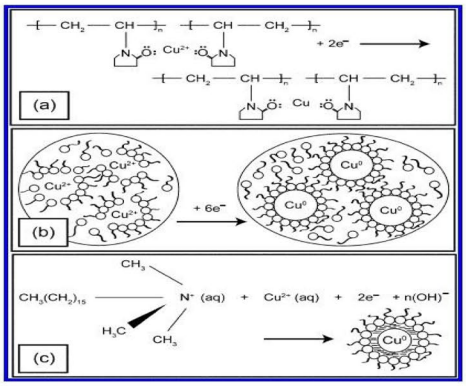
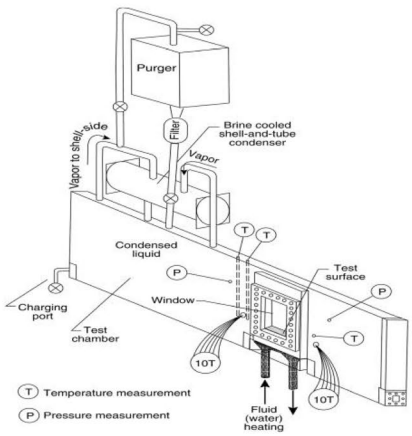
				temperature.
2012	[32]	DI-water (De-ionized) $C_2H_6O_2$ PAO lubricant Graphite suspensions		Developed a method to improve the thermal conductivity of oil and water with use of graphite flakes as an additive for suspension
2012	[53]	Cu; 80,100 nm Polyvinylpyrrolidone (PVP) Cetyl trimethyl ammonium bromide (CTAB)		Copper powder of 0.55 vol% can improve the thermal conductivity by 22% and 48% improvement by 1.0 vol % over water.
2012	[40]	R134a; RL68H2C; Diamond;10nm and 250µm		The utilization of a nano-lubricant with a mass fraction of 0.5% can enhance the heat transfer rate by as much as 98%.
2012	[29]	R-123; TiO_2;		The viscosity of nanofluids rises with an increase in the

Table 1 (continued)

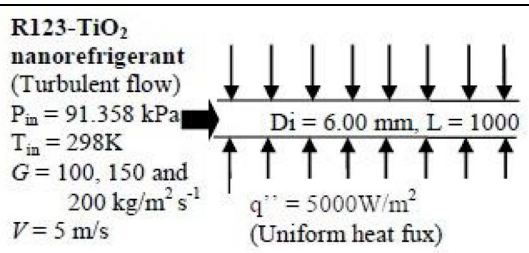
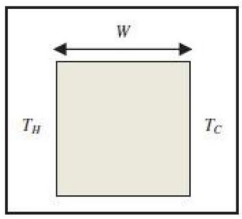
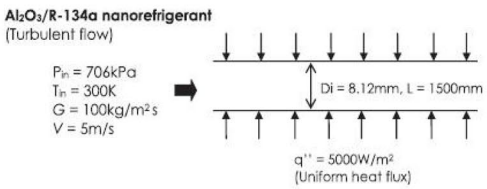
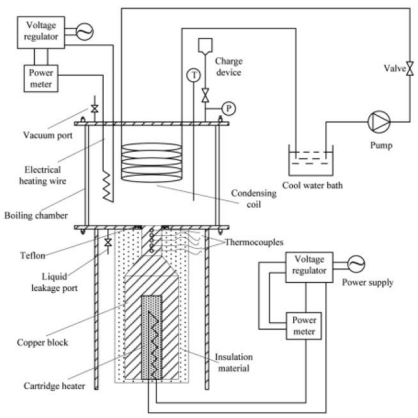
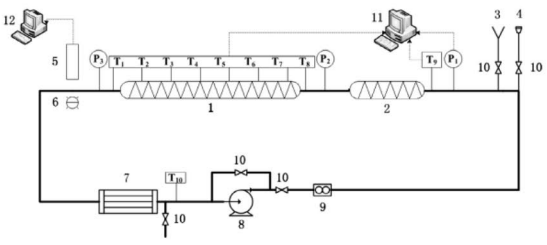
		21 nm		volume fraction of particles measuring 3 to 10 nanometers, but it diminishes with a rise in temperature.
2013	[23]	Al_2O_3–water nanofluid;		The convective heat transfer rate experiences enhancement within the range of 1% to 3% of the solid volume fraction.
2013	[30]	R-134a; Al_2O_3; 30 nm;		Al_2O_3 nanoparticles of concentration from 1% to 5% of volume in R-134a refrigerant can improve the thermal conductivity by increasing the volume fraction and temperature
2013	[27]	R113; SDS; CTAB; Span-80; Cu; 20 nm;		The surfactant impact factor (SIF) increases with increase of surfactant concentration, due to increase of surfactant concentration. SIF also increases and reach the maximum value.
2013	[90]	R141b; Cu; Al; Al_2O_3; CuO;		As the quality of nano-refrigerants increases, the heat transfer coefficient also experiences a significant rise, reaching its maximum value in the range of x = 0.5 to 0.8, respectively.
2013	[89]	R141b; Cu; Al; Al_2O_3; CuO;		Cu nanoparticles are best among others to enhance the heat transfer performance of nano-refrigerants.

Table 1 (continued)

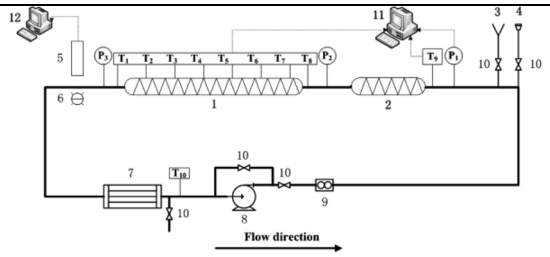
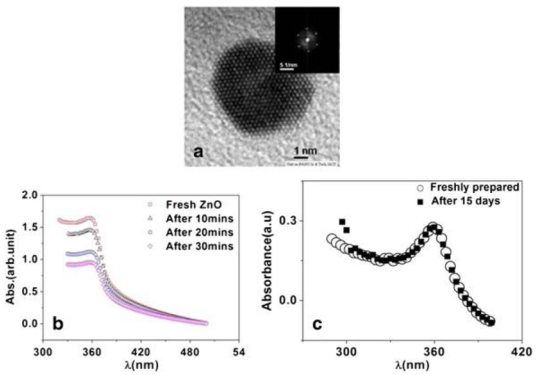
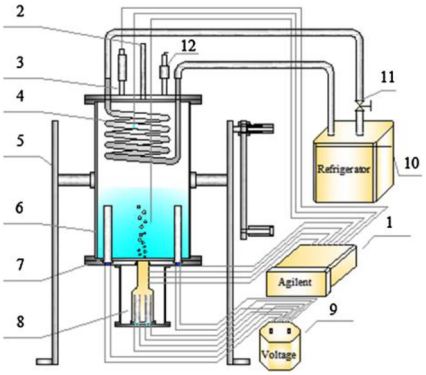
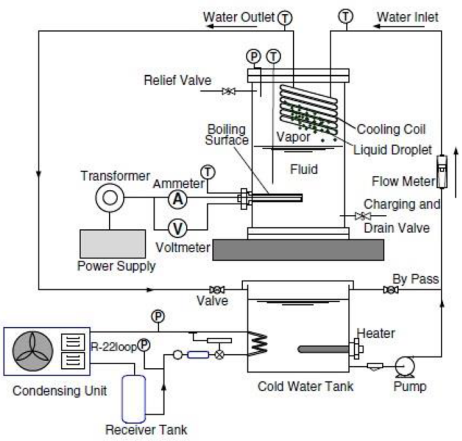
				
2013	[81]	ZnO; 10 nm Polyvinylpyrrolidone (PVP)		Addition of stabilizer local aggregation increased and reduced the thermal parameters results in decrease in thermal transport properties to control the nanofluids.
2014	[92]	R141b; Al_2O_3; 20 nm		When the nanoparticle concentration is at 0.001 vol%, the use of SDBS surfactant is unnecessary, as it actually decreases performance. However, at concentrations of 0.01 vol% and 0.1 vol%, the addition of SDBS enhances performance.
2014	[71]	R141b; TiO_2; 21 nm		At elevated heat flux levels, an increase in particle volume concentration leads to a reduction in the boiling heat transfer coefficient. Conversely, at higher boiling pressures, the boiling heat transfer coefficient increases.
2014	[5]	R-600a; CuO; 50 nm		Maximum heat transfer is increased up to 63% with the use of CuO nanoparticles in R-600a/oil.

Table 1 (continued)

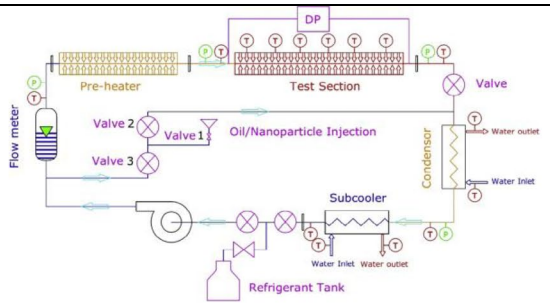
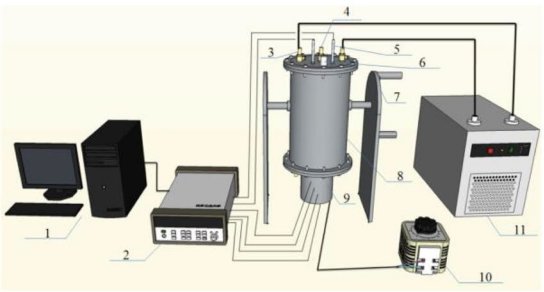
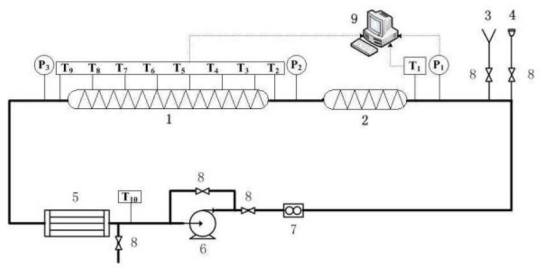
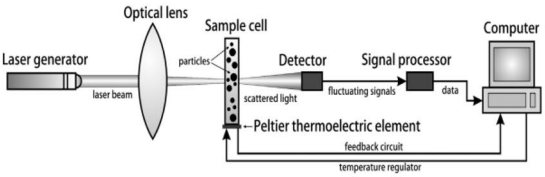
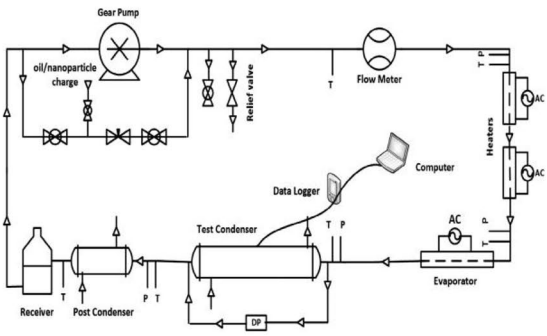
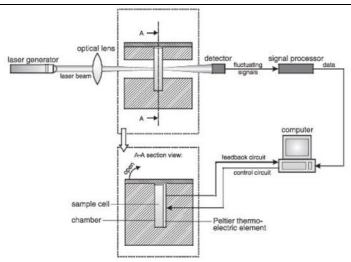
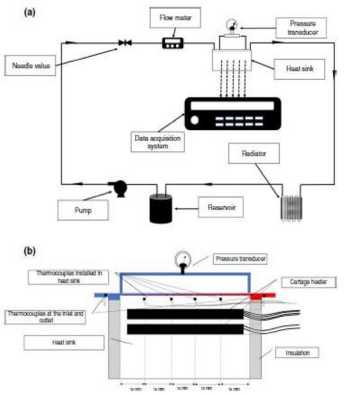
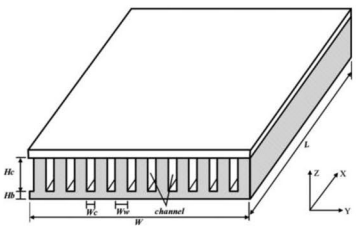
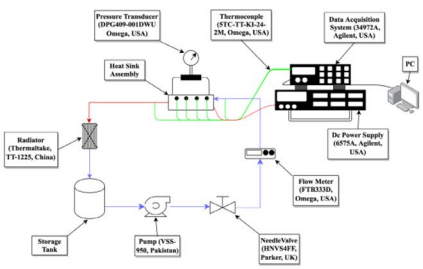
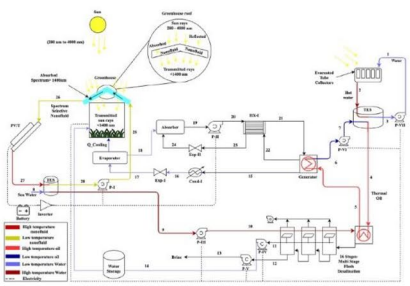
				
2015	[15]	R141b; Cu; 30 nm SDBS (Sodium dodecyl benzene sulfonate)		Nano-refrigerants containing varying concentrations of 0.008 vol%, 0.015 vol%, and 0.05 vol% when combined with pure R141b can enhance pool boiling heat transfer.
2015	[98]	MWCNT-R141b; Span-80; 20 nm		MWCNT nano-refrigerant R141b with 0.3 wt% had improvement in optimal heat transfer as compared to pure refrigerants.
2015	[79]	R141b; TiO_2; 25nm 40nm 60nm 100nm		By increasing the primary particle size from 25 to 100 nm, the steady-state hydrodynamic diameter exhibited a substantial increase of 127.6%.
2015	[6]	R600a; CuO ; 50nm		1.5 mass fraction of Cu nanoparticle can improve the heat transfer up to 83%.
2016	[44]	R141b; ATMOS NM56; TiO_2; 25 to 100 nm		Lubricating oil ATMOS NM56 present in nano-refrigerant reduces the accumulation of TiO_2 nanoparticles by increasing the particle concentration,

Table 1 (continued)

				primary particle size and temperature boost the accumulation of nanoparticles.
2019	[67]	TiO_2–water nanofluid; 0.006, 0.008, 0.01 and 0.012 vol%		The maximum improvement in Nusselt number notes is 40.57% as compared with distilled water.
2020	[50]	Cu; Al_2O_3; SiC-W; TiO_2		Alumina, titania and copper are used for nanofluid can improve the heat transfer characteristics
2020	[86]	MgO ; 20 nm		MgO nanoparticles can improve the rate of heat transfer and Exergy efficiency is high with nanofluid and can be improved by increasing Reynold's number and concentration of nanoparticles.
2020	[69]	CuO (with 0.0005 vol% concentration)		Use of nanofluid CuO (with 0.0005 vol% concentration) in the PV/T system for electricity generation in which visible spectrum of light is used in photosynthesis of the plants, thermal energy of the of PV/T system

The results show that 0.55 volume fraction of iron in nanofluid exhibits an 18% increment in thermal conductivity. Dossat and Horan (2002) discussed the effect of orientation of refrigeration cycle with the help of different equipment's, operating conditions, and attached components. Gigiel (2004) discussed on circuit protection from damage of the refrigeration system through different testing methods. Tsai et al. (2004) found that thermal resistance is reduced from 7 to 20% of in the evaporator and condenser. In He et al. (2005), HFC152a/HFC125 was proposed as alternative to replace CFC12. In Hong et al. (2005), iron with 0.55 volume fraction in nanofluid exhibits 18% increment in thermal conductivity. Fatouh and El Kafafy (2006) have taken mixture of propane (60%), isobutane, and n-butane and found that it can easily be replaced by R-134 with same performance. In the analysis in Brown et al. (2002), it was found that R134a have higher COP than CO affected by speed of compressor and ambient temperature. Joseph Sekher (2004) reported consumption reduced by 4 to 11% and actual COP improved by 3 to 8%. Sekhar and Lal (2005) recorded improvement in performance of 10–30% and energy consumption reduced to 5–15%. Hozod (2005) results show low temperature in evaporator, low temperature in discharge from compressor, Low discharge to suction pressure ratio, and low COP with high volumetric efficiency of the compressor respectively for cascade systems. Wongwises and Chimres (2005) found that Propane/butane 60%/40% mixture consumes 0.69% amount of energy to run the system as compared to HFC-134a. The effect of nanoparticles on bubble-type absorption systems was experimentally examined by Kang et al. (2006) to design a highly compact absorber for an aqua-ammonia absorption system. The combination of surfactants and nanoparticles yields a remarkable 5.32-fold enhancement in absorption performance, as demonstrated by the results. They also investigated that a high amount of silver-nanoparticle suspended in the fluid can increase the heat pipe wall temperature. Lee and Kim et al. (2006) found that higher surface charge state can affect the suspension stability and improve the desirable thermal conductivity of nanofluids. Misheck et al. (2006) compared the performance of Heat pipes on the basis of course pore size, fine pore size, composite of course and fine pore and, different wick structures. A surface complexation model to measure the experimental data of hydrodynamic size, potential, and thermal conductivity by Lee et al. (2006) was applied in the investigation and found that the pH value of colloidal fluid strongly affects the performance of the thermal fluid, and higher surface charge can affect suspension stability and improve the desirable thermal conductivity of nanofluids. CuO nanofluid was prepared by Liu et al. (2006) through a two-step method without any surfactant suspended in ethylene glycol liquid passed through XRD to study the crystalline phase to study the microstructure scanning electron microscopy (SEM) used. The hot wire method is employed to measure thermal conductivity, and it exhibits a notable 22.4% improvement at a volume fraction of 0.05 (equivalent to 5 vol.-%). Bi et al. (2007) shows experimental results that mixture of R-134a, mineral oil and TiO_2 nanoparticles performed better than the R-134a mixed polyester system, and with 0.06% mass concentration saves 21.2% energy. Ding et al. (2007) complete summary is prepared with the help of different simulation models for components of VCRS system like evaporator, condenser, compressor, capillary tube. The flow heat transfer coefficient and friction factor of Al_2O_3 nanoparticles dispersed in water and ethylene glycol in a circular tube with the help of CFD software FLUENT 6.0 was studied by Sundar et al. (2007). Park and Jung (2007a) Nucleate boiling heat transfer coefficients of R134 & R123 can rise up to 36.6% by use of CNTs. Park and Jung et al. (2007b) Boiling heat transfer coefficients of R22 and water can rise up to 28.7% by use of CNTs Syam Sundar et al. (2007) investigated results of hybrid nanofluids shows maximum effectiveness and maximum rate of heat transfer at volume concentration of 2 % for TiO_2 and for ZnO with 1.5% respectively. The analysis covers a volume concentration range from 0.1 to 4%. Utilizing nanofluid as a working medium results in a maximum heat transfer enhancement of 2.25 times and a friction increase of 1.42 times. Experimental results of Park and Jung (2007a) show that nucleate boiling heat transfer coefficients of R134 and R123 can be improved up to 36.6% by the use of CNTs without fouling and can reduce energy consumption resulting in low greenhouse warming. The effect of carbon nanotubes on boiling heat transfer coefficients of R22 was investigated by Park and Jung (2007a) who found that the addition of 1.0% of CNT can improve the temperature of water rise to 28.7% without fouling at the surface. A performance and reliability study of a domestic refrigerator with TiO_2 nanoparticles suspended in the working fluid R134 was performed by Bi et al. (2008), and it was found that it saves energy consumption by up to 26.1% as compared to HFC134a and POE oil systems. Bi et al. (2008) R134 mixed with TiO_2 nano particles as lubricant can saves energy consumption upto 26.1%. Jwo (2008) performed experiments and concluded that Al_2O_3 nanoparticles can be suspended directly in the lubricants, and they can improve thermal conductivity by 2%, 4.6%, and 2.5% by use of Al_2O_3 nanoparticles by 1, 1.5, and 2 wt% at 40 °C and reported better thermal conductivity at higher temperatures. Carbon nanotube nanorefrigerants with R113 as host refrigerant was investigated by Ding et al. (2009) and found that thermal conductivity can be increased by increasing the volume fraction of CNT. They also noted that both the diameter and aspect ratio of CNTs can influence thermal conductivity. Specifically, a smaller CNT diameter leads to a greater enhancement in thermal conductivity. Jiang et al. (2009b) did experimental and model investigations on nanorefrigerants to enhance the thermal conductivity; they found that thermal conductivity increases sharply with the

increase of nanoparticle volume fraction, and thermal conductivity is also highly affected by aggregation and forming of nanoparticle clusters. Jwo et al. (2009) fount that COP of the system is improved by 4.4 % by replacement of R-134a into hydrocarbon refrigerant with suspended Al_2O_3 nanoparticles in mineral lubricant Ding et al. (2009) studied the migration characteristics of nanoparticles in the pool boiling process through experimental and simulation results. By conducting a comparative analysis of experimental data and predicted outcomes, they conclude that the thermal conductivity of the nanofluid can be maximally enhanced by up to 29.7% as a result of nanoparticle disturbance, which reduces the boundary layer height. Ding et al. (2009) investigated the effect of copper nanoparticles suspended in R113 refrigerant on heat transfer characteristics in a smooth horizontal tube; due to the reduction of boundary layer height and disturbance of nanoparticles, they create a protective layer over the nanoparticles resulting in a maximum enhancement in the heat transfer coefficient up to 29.7%. Jiang et al. (2009a) found thermal conductivity of nano refrigerants are highly affected by aggregation & forming of nanoparticle clusters. The optimal mixture, consisting of 60% R-134a and 0.1 wt% Al_2O_3, has been determined. This blend results in a 4.4% improvement in the system COP (Coefficient of Performance) and a 2.4% reduction in power consumption. They performed TEM and XRD spectroscopy for the analysis of Al_2O_3 nanoparticles. Mohanraj et al (2009a) performed experiments and results show that HCM improves efficiency up to 25% by increasing the capillary tube length, consumes 11.1% lesser energy, domestic refrigerator COP can improve by 3.6%. Overall it is better alternative for R-134a. Mohanraj et al. (2009b) found that refrigerators using HC mixtures and R152a are better options to replace R134a and R12. For air conditioning R22 can be replaced by R290, R1270, R290, R152a, R744 For Chiller units R123 can be replaced by R11, R12 and R22. For automobile ACs HC mixtures and R152a are found as best alternatives. An investigation on a straight copper tube heat pipe to improve the heat pipe efficiency by using R11 refrigerant as a base and suspended titanium nanoparticles was performed by Naphon et al. (2009). The findings indicate that the highest heat pipe efficiency is attained when the tilt angle is set at 600, and when 50% of the charge consists of pure refrigerant with a nanoparticle concentration of 0.1%, the performance is enhanced by up to 1.40 times compared to using pure refrigerant alone. Peng et al. (2009) Thermal conductivity of the nanofluid can maximum enhance upto 29.7% by disturbance in nanoparticles due to reduction in boundary layer height. Additionally, they developed a set of parametric models to forecast thermal conductivity and determine the optimal fractional dimension of nanoparticles that yield the maximum clustering effect. The effect of CuO-based nanofluids through an experimental setup of a horizontal mesh heat pipe was done by Liu and Zhu (2011) who found that heat transfer coefficients of condenser and evaporator are improved by the use of CuO nanoparticles along with deionized water in the system. The highest heat flux is achieved at a mass concentration of 1.0 wt%. Heat transfer coefficients are influenced by the operating pressure of the system; specifically, at lower pressures, the heat transfer coefficient experiences an increase. An experimental investigation was performed by Bobbo et al. (2010) on R-134a refrigerant mixed with POE oil and TiO_2 suspended nanoparticles in single-wall carbon nanohorns (SWCNH). Nanooil properties can change the tribological behavior of base lubricants at different temperatures and depending upon the nanomaterial. When nanoparticles are suspended in the base oil, they do not change their solubility. Bolaji (2010) R152 can improve the COP upto 4.7 % as compared with R134a, Using R32 gives COP reduction of 8.5% as compared to R134a. Dalkilic and Wongwises (2010) R12 & R22 can replaced by mixed blends of HC290 & HC600a (40 & 60 by wt.%) and HC290 & HC1270 (20 & 80 by wt.%) ratio. The performance and reliability of a retrofitted residential air conditioner (RAC) were experimentally investigated by Wang et al. (2010) with a mixture of R410a and mineral-based nanorefrigeration oil (MNRO) with $NiFe_2O_4$ nanoparticle. They replaced Polyol-Easter oil (VG32) lubricant with naphthenic-based mineral nanorefrigeration oil and found an energy efficiency ratio improvement of up to 6%. The effect of carbon nanotubes on nucleate pool boiling heat transfer of refrigerant oil mixture was experimentally investigated by Peng (2010a) in R-113 oil mixture with CNT which can improve the nucleate pool boiling heat transfer coefficient by up to 61%; this can be also enhanced by reducing the outside diameter of the CNT tube or increasing the length respectively. They also established a correlation between the mixture of refrigerant and CNTs to predict the nucleate pool boiling heat transfer coefficient. The test results exhibited a strong agreement of 96%, with only a 10% deviation. An experimental investigation was done to study the nucleate pool boiling heat transfer characteristics of refrigerant/oil mixture with diamond nanoparticles (Peng, 2010b). Test results show that an increase in the concentration of nanoparticle results increase in nucleate pool boiling heat transfer coefficient, which also increases with diamond nanoparticles in the mixture of oil and R113; the maximum enhancement was found to be around 63.4%. They have additionally put forth a correlation for predicting the nucleate pool boiling heat transfer coefficient in refrigerant/oil mixtures containing diamond nanoparticles. To find out the importance of particle size distribution on copper-based nanofluid. Wu et al. (2010) Experiments done & found that clustering of suspended nano-particles of Al_2O_3 gives best result for improvement in thermal conductivity. They used two surfactants oleic acid and CTAB for the preparation of copper nanofluids by chemical reduction process; water serves as the base fluid for copper nanoparticle concentrations of 0.55 and 1.0 vol%. To achieve homogeneous mixing, sodium

dodecylbenzene sulfonate (SDBS) was employed, followed by an ultrasonication process. A dynamic light scattering process was used for stable particle size distribution. Results show that CTAB surfactant-based 1.0% Cu nanopowder mixed in the fluid can increase by 48% of thermal conductivity. Liu and Zhu (2010) performed experiments on cylindrical micro-grooved heat pipe with water base nanofluid solution using Cu and CuO nanoparticles, respectively, for different diameters. Five distinct nanofluids are formulated by utilizing copper (Cu) nanoparticles with average diameters of 40 nm and 20 nm, copper oxide (CuO) nanoparticles with average diameters of 50 nm and 20 nm, and silicon dioxide (SiO) nanoparticles with an average diameter of 30 nm. The experiment shows that heat transfer in Cu nanoparticles is better than in CuO nanoparticles, and thermal performance was decreased by SiO nanofluid. Mahlia and Saidur (2010) found that energy consumption of air conditioners and refrigerators can be reduced by using different regulation tools efficiently. Padilla et al. (2010) conducted exergy analysis and found that R413A could be a safe viable alternative to R12. Peng et al. (2010a) performed experiments on different combinations and found that carbon nanotubes can improve the characteristics of the refrigerant for nucleate pool boiling. In Peng et al. (2010b), increase of concentration of nanoparticles results in increase in nucleate pool boiling heat transfer coefficient also increases with diamond nanoparticles in the mixture of oil and R113; the maximum enhancement was found around 63.4%. Yang (2010) found by designing the duct and evaporator section that the maximum temperature difference and variation can be reduced by 7 to 3.5 ^{0}C and from 3.17 to 1.55 ^{0}C, respectively. In Abdel-Hadi et al. (2011), CuO nanoparticles can enhance the evaporative heat transfer coefficient upto 40 KW/m^2 by increasing the heat flux. The light absorbent ratio indexing method was used for the investigation of the dispersion stability of nanofluids. They found that carbon nanofluids are stable with an increase in mass fraction of the surfactant; Al_2O_3 nanofluid is disturbed initially than stable. Bolaji et al. (2011) As compared with R32 and R134a the COP is better with R152 by 2.5% and 14.7% with zero ODP & low GWP respectively. Laminar flow and convective heat transfer coefficient of water-alumina nanofluid on a rectangular was investigated by Aminfar and Maroofiazar (2011) on micro-channel numerically. A mixture model of nanofluids was used and compared with the single-phase model as they have slip velocity between the nanoparticles and base fluid. Results show that nanofluid as a coolant can enhance the pressure drop and heat transfer coefficient significantly. Their findings revealed that the combination of R134a refrigerant and mineral oil with nanoparticles operates seamlessly within the system. Moreover, the introduction of nanoparticles into the refrigerants leads to a remarkable reduction in power consumption, with potential savings of up to 25%, while simultaneously enhancing the COP (Coefficient of Performance) by 33%. An experimental study to investigate the performance by using TiO_2 nanoparticles mixed with R600a refrigerant in a domestic refrigerator was done by Bi et al. (2011) without any change in the circuit. TiO_2 mixed with R600a can save energy consumption up to 5.94% more than pure R600, and freezing velocity is also enhanced. The results show that TiO_2 mixed with R600a refrigerant works normally, safely, and efficiently. Mahabubul et al. (2011) found that significant affect is shown by increase in TiO_2 nanoparticle volume fraction results increase in the pressure gradient respectively. Peng et al. (2011) Nucleate pool boiling rate of heat transfer is improved by suspending Copper nanoparticles in the mixture of R-113 and oil. The fractional pressure drop characteristics of TiO_2 nanoparticles suspended in R-123 refrigerant was investigated and studied by Mahbubul et al. (2012). They found that volume fractions have an effect on pressure drop; pressure gradient increases with an increase in particle volume fraction. They used a mathematical relationship model to investigate thermal conductivity, viscosity, density, surface tension, specific heat, pressure drop, pumping power, and heat capacity of other fluids and compared them with nanorefrigerant. Saidur et al. (2011) studied the thermos-physical properties of suspended nanoparticles in refrigerants and lubricants. They also investigated the heat transfer performance of various nanorefrigerants at different concentration levels. Their study encompassed an examination of the impact on pumping power, pressure drop, and pool boiling heat transfer performance of carbon nanotubes (CNTs) within the system using different sources. They found that, at very low particle concentrations, the nanorefrigerants have very strong thermal conductivity that can improve the efficiency of the system. They conclude that HFC 134a and TiO_2 nanoparticle mixed in mineral oil can save 26.1% energy and gives better performance. To comprehend the influence of Al_2O_3 nanoparticles on the pool boiling performance of the R-134a/polyester mixture. Subramani and Prakash, (2011) found that Al_2O_3 nanoparticles with mineral oil power consumption is reduced upto 25 %, results improvement in COP by 33%. Kedzierski (2011) performed an experimental investigation on a roughened horizontal flat surface. They formulated a semi-empirical model for various mass fractions of 0.5%, 1%, and 2% lubricants, revealing average heat flux enhancements of 105%, 49%, and 155%, respectively. The effect of the size of nanoparticles on nucleate pool boiling heat transfer of refrigerants was studied experimentally by Peng et al. (2011). The experiment was conducted using R-113 and VG68 ester oil, along with copper nanoparticles of average diameters measuring 20 nm, 50 nm, and 80 nm. This testing occurred under operating conditions with a saturation pressure of 101.3 kPa and a heat flux range of 10 to 80 kW m^{-2}. The results indicated a remarkable enhancement in nucleate pool boiling heat transfer coefficient, reaching a maximum of 23.8% when the particle size was reduced from 80 to 20 nm. Liu et al.

(2011) experiment shows that heat transfer in Cu nanoparticles are better than CuO nanoparticles; thermal performance was decreased by SiO nanofluid. Yang et al (2011)found stability in carbon nano fluids with increase in mass fraction of the surfactant, Al_2O_3 nanofluid is disturbed initially then stable. Jwo et al. (2012) found that R-290 and R-600a with mass 90 gives best results in refrigeration instead of R-134a with 150g for the same circuit is reduced by 40%. total energy conserved is 4.4%. Study of the effect of viscosity of nanorefrigerant for R-123 refrigerant with TiO_2-suspended nanoparticles was done by Mahbubul et al. (2012). Their findings revealed that viscosity rises with an increase in particle volume fraction but decreases as temperature increases. Additionally, the pressure drop characteristics are contingent on viscosity, as an increase in pressure drop corresponds to an increase in volume concentration and vapor quality. They developed a method to improve the thermal conductivity of oil and water with use of graphite flakes as an additive for suspension; through combined study of optical and AC impedance spectroscopy, they explained the phenomenon of thermal percolation. PVP and CTAB can reduce the size of the crystal results; as surface area increases, 0.55 vol% of copper powder can improve the thermal conductivity by 22% and 48% improvement by 1.0 vol% over water. On investigation on the effect on boiling performance of R-134a mixed with diamond nanoparticles on a rough and flat surface, they mixed diamond nanoparticles of 2.6% volume fraction along with polyol ester lubricant in R-134a done by Kedzierski (2012). Based on the results obtained, it was observed that a 0.5% mass fraction of nanolubricant can enhance the heat transfer rate by as much as 98%. They draw the conclusion that, for improved boiling heat transfer, it is crucial to ensure that nanoparticles are effectively dispersed and remain suspended in the fluid (Mahbubul et al. 2012). Viscosity of nanofluids increases by increasing the particle volume fraction and decreases with temperature increase presented by Mahbubul et al. (2012). Their findings indicated that an increase in pressure drop correlates with an increase in the quality and concentration of vapor. Their investigation was aimed at examining the improvement of the convective heat transfer process in a square vertical cavity using an Al_2O_3 water-based nanofluid. Prapainnop and Suen (2012), performance of the system by the effect of derived parameters and thermo-physical properties of different refrigerants is studied. Saterile et al. (2012), experiments shows that nanofluid's thermal conductivity is improved by X-ray line broadening method to control the size of the particle by dynamic and crystalline light scattering, respectively. Wang et al. (2012) results show that cluster formation and percolation can enhance the thermal conductivity. The numerical findings closely align with the experimental data. They draw the conclusion that within the range of 1 to 3% solid volume fraction, there is an enhancement in the convective heat transfer rate. It was also found that the thermal conductivity of fluid is improved with the use of nanoparticles significantly. Aminfar and Haghgoo (2012) Addition of nanofluid can improve the heat transfer rate of natural convection process. Hu et al. (2013) SDS, CTAB and Span-80 can improve the nucleate pool boiling heat transfer of R-113 with copper nano lubricant. Experiments performed over a horizontal smooth tube with Al_2O_3 nanoparticles of concentration from 1 to 5% of volume in R-134a refrigerant were performed by Mahbubul et al. (2013). Their findings revealed that thermal conductivity can be enhanced by increasing both the volume fraction and temperature. Conversely, thermal conductivity decreases as the particle size increases. Additionally, an increase in nanoparticle concentration leads to higher viscosity, resulting in frictional pressure drop. The pumping power is directly related to the pressure drop in nanorefrigerants. Furthermore, due to the rise in nanoparticle concentration, the convective flow boiling heat transfer rate experiences a significant increase. They conducted experiments involving three surfactants: SDS, CTAB, and Span 80, which were mixed with a nanolubricant, blend to explore the characteristics of nucleate pool boiling heat transfer. The surfactant impact factor (SIF) increases with the increase of surfactant concentration, due to the increase of surfactant concentration. SIF also increases and reaches the maximum value. Sarkar et al. (2013) from experiments found that NH_3 is best for high temperature & C_3H_6 is the best as low temperature refrigerant for maximum COP. Sun and Yang (2013a) investigated flow boiling heat transfer characteristics of four nanorefrigerants, Cu-R141b, Al-R141b, Al_2O_3-R141b, and CuO-R141b, experimentally in an internal copper tube. They found that the mass fraction of nanoparticles can change the rate of heat transfer; high conductivity of nanoparticles gives better performance than another refrigerant. They also observed that with an increase in the quality of nanorefrigerants, the heat transfer coefficient experiences a substantial increase, reaching its maximum value in the range of $x = 0.5$–0.8. To understand the effect of nanoparticles on heat transfer performance of nanorefrigerants, Sun and Yang (2013b) conducted experiments using Cu-R141b, Al-R141b, Al_2O_3-R141b, and CuO-R141b nanorefrigerants. Their observations revealed that an increase in the mass fraction of nanoparticles led to an enhancement in the heat transfer coefficient. However, due to a decrease in the impact factor of nanoparticles, the improved heat transfer performance decreased as well. Furthermore, increasing the mass velocity also resulted in an improvement in the heat coefficient. Notably, Cu nanoparticles outperformed the others in enhancing the heat transfer performance of nanorefrigerants. The effect of stabilizer on nanofluids was revealed by Neogy and Raychaudhuri (2013), PVP stabilizer is added to the nanofluid with ZnO nanocrystals of an average diameter of 10 nm. Through practical investigation with and without stabilizer, they found that by addition of stabilizer local aggregation increased and reduced the thermal parameters.

They compared the effect of the stabilizer with and without adding nanofluids. PVP stabilizers resist the thermos-diffusion-assisted local aggregation which results in a decrease in thermal transport properties to control the nanofluids. The performance of vapor refrigeration systems by using R-134a in the primary circuit and Al_2O_3 water-based nanofluid in the secondary circuit by thermal modeling was investigated by Mishra (2014). The model extracts information of secondary fluid like the size of nanoparticles, compressor speed, and geometric conditions to predict the output of the secondary fluid like operating pressure, temperature, power consumption, and overall system performance. According to the results obtained from the simulation investigation, the system's performance can be enhanced by approximately 17–20%; design of components like compressor, evaporator, condenser, and throttle valve can be done. Before boiling, the surface of the plate was finished with #2000 grade of sandpaper; then, at three different concentrations of 0.001 vol%, 0.01 vol%, and 0.1 vol%, δ-Al_2O_3 nanoparticles are suspended in R141b base refrigerant with and without SDBS surfactant. It is observed that 0.001 vol% and 0.01 vol% nanoparticle mixed refrigerant improves the pool boiling heat transfer coefficient; it was noted that, at 0.1 vol% concentration of δ-Al_2O_3 nanoparticles, the pool boiling heat transfer characteristics reduced without the surfactant due to segregation of nanoparticles at the bottom level. They conclude that for 0.001 vol% of nanoparticle, SDBS surfactant is not required as it decreases the performance but at 0.01 vol% and 0.1 vol% of concentration SDBS with increase the performance. To investigate the pool boiling heat transfer characteristics of nanofluid experiments on the cylindrical surface of brass was performed by Naphon and Thongjing (2014). R141b is used as base fluid in which TiO_2 nanoparticles are suspended and ethyl alcohol is used as working fluid. The investigation was performed under desired boiling pressure controlled by the coolant flow rate through the control valve. The cylindrical heater is connected to electric power at input, voltage, and current supplied to the heater to compute the heat flux, at high heat flux the particle volume concentration increases result to decrease in the boiling heat transfer coefficient, and at higher boiling pressure, the boiling heat transfer coefficient increases. The effect of CuO nanoparticles mixed with R-600a refrigerant on flow boiling was investigated by Behabadi et al. (2014). Tang et al (2014) shows that pool boiling of R141b is enhanced by suspending Al_2O_3 nanoparticles & SDBS (sodium dodecyl benzene sulphonate) as surfactant through experimental results. An experimental test rig was meticulously designed and constructed, comprising key components such as a flow meter, pump, preheater, condenser, sub-cooler, bypass, and an evaporator constructed from a smooth horizontal copper tube measuring 9.52 mm in diameter, 0.63 mm in thickness, and 103 mm in length. To ensure even heating, the test section was enclosed by a heater coil, and the preheater was insulated with glass wool to create an isolated system for measuring the flow boiling heat transfer of the mixture. To achieve uniform dispersion of nanoparticles in the refrigerant mixture, an ultrasonic instrument was employed. The investigation revealed that the utilization of CuO nanoparticles in R-600a/oil resulted in a significant increase in maximum heat transfer, with improvements of up to 63%. Experimental results showed a 2–6% improvement in COP with R1234yf and R1234ze as refrigerants compared to R-134 by Mishra (2015). Investigation of pool boiling heat transfer and critical heat flux of nanorefrigerant mixture of Cu-R141b-SDBS and R141b-SDBS under atmospheric pressure on flat surface was performed by Diao et al. (2015). The experimental setup was consisted of a test vessel, a condensation loop, and a heating unit, Agilent data acquisition system connected to a computer. They conducted experiments using nanorefrigerants with varying concentrations of 0.008 vol%, 0.015 vol%, and 0.05 vol%, in addition to pure R141b. The findings indicate that Cu-R141b-SDBS refrigerant can enhance pool boiling heat transfer when compared to pure R141b. Nanorefrigerants have lower CHF at bare surface than pure R141b. Nanorefrigerants are synthesized using a two-step method, resulting in an average particle diameter of 20 nm. Surfactant span-80 is used with the transmittance method to check the stability of nanorefrigerants. Their findings demonstrated that R141b nanorefrigerants infused with 0.3 wt% MWCNTs exhibited enhanced optimal heat transfer when compared to pure refrigerants. To understand the aggregation behavior of nanoparticles which affects the stability and heat transfer performance for a long period and changes in primary particle parameters with surfactants, an experimental investigation was done by Peng et al. (2015). They selected R-141b refrigerant and TiO_2 nanoparticles of 25, 40, 60, and 100 nm particle size for preparation of nanorefrigerant. The light scattering method was employed to monitor the time evolution of particle size. It was observed that as the primary particle size increased from 25 to 100 nm, the steady-state hydrodynamic diameter increased by a substantial 127.6%. An experimental investigation on the plane horizontal tube was done to study the thermal characteristics of nanorefrigerant performed by Behabadi et al. (2015). The experimental setup was constructed using various components, including a gear pump, heater, test condenser, post condenser, bypass path, and flow meter. R-600a was selected as the primary refrigerant, and three different combinations were investigated: pure R-600a, R-600a with POE oil, and R-600a with 50 nm CuO nanoparticles. Their conclusion indicates that incorporating nanoparticles into the refrigerant/oil mixture leads to an enhancement in the heat transfer coefficient compared to using pure refrigerant alone;1.5 mass fraction of Cu nanoparticle can improve the heat transfer up to 83%. Peng et al. (2015) recorded that there is rise in stable-state hydrodynamic diameter of about 127 %. They performed experiments at ambient pressure, particle size

from 25 to 100 nm, temperature from 6^0 to 27^0C, and oil concentration from 50 to 500 mg L^{-1} on the DLS apparatus. They found that lubricating oil present in nanorefrigerants reduces the accumulation of nanoparticles by increasing the particle concentration; primary particle size and temperature boost the accumulation of nanoparticles. Yang et al. (2015) recorded optimal rate of heat transfer is obtained by R141b and MWCNT (0.3 wt %) nanorefrigerants as compared to pure refrigerants. Lin et al. (2016) Extraction of T_iO_2 in refrigerant is reduced by use of lubricating oil. It also improves the particle size, concentration & temperature in the system. Sajid and Bicer (2020a) investigated hydrodynamic characteristics of TiO_2 nanofluid as a coolant in way channel heat sink under variation of nanoparticles volume fraction, Reynold's number, and heating power. They found that the geometry of the heat sink affects the heating power, Nusselt number increases, and the maximum improvement noted is 40.57% as compared with distilled water. The results indicate that under identical input conditions, nanofluids exhibit superior heat transfer characteristics compared to distilled water for all heat sinks. Sajid et al. (2019) experimental results shows that nanofluids have better heat transfer properties than distilled water, thermal properties of nanofluid is decreased by increasing the heating power. The performance is improved better by CuO as compared to Al_2O_3, and TiO_2 reported by Mishra (2020a). Correlations are used to predict the properties of nanofluid, and metallic oxide nanoparticles from alumina, titania, and, copper and are used for nanofluids which can improve the heat transfer characteristics. They discussed the numerical and experimental investigation to understand the effect and stability of nanofluids by experimental as well as numerical approaches. A square-shaped pin fin heat sink with a fin angle of 22.50 with graphene oxide nanoparticles was investigated by Nawas et al. (2022) and they found that thermal conductivity is improved by increasing the Reynolds number and heating power. The influence of various factors, including nanoparticle size, shape, concentration, choice of base fluid, surfactant, and optical path, is briefly discussed. Different models and correlations are used to predict the behavior of the nanofluids. Sajid and Bicer et al. (2020b) studied and proposed a self-sustainable greenhouse model for utilizing solar energy in power generation, water requirement, and cooling of space. They proposed the use of nanofluids in the PV/T system for electricity generation in which the visible spectrum of light is used in the photosynthesis of the plants, thermal energy of the PV/T system is used for the desalination process of seawater to produce fresh water, and the requirement of heat for absorption cooling system is fulfilled by thermal energy storage feed from evacuated tube collectors of solar heater. In Soudagar et al. (2020), cooling mechanisms of mini-channels for devices are discussed with water-based and ethylene glycol-based nanofluids of Cu, Al_2O_3 SiC-W and TiO_2 to absorb heat. Mishra (2020b) found that the performance is improved better by CuO as compared to Al_2O_3, and TiO_2. In Mishra (2020c), R1234yf gives better performance over other HFO refrigerants. Mishra (2020d) proposed that system performance improves from the range of -50 to 150^0C with replacement of R134a. Mishra (2020e) found that R-245fa, R-32, and other HFO refrigerants can be replaced by R-134a. Mishra (2020f) noted that R-134 is replaced by refrigerants like R-245fa and other HFO refrigerant blends. Sajid and Bicer (2020a) discussed the characteristics like absorbance, transmittance, and scattering of nanofluids as solar spectrum splitters. Effect of different factors such as size, shape, and concentration of nano-particles, type of base fluid, surfactant, and optical path is discussed briefly. Different models and correlations are used to predict the behavior of the nanofluid. Sajid et al. (2020a) did comparative study to investigate thermal characteristics and exergy efficiency of distilled water and MgO-water nano-fluid as coolant on corrugated mini-channel heat sink and found that nano-particles can improve the rate of heat transfer. Exergy efficiency is noted high with nano-fluid at all concentration as compared to distilled water and can be improved by increasing Reynold's number and concentration of nanoparticles. Sajid and Bicer et al. (2020b) studied and proposed a self-sustainable greenhouse model for utilizing solar energy in power generation, water requirement, and cooling of the space. They proposed use of nanofluid in the PV/T system for electricity generation in which visible spectrum of light is used in photosynthesis of the plants, thermal energy of the of PV/T system is used for desalination process of seawater to produce fresh water, and requirement of heat for absorption cooling system is fulfilled by thermal energy storage feed from Evacuated tube collectors of solar heater. Shaik et al. (2020) explore novel refrigerant mixtures as environmentally friendly R22 replacements for air conditioners, with R1270, RM30, and RM50 emerging as promising alternatives that offer lower CO_2 emissions, good performance, and favorable toxicity and flammability classifications. Sharma et al. (2020) review emphasizes the importance of thermal energy storage in maintaining the reliability of solar thermal-powered vapor absorption systems, with a focus on selecting suitable solar collectors, thermal energy storage systems, and economic feasibility for sustainable and consistent operation. In Li et al. (2021), in the ejector refrigeration system, R1234yf emerges as a promising environmentally friendly replacement for R134a, offering a greater entrainment ratio despite its slightly closer shocking position upstream, making it a favorable alternative with minimal critical temperature difference. Padmavathy et al. (2021) explored low-GWP refrigerant mixtures as eco-friendly alternatives to R134a in a domestic refrigerator, highlighting R1234ze/R134a (90/10) as a promising replacement, with improved exergy efficiency and suitability for lower temperature applications to align with environmental protocols. Selvaraj and Victor (2021) explore the need for alternative refrigeration systems due to

environmental concerns, highlighting power quality challenges in vapor compression and vapor absorption refrigeration, along with a comparative analysis of these technologies. In Esfanjani et al. (2022), recent studies highlight the potential of solar-driven cooling systems using various solar collectors, absorption chillers, and cryogenic systems to reduce carbon emissions and increase freshwater and LNG production by harnessing sustainable solar energy. Kumma and Kruthiventi et al. (2022) The study explores eco-friendly HC-HFC ternary mixtures as promising alternatives to R134a in domestic refrigeration, with R600a/R13I1/R125 identified as a superior, more efficient replacement, offering over 10% higher COP compared to R134a under similar conditions. In Li and Jun (2022), complete summary is prepared with the help of different simulation models for components of VCRS system like evaporator, condenser, compressor, capillary tube. Nawaz et al. (2022) investigated square shaped pin fin heat sink of fin angle of 22.50 with graphene oxide nanoparticles and found that thermal conductivity is improved by increase of Reynolds number and heating power. Noman et al. (2022) review geothermal heating and cooling systems, emphasizing Earth-Air Heat Exchangers (EAHE) and Ground Source Heat Pumps (GSHP), highlighting key factors such as soil thermal conductivity and moisture content that enhance system performance and the potential for hybrid energy integration. In Prabakaran et al. (2022), in replacing R22 for split AC units, a multi-criteria analysis suggests R290 and R1123 as the top and least favorable alternatives, considering factors like COP, TEWI, and lifetime cost alongside thermal properties. Tan et al. (2022) experimentally examine the performance of an ultralow-temperature cascade refrigeration freezer using natural refrigerants, demonstrating its suitability for storing and transporting COVID-19 vaccines but not ideal for high freezing temperatures due to high power consumption and frequent start-stop cycles.

Results and discussion

Numerous studies have been carried out to evaluate the performance of low global warming potential (GWP) refrigerants, including hydrofluorocarbons (HFCs) and hydrofluoroolefins (HFOs), in both the primary and secondary circuits of the evaporator. It has been observed that R-178 is predominantly utilized in the secondary circuit of the evaporator. Among the eco-friendly refrigerants studied, R1224yd(Z) exhibits superior first law performance under comparable operating conditions compared to R-1223zd(E). Furthermore, in the primary circuit, employing HFO1336mzz(Z) in the evaporator leads to enhanced performance compared to R1243zf, R1234yf, R1225ye(Z), and R-134a, albeit slightly inferior to R1233zd(E). The introduction of low GWP refrigerants blended with R-178 in the secondary circuit of the evaporator resulted in discernible variations in first law performance, both in terms of improvement and decrement. Incorporating nanomaterials with R-1233zd(E), R-1224yd(Z), and HFO-1336mzz(Z) as substitutes for R134a yields notable enhancements in second law efficiency, with improvements of 8.16%, 5.86%, and 3%, respectively. Notably, the compressor's isentropic efficiency is highest when using CuO and lowest with TiO2 in conjunction with HFO1336mzz(Z). Similarly, the compressor's volumetric efficiency follows the same trend, being highest with CuO and lowest with TiO_2 in combination with HFO1336mzz(Z).

The integration of nanomaterials, specifically CuO, Al_2O_3, and TiO_2, leads to significant improvements in first law efficiency, registering increments of approximately 18.5%, 17.5%, and 15.95%, respectively. The use of CuO, Al_2O_3, and TiO_2 for condenser tube enhancements also results in variations in exergetic efficiency, with improvements of about 13.9%, 12.9%, and 11.5%, respectively.

Moreover, the overall evaporator heat transfer coefficient is highest when utilizing CuO and lowest with TiO_2 in combination with HFO1336mzz(Z). Similarly, the overall condenser heat transfer coefficient exhibits the same trend, with CuO yielding the highest coefficient and TiO_2 yielding the lowest when paired with HFO1336mzz(Z). Investigations are done and it was found that different investigations are done for low GWP HFC and HFO refrigerants used in the primary circuit, and R-178 is mostly used in the secondary circuit of the evaporator, and it was found that refrigerant R1224yd(Z) gives better 1st law performance than R-1223zd (E) for the same operating conditions as compared to other eco-friendly refrigerants. Evaporator filled with the HFO1336mzz (Z) fluid in the primary circuit gives better performance than R1243zf, R1234yf, and R1225ye (Z) and R-134a but lower than R1233zd (E). Variation in first law performance (improvement/decrement) was found by using low GWP refrigerants mixed with R-178 refrigerant in the evaporator's secondary circuit. Refrigerant R-1233zd(E) with nanomaterials can improve the 2nd law efficiency by 8.16%, R-1224yd (Z) by 5.86% and HFO-1336mzz (Z) by 3% as a replacement of R134a. Compressor's isentropic efficiency was found highest with CuO and lowest with TiO_2 with HFO1336mzz (Z). Compressor's volumetric efficiency was found highest with CuO and lowest with TiO_2 with HFO1336mzz (Z). The 1st law efficiency is improved by using nanomaterials CuO, Al_2O_3, and TiO_2 about 18.5%, 17.5%, and 15.95%, respectively. Improvements in the second law efficiency were observed in the condenser tube through the utilization of CuO, Al_2O_3, and TiO_2 nanoparticles, resulting in enhancements of approximately 13.9%, 12.9%, and 11.5%, respectively. Overall evaporator heat transfer coefficient was found highest with CuO and lowest with TiO_2 with HFO1336mzz (Z). Overall condenser heat transfer coefficient was found highest with CuO and lowest with TiO_2 with HFO1336mzz (Z).

Conclusions

In this research article, with various refrigerants with suspended nanoparticles in the circuit, the performance is summarized with the help of previous research on system performance enhancement (COP), heat transfer characteristics, and solubility of nanorefrigerants into the base refrigerant. From the present review, following points are extracted:

- The addition of nanolubricants and nanorefrigerants enhances the heat transfer rate in both pool boiling and nucleate boiling processes, leading to improved heat transfer performance in Variable Capacity Refrigeration Systems (VCRS).
- Nanolubricants and nanorefrigerants reduce the power consumption in refrigeration and air conditioning applications.
- It was found that maximum improvement in thermal conductivity (46%) is achieved by the use of CNT-based nanorefrigerants.
- Nanoparticles are easily soluble and miscible in the refrigerant without any change in viscosity, resulting in smooth operation in VCRS with a constant oil return ratio.
- Compressor life is increased due to low friction and wear by working fluid, as nanolubricants have good tribological characteristics.
- Incorporating nanolubricants and nanorefrigerants can enhance the freezing speed, freezing capacity, and overall Coefficient of Performance (COP) of the system.
- Optimal performance is achieved by using R-134a (60%) and Al_2O_3 0.1 wt%. With such mixture, the COP is improved by 4.4%, and the power consumption is improved by 2.4%.
- In flow boiling and condensation processes, key parameters such as heat flux, mass flux, vapor quantity, and the weight percentage of nanoparticles exert a significant influence.
- Compared to pure refrigerants, the utilization of nanorefrigerants under identical input parameters in the system can lead to a remarkable improvement in heat transfer rates, ranging from 21 to 275%.

Availability of data and materials The data supplied in this research can be attained on request from the corresponding author.

Author contribution USP performed the idea development, methodology, and analysis and wrote down the original document. Dr. RSM inspected the study and writing for the draft manuscript. RKD corrected the documents for final submission. All authors have understood and agreed to the published edition of the document. All authors have read and agreed to the published version of the manuscript.

Declarations

Ethical approval Not applicable.

Consent to participate I have read and I understand the provided information and have had the opportunity to ask questions. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason and without cost.

Consent for publication I, the undersigned, give my consent for the publication of identifiable details, which can include photograph(s) and/or videos and/or case history and/or details within the text (Study of vapor compression refrigeration system with suspended nanoparticles in the low GWP refrigerant) to be published in the above journal and article.

Competing interests The authors declare no competing interests.

References

Abdel-Hadi EA-H, Taher SH, Torki AHM, Hamad SS (2011) Heat transfer analysis of vapour compression system using nano CuO-R-134a. In: International Conference on Advanced Materials Engineering, IPCSIT 15. https://api.semanticscholar.org/CorpusID:138078734

Aminfar H, Haghgoo MR (2012) Brownian motion and thermophoresis effects on natural convection of alumina–water nanofluid. Proc Inst Mech Eng C J Mech Eng Sci 227(1):100–110. https://doi.org/10.1177/0954406212445683

Aminfar H, Maroofiazar R (2011) A numerical study of the hydro- thermal behavior of nano fluids in rectangular micro channels using a mixture model. Proc Inst Mech Eng Part C: J Mech Eng Sci 225(4):791–798. https://doi.org/10.1243/09544062JMES2419

Behabadi MA, Nasr M, Baqeri S (2014) Experimental investigation of flow boiling heat transfer of R-600a/oil/CuO in a plain horizontal tube. Exp Therm Fluid Sci 58:105–111. https://doi.org/10.1016/j.expthermflusci.2014.06.013

Behabadi MA, Sadoughi MK, Darzi M, Fakoor-Pakdaman M (2015) Experimental study on heat transfer characteristics of R600a/POE/CuO nano-refrigerant flow condensation. Exp Therm Fluid Sci 66:46–52. https://doi.org/10.1016/j.expthermflusci.2015.02.027

Bi S, Shi L, Zhang L (2007) Performance study on domestic refrigerator using R134a/mineral oil/nano TiO as working fluid. Int J Refrig 106:184–190 https://iifiir.org/en/fridoc/performance-study-of-a-domestic-refrigerator-using-r-134a-mineral-24631

Bi S-s, Shi L, Zhang L-l (2008) Application of nanoparticles in domestic refrigerators. Appl Therm Eng 28(14–15):1834–1843. https://doi.org/10.1016/j.applthermaleng.2007.11.018

Bi S, Guo K, Liu Z (2011) Performance of domestic refrigerator using TiO2- R600a nanorefrigerant as working fluid. Energy Conserv Management 52(1):733–737. https://doi.org/10.1016/j.enconman.2010.07.052

Bobbo S, Fedele L, Fabrizio M, Barison S, Battiston S, Pagura C (2010) Influence of nanoparticles dispersion in POE oils on lubricity and R134a solubility. Int J Refrig 33(6):1180–1186. https://doi.org/10.1016/j.ijrefrig.2010.04.009

Bolaji BO (2010) Experimental study of R152a and R32 to replace R134a in a domestic refrigerator. Int J Energy Eng 35:3793–3798. https://doi.org/10.1016/j.energy.2010.05.031

Bolaji BO, Akintunde MA, Falade TO (2011) Comparative analysis of performance of three ozone-friends HFC refrigerants in a vapour compression refrigeration. J Environ Sustain 2:61–64 http://repository.fuoye.edu.ng/handle/123456789/1231

Choi SUS (1995) In: Singer DA, Wang HP (eds) Enhancing thermal conductivity of fluids with nanoparticles, in Developments and applications of non-Newtonian flows, vol FED-231/MD-66. ASME, New York, pp 99–105 https://www.osti.gov/servlets/purl/196525

Dalkilic AS, Wongwises S (2010) A performance comparison of vapor-compression refrigeration system using various alternatives refrigerants. Int Commun Heat Mass Transf 37:1340–1349. https://doi.org/10.1016/j.icheatmasstransfer.2010.07.006

Diao YH, Li CZ, Zhao YH, Liu Y, Wang S (2015) Experimental investigation on the pool boiling characteristics and critical heat flux of Cu-R141b nanorefrigerant under atmospheric pressure. Int J Heat Mass Transf 89:110–115. https://doi.org/10.1016/j.ijheatmasstransfer.2015.05.043

Ding G-l (2007) Recent developments in simulation techniques for vapour-compression refrigeration systems. Int J Refrig 30:1119–1133. https://doi.org/10.1016/j.ijrefrig.2007.02.001

Ding G, Peng H, Jiang W, Gao Y (2009) The migration characteristics of nanoparticles in the pool boiling process of nanorefrigerant and nanorefrigerant–oil mixture. Int J Refrig 32(1):114–123. https://doi.org/10.1016/j.ijrefrig.2008.08.007

Dossat RJ, Horan TJ (2002) Principles of refrigeration. Prentice hall International Inc., New Jersey, p 454 https://www.worldcat.org/title/principles-of-refrigeration/oclc/464217230

Eastman JA, Choi US, Thompson LJ, Lee S (1996) Enhanced thermal conductivity through the development of nano fluids. Mater Res Soc Symp Proc 457:3–11. https://doi.org/10.1557/PROC-457-3

Esfanjani P, Jahangiri S, Heidarian A (2022) A review on solar-powered cooling systems coupled with parabolic dish collector and linear Fresnel reflector. Environ Sci Pollut Res 29:42616–42646. https://doi.org/10.1007/s11356-022-19993-3

Fatouh M, El Kafafy M (2006) Assessment of propane / commercial butane mixtures as possible alternatives to R134a in domestic refrigerators. Energ Conver Manage 47:2644–2658. https://doi.org/10.1016/j.enconman.2005.10.018

Gigiel A (2004) Safety testing of domestic refrigerators using flammable refrigerants. Int J Refrig 27:621–628. https://doi.org/10.1016/j.ijrefrig.2004.03.001

Hamilton RL, Crosser OK (1962) Thermal conductivity of heterogeneous two-component systems. Ind Eng Chem Fundamentals 1(3):187–191. https://doi.org/10.1021/i160003a005

Havelsky V (2000) Investigation of refrigerating system with R12 refrigerant replacements. Appl Therm Eng 20:133–140. https://doi.org/10.1016/S1359-4311(99)00016-2

He M-G, Li T-C, Liu Z-G, Zhang Y (2005) Testing of the mixing refrigerants HFC152a/HFC125 in domestic refrigerator. Appl Therm Eng 25:1169–1181. https://doi.org/10.1016/j.applthermaleng.2004.06.003

Hong TK, Yang HS, Choi CJ (2005) Study of the enhanced thermal conductivity of Fe nanofluids. J Appl Phys 97:064311. https://doi.org/10.1063/1.1861145

Hozod M (2005) Performance comparison of single-stage and cascade refrigeration systems using R134a as the working fluid. Turk J Eng Environ Sci 29:285–269

Hu H, Peng H, Ding G (2013) Nucleate pool boiling heat transfer characteristics of refrigerant/nanolubricant mixture with surfactant. Int J Refrig 36(3):1045–1055. https://doi.org/10.1016/j.ijrefrig.2012.12.015

Jiang W, Ding G, Peng H (2009a) Measurement and model on thermal conductivities of carbon nano tube nano refrigerant. Int J Therm Sci 48(2009):1108–1115. https://doi.org/10.1016/j.ijthermalsci.2008.11.012

Jiang W, Ding G, Peng H, Gao Y, Wang K (2009b) Experimental and model research on nanorefrigerant thermal conductivity. HVAC&R Research 15(3):651–669. https://doi.org/10.1080/10789669.2009.10390855

Joseph Sekher S (2004) Improved energy efficiency for CFC domestic refrigerators retrofitted with ozone friendly HFC134a. Int J Therm Sci 43:307–314. https://doi.org/10.1016/j.ijthermalsci.2003.08.002

Jwo C-S (2008) Experimental study on thermal conductivity of lubricant containing nanoparticles. Rev Adv Mater Science 18:660–666 https://www.ipme.ru/e-journals/RAMS/no_71808/jwo.pdf

Jwo C-S, Jeng L-Y, Teng T-P, Chang H (2009) Effects of nanolubricant on performance of hydrocarbon refrigerant system. J Vac Sci Technol B 27(3):1473–1477. https://doi.org/10.1116/1.3089373

Jwo C-S, Ting C-C, Wang W-R (2012) Efficiency analysis of home refrigerators by replacing hydrocarbon refrigerators. Int J Meas 42:697–701. https://doi.org/10.1016/j.measurement.2008.11.006

Kang SW, Wei WC, Tsai SH, Yang SY (2006) Experimental investigation of silver nano fluid on heat pipe thermal performance. Appl Therm Eng 26:2377–2382. https://doi.org/10.1016/j.applthermaleng.2006.02.020

Kedzierski MA (2011) Effect of Al2O3 nanolubricant on R134a pool boiling heat transfer. Int J Refrig 34(2):498–508. https://doi.org/10.1016/j.ijrefrig.2010.10.007

Kedzierski MA (2012) Effect of diamond nanolubricant on R134a pool boiling heat transfer. ASME J Heat Transfer 134:051001–051001. https://doi.org/10.1115/1.4005631

Kumma N, Kruthiventi SSH (2022) Effect of dilutant concentration on the performance of environment-friendly mixed refrigerants used in domestic refrigerators. Environ Sci Pollut Res 29:71988–72000. https://doi.org/10.1007/s11356-022-20965-w

Lee D, Kim JW, Kim BG (2006) A new parameter to control heat transport in nanofluids: surface charge state of the particle in suspension. J Phys Chem B 110(9):4323–4328. https://doi.org/10.1021/jp057225m

Li S, Jun L (2022) A theoretical comparative study of vapor-compression refrigeration cycle using Al2O3 nanoparticle with low-GWP refrigerants. Entropy 24(12):1820. https://doi.org/10.3390/e24121820

Li S, Liu Y, Liu Y (2021) Performance comparison of ejectors in ejector-based refrigeration cycles with R1234yf, R1234ze(E) and R134a. Environ Sci Pollut Res 28:57166–57182. https://doi.org/10.1007/s11356-021-14626-7

Lin L, Peng H, Ding G (2016) Experimental research on particle aggregation behavior in nanorefrigerant-oil mixture. Appl Therm Eng 98:944–953. https://doi.org/10.1016/j.applthermaleng.2015.12.052

Liu Z, Imam Haider BY, Liu RR (1994) Test result of hydro carbon mixtures in domestic refrigerators freezers Centre Environmental Energy Engineering. University of Maryland College Park, Maryland

Liu Z-H, Li Y-Y, Bao R (2011) Composite effect of nano particle parameter on thermal performance of cylindrical micro-grooved heat pipe using nano fluids. Int J Therm Sci 50:558–568. https://doi.org/10.1016/j.ijthermalsci.2010.11.013

Liu M-S, Lin MC-C, Huang I-T, Wang C-C (2006) Enhancement of thermal conductivity with CuO for nanofluids. Chem Eng Technol 29(1):72–77. https://doi.org/10.1002/ceat.200500184

Liu ZH, Zhu QZ (2011) Application of aqueous nano fluids in a horizontal mesh heat pipe. Energ Conver Manage 52:292–300. https://doi.org/10.1016/j.enconman.2010.07.001

Mahbubul IM, Fadhilah SA, Saidur R, Leong KY, Amalina MA (2013) Thermo physical properties and heat transfer, performance of Al2O3/R-134a nano refrigerants. Int J Heat Mass Transf 57:100–108. https://doi.org/10.1016/j.ijheatmasstransfer.2012.10.007

Mahabubul IM, Saidur R, Amalina MA (2011) Pressure drop characteristics of TiO2-R123 nano refrigerant in a circular tube. Eng e-trans 6(2):124–130

Mahabubul IM, Saidur R, Amalina MA (2012) Investigation of viscosity of R123-TIO2 nano refrigerant. Int J Mech Mater Eng 7(2):146–151

Mahlia TMI, Saidur R (2010) A review on test procedure, energy efficiency standards and energy labels for room air conditioners and refrigerator-freezers. Renew Sustain Energy Rev 14:1888–1900. https://doi.org/10.1016/j.rser.2010.03.037

Soudagar MEM, Kalam MA, Sajid MU, Afzal A, Banapurmath NR, Akram N, Mane SD, Ahamed Saleel C (2020) Thermal analyses of minichannels and use of mathematical and numerical models. Numer Heat TR A-Appl 77(5):497–537. https://doi.org/10.1080/10407782.2019.1701883

Maxwell JC (1873) Treatise on electricity and magnetism. Clarendon Press, Oxford

Misheck G, Mwaba XH, Junjie G (2006) Influence of wick characteristics on heat pipe performance. Int J Energy Res 30:489–499. https://doi.org/10.1002/er.1164

Mishra RS (2014) Methods for improving thermodynamic performance of vapour compression refrigeration system using twelve eco friendly refrigerants in primary circuit and nanofluid (water-nano particles based) in secondary circuit. Int J Emerg Technol Adv Eng 4(6)

Mishra RS (2015) Irreversibility reduction in vapour compression refrigeration systems using Al O nano material mixed in R718 as Secondary fluid. Int J Adv Res Innovation 3(2):321–327

Mishra RS (2020a) Energy-exergy performance evaluation of new HFO refrigerants in the modified vapour compression refrigeration systems using liquid vapour heat exchanger. International Journal of Advance Research and Innovation 4(2):77–85. https://doi.org/10.36037/ijrei.2021.5204

Mishra RS (2020b) Methods for enhancing energetic-exergetic performances of vapour compression refrigeration system using HFO-1336mzz(Z) refrigerant through nano materials mixed with R-718 fluid in secondary circuit of evaporator. Int J Adv Res Innovation 4(3):143–148

Mishra RS (2020c) New & low GWP eco-friendly refrigerants used for predicting thermodynamic (energy-exergy) performances of cascade vapour compression refrigeration system using for replacing R134a, R245fa and R32. Int J Adv Res Innovation 4(3):124–130. https://doi.org/10.36037/IJREI.2020.4301

Mishra RS (2020d) Single and multiple cascading of VCRS in NH H O vapour absorption refrigeration systems for improving thermodynamic (energy-exergy) performances using five ecofriendly new HFOs and other low GWP refrigerants for replacing R134a. Int J Adv Res Innovation 4(2):96–104

Mishra RS (2020e) Thermal performance of multi cascaded vapour compression NH H O absorption systems for ultra-low temperature applications. Int J Adv Res Innov 4(1):39–55. https://doi.org/10.36037/IJREI.2020.4104

Mishra RS (2020f) Thermal performance of three stage cascade vapour compression refrigeration systems using new HFO in high and intermediate temperature cycle and R32 ethylene and hydrocarbons in ultra-low temperature cycle refrigerants. Int J Adv Res Innov 4(2):109–123. https://doi.org/10.36037/IJREI.2020.4206

Mohanraj M, Jayaraj S, Muraleedharan C (2009a) Environment friendly alternatives to halogenated refrigerants – a review. Int J Greenhouse Gas Control 3:108–119. https://doi.org/10.1016/j.ijggc.2008.07.003

Mohanraj M, Jayaraj S, Muraleedharan C, Chandrasekar P (2009b) Experimental investigation of R290/R600a mixture as an alternative to R134a in a domestic refrigerator. Int J Therm Sci 48:1036–1042. https://doi.org/10.1016/j.ijthermalsci.2008.08.001

Naphon P, Thongjing C (2014) Pool boiling heat transfer characteristics of refrigerant- nanoparticle mixtures. Int Commun Heat Mass Transf 52:84–89. https://doi.org/10.1016/j.icheatmasstransfer.2014.01.014

Naphon P, Thongkum D, Assadamongkol P (2009) Heat pipe efficiency enhancement with refrigerant-nano particles mixtures. J Energ Convers Manage 50:772–776. https://doi.org/10.1016/j.enconman.2008.09.045

Nawaz S, Babar H, Ali HM, Sajid MU, Janjua MM, Said Z, Tiwari AK, Syam Sundar L, Li C (2022) Oriented square shaped pin-fin heat sink: Performance evaluation employing mixture based on ethylene glycol/water graphene oxide nanofluid. Appl Therm Eng 206:118085. https://doi.org/10.1016/j.applthermaleng.2022.118085

Neogy RK, Raychaudhuri AK (2013) Effect of stabilizer on dynamic thermal property of ZnO nanofluid. Nanoscale Res Lett 8:12. https://doi.org/10.1186/1556-276X-8-125

Noman S, Tirumalachetty H, Athikesavan MM (2022) A comprehensive review on experimental, numerical and optimization analysis of EAHE and GSHP systems. Environ Sci Pollut Res 29:67559–67603. https://doi.org/10.1007/s11356-022-22228-0

Padilla M, Revellin R, Bonjour J (2010) Exergy analysis of R413A as replacement of R12 in a domestic refrigeration system. Energ Conver Manage 51:2195–2201. https://doi.org/10.1016/j.enconman.2010.03.013

Padmavathy R, Chockalingam MP, Kamaraj N (2021) Performance studies of low GWP refrigerants as environmental alternatives for R134a in low-temperature applications. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-021-15875-2

Park KJ, Jung D (2007a) Boiling heat transfer enhancement with carbon nanotubes for refrigerants used in building air-conditioning. Energ Buildings 39(9):1061–1064. https://doi.org/10.1016/j.enbuild.2006.12.001

Park KJ, Jung D (2007b) Enhancement of nucleate boiling heat transfer using carbon nanotubes. Int J Heat Mass Transf 50:4499–4502. https://doi.org/10.1016/j.ijheatmasstransfer.2007.03.012

Peng H, Ding G, Jiang W, Hu H, Gao Y (2009) Heat transfer characteristics of refrigerant-based nanofluid flow boiling inside a horizontal smooth tube. Int J Refrig 32(6):1259–1270. https://doi.org/10.1016/j.ijrefrig.2009.01.025

Peng H, Ding G, Hu H, Jiang W (2010a) Influence of carbon nanotubes on nucleate pool boiling heat transfer characteristics of refrigerant–oil mixture. Int J Therm Sci 49(12):2428–2438. https://doi.org/10.1016/j.ijthermalsci.2010.06.025

Peng H, Ding G, Hu H, Jiang W, Zhuang D, Wang K (2010b) Nucleate pool boiling heat transfer characteristics of refrigerant/oil mixture with diamond nanoparticles. Int J Refrig 33(2):347–358. https://doi.org/10.1016/j.ijrefrig.2009.11.007

Peng H, Ding G, Hu H, Jiang W (2011) Effect of nanoparticle size on nucleate pool boiling heat transfer of refrigerant/oil mixture with nanoparticles. Int J Heat Mass Transf 54:1839–1850. https://doi.org/10.1016/j.ijheatmasstransfer.2010.12.035

Peng H, Lin L, Ding G (2015) Influences of primary particle parameters and surfactant on aggregation behavior of nanoparticles in nanorefrigerant. Energy 89:410–420. https://doi.org/10.1016/j.energy.2015.05.116

Prabakaran R, Sivalingam V, Kim SC (2022) Future refrigerants with low global warming potential for residential air conditioning system: a thermodynamic analysis and MCDM tool optimization. Environ Sci Pollut Res 29:78414–78428. https://doi.org/10.1007/s11356-022-21263-1

Richardson RN, Butterworth JS (1995) The performance of propane/isobutene mixtures in vapour-compression refrigerant system. Int J Refrig 18:58–62. https://doi.org/10.1016/0140-7007(94)P3712-A

Prapainnop R, Suen KO (2012) Effects of refrigerant properties on refrigerant performance comparison: a review. Int J Eng Res Appl 2:486–493

Saidur R, Kazi SN, Hossain MS, Rahman MM, Mohamed HA (2011) A review on the performance of nanoparticles suspended with refrigerants and lubricating oils in refrigeration systems. Renew Sustain Energy Rev 15:310–323. https://doi.org/10.1016/j.rser.2010.08.018

Sajid MU, Bicer Y (2020a) Nanofluids as solar spectrum splitters: a critical review. Sol Energy 207:974–1000. https://doi.org/10.1016/j.solener.2020.07.009

Sajid MU, Bicer Y (2020b) Performance assessment of spectrum selective nanofluid-based cooling for a self-sustaining greenhouse. Energ Technol 9(1):2000875. https://doi.org/10.1002/ente.202000875

Sajid MU, Ali HM, Sufyan A, Rashid D, Zahid SU, Rehman WU (2019) Experimental investigation of TiO2–water nanofluid flow and heat transfer inside wavy mini-channel heat sinks. J Therm Anal Calorim 137(4):1279–1294. https://doi.org/10.1007/s10973-019-08043-9

Sarkar J, Bhattacharya S, Lal A (2013) Selection of suitable natural refrigerants pairs for cascade refrigeration system. Proc Inst Mech Eng: J Power Sources 227(5):612–622. https://doi.org/10.1177/0957650913487730

Saterlie MS, Sahin H, Kavlicoglu B, Liu Y, Graeve OA (2012) Surfactant effects on dispersion characteristics of copper-based nanofluids: a dynamic light scattering study. Chem Mater 24(17):3299–3306. https://doi.org/10.1021/cm203853f

Sekhar SJ, Lal DM (2005) HFC 134a/HC600a/HC290 mixture a retrofit for CFC12 systems. Int J Refrig 28:735–743. https://doi.org/10.1016/j.ijrefrig.2004.12.005

Selvaraj DA, Victor K (2021) Vapour absorption refrigeration system for rural cold storage: a comparative study. Environ Sci Pollut Res 28:34248–34258. https://doi.org/10.1007/s11356-020-11214-z

Shaik SV, Shaik S, Gorantla K (2020) Investigation on thermodynamic performance analysis and environmental effects of various new refrigerants used in air conditioners. Environ Sci Pollut Res 27:41415–41436. https://doi.org/10.1007/s11356-020-09478-6

Sharma DK, Sharma D, Ali AHH (2020) A state of the art on solar-powered vapor absorption cooling systems integrated with thermal energy storage. Environ Sci Pollut Res 27:158–189. https://doi.org/10.1007/s11356-019-06941-x

Spaushus HO (1988) CF-134a as a substitute refrigerant for CFC 12. Int J Refrig 11:389–392

Steven Brown J, Yana-Motta SF, Domanski PA (2002) Comparative analysis of an automotive air conditioning systems operating with CO2 and R134a. Int J Refrig 25:19–32. https://doi.org/10.1016/S0140-7007(01)00011-1

Subramani N, Prakash MJ (2011) Experimental studies on a vapour compression system using nanorefrigerants. Int J Eng Sci Technol 3(9):95–102. https://doi.org/10.4314/ijest.v3i9.8

Sun B, Yang D (2013a) Flow boiling heat transfer characteristics of nano-refrigerants in a horizontal tube. Int J Refrig 38:206–214. https://doi.org/10.1016/j.ijrefrig.2013.08.020

Sun B, Yang D (2013b) Experimental study on the heat transfer characteristics of nanorefrigerants in an internal thread copper tube. Int J Heat Mass Transf 64:559–566. https://doi.org/10.1016/j.ijheatmasstransfer.2013.04.031

Sundar LS, Ramanathan S, Sharma KV, Babu PS (2007) Temperature dependent flow characteristics of Ai2o3 nano fluid. Int J Nanotechnol Appl 1(2):35–44

Tan H, Xu L, Yang L (2022) Operation performance of an ultralow-temperature cascade refrigeration freezer with environmentally friendly refrigerants R290-R170. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-022-24310-z

Tang X, Zhao YH, Diao YH (2014) Experimental investigation of the nucleate pool boiling heat transfer characteristics of δ-Al2O3-R141b nanofluids on a horizontal plate. Exp Therm Fluid Sci 52:88–96. https://doi.org/10.1016/j.expthermflusci.2013.08.025

Tsai CY, Chien HT, Ding PP, Chan B, Luh TY, Chen PH (2004) Effect of structural character of gold nanoparticles in nano fluid on heat pipe thermal performance. Mater Lett 58:1461–1465. https://doi.org/10.1016/j.matlet.2003.10.009

Wang R, Wu Q, Wu Y (2010) Use of nanoparticles to make mineral oil lubricants feasible for use in a residential air conditioner employing hydro-fluorocarbons refrigerants. Energ Buildings 42:2111–2117. https://doi.org/10.1016/j.enbuild.2010.06.023

Wang JJ, Zheng RT, Gao JW, Chen G (2012) Heat conduction mechanisms in nano fluids and suspensions. Nano Today 7:124–136. https://doi.org/10.1016/j.nantod.2012.02.007

Wongwises S, Chimres N (2005) Experimental studies of hydrocarbon mixture to replace HFC134a in a domestic refrigerator. Energ Conver Manage 46:85–100. https://doi.org/10.1016/j.enconman.2004.02.011

Wu C, Cho TJ, Xu J, Lee D, Yang B, Zachariah MR (2010) Effect of nanoparticle clustering on the effective thermal conductivity of concentrated silica colloids. Phys Rev E 81(1):011406. https://doi.org/10.1103/physreve.81.011406

Yang K-S et al (2010) An investigation of a top-mounted domestic refrigerator. Energy Convers Manag 51:1422–1427. https://doi.org/10.1016/j.enconman.2010.01.016

Yang L, Du K, Niu X, Li Y, Zhang Y (2011) An experimental and theoretical study of the influence of surfactant on the preparation and stability of ammonia-water nanofluids. Int J Refrig 34(8):1741–1748. https://doi.org/10.1016/j.ijrefrig.2011.06.007

Yang D, Sun B, Li H, Fan X (2015) Experimental study on the heat transfer and flow characteristics of nanorefrigerants inside a corrugated tube. Int J Refrig 56:213–223. https://doi.org/10.1016/j.ijrefrig.2015.04.011

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.

Sādhanā (2023) 48:266
https://doi.org/10.1007/s12046-023-02341-7
© Indian Academy of Sciences

Synergistic effect of Al_2O_3–40%TiO_2 coating on thermal conductivity and corrosion rate of SS 304 substrate

DEEPAK KUMAR[1,*], PUSHPENDRA SINGH[1], QASIM MURTAZA[1] and R S WALIA[2]

[1]Department of Mechanical Engineering, Delhi Technological University, New Delhi 110042, India
[2]Department of Production and Industrial Engineering, Punjab Engineering College, Chandigarh 160012, India
e-mail: Deepak.kumar@dtu.ac.in

MS received 3 May 2023; revised 25 September 2023; accepted 6 October 2023

Abstract. The present research investigates the thermal and corrosive behaviour of alumina-based ceramic coatings deposited on stainless steel grade 304 through the flame spray method. The microstructure of the coatings was examined using various techniques such as FESEM, EDS, and a 3D optical profilometer. The thermal conductivity of the coatings was determined by employing the hot plate method to calculate the temperature difference across the exposed coatings. Furthermore, an electrochemical test was conducted to evaluate the corrosion behaviour of the deposited coatings. The outcomes of the study reveal that the Al_2O_3–40%TiO_2 coating exhibits significantly higher thermal and corrosive resistance characteristics in comparison to the stainless-steel substrate.

Keywords. Thermal conductivity; corrosion rate; Al_2O_3–40%TiO_2; flame spray; coatings.

1. Introduction

The need for highly functional machinery parts is much desired for the excellent performance of the system. The part's durability, corrosive resistance and thermal stability are some of the systems' main properties that could be considered for high performance. The issue of poor thermal efficiency and corrosion parts has been a long-standing challenge in the industrial sector and continues to be a significant problem today [1].

Thermal conductivity is a critical characteristic that defines a material's thermal behaviour. In terms of coating, it is an important property that determines the ability of a coating material to transfer heat. For example, surface modifications have been made in steam gas turbine engines to reduce heat transfer, making them suitable for high-temperature environments. Furthermore, the costs incurred due to corrosion are significant, often amounting to billions of dollars for developed nations. The oil and gas industry is particularly vulnerable to corrosion, as equipment and devices are constantly exposed to corrosive environments, resulting in substantial costs for this sector [2, 3]. Therefore, improving the thermal efficiency and corrosion resistance of products can significantly reduce the costs associated with inspection and maintenance. Therefore, it is essential to take anticipatory actions to minimise the occurrence of corrosion. Additionally, it is important to ensure that the coating material used has an appropriate thermal conductivity for its intended application.

To improve the performance and characteristics of the materials, thermal spraying coatings act as a promising technology. Among various protective coatings, thermal barrier coatings (TBCs) like ceramics, carbides and many more have gotten attention and are widely used to enhance surface properties and result in an increase in the working cycle of the engine components. It is the most promising technology, and it has attracted much attention because of its thermal, corrosion and wear resistance properties studied [4]. TBCs have been censorious in enhancing the proficiency of modern propulsion and gas turbines associated with energy production [5].

Ceramic coatings deposited by different methods are widely used as a protective layer to surface exposed. In addition, Al-Ti coatings are known as protective layers for industrial parts due to their excellent corrosion resistance and thermal stability of coatings [6]. Ceramics made from Al_2O_3–TiO_2 composite powder are becoming increasingly popular, as they exhibit excellent heat and corrosion resistance, favourable fracture toughness and moderate frictional coefficients. Aluminium titanate ceramic composites are particularly significant due to their superior abrasion, heat, and corrosion resistance properties [7, 8]. However, alumina exhibits superior mechanical properties, such as hardness and toughness, compared to aluminium titanates. This means that increasing the amount of TiO_2 in the feedstock powder can decrease the hardness of alumina-titania coatings [9, 10].

*For correspondence
Published online: 23 November 2023

The hot plate guarded method is used for thermal conductivity for different applications [11–13]. Tan *et al* [14] studied the steady-state heat flow method and evaluated the thermal conductivity of thick alumina coatings of around 100 µm. Wondu *et al* [15] studied that the addition of alumina particles resulted in a substantial enhancement in the thermal conductivity of the composite material. The corrosion rate of aluminium composites in an atmosphere encompassing 3.5 wt.% NaCl at room temperature has been widely examined through electrochemical corrosion studies conducted by numerous researchers [16–18]. In the presence of a 3.5% sodium chloride (NaCl) solution, Al_2O_3 coatings exhibit excellent corrosion resistance properties when compared to the uncoated mild steel substrate [19]. Shen *et al* [20] investigated the electrochemical corrosion resistance of uniform TiO_2 coatings applied to steel surfaces. Their findings reveal that TiO_2 particle coatings, resembling ceramic protective films, exhibit outstanding corrosion resistance when exposed to NaCl solution. Researchers reported that adding TiO_2 in Al_2O_3 decreases the thermal conductivity and improves the corrosion-resistant properties of the material [21, 22]. Therefore, Al_2O_3–TiO_2 composite coatings hold the potential for effective anticorrosion protection and thermal conductive properties. This is because they combine the beneficial properties of both materials, namely thermal stability and favourable corrosion resistance of TiO_2, and the outstanding sealing properties of Al_2O_3.

The novelty of the current research is addressing crucial challenges in machinery performance by exploring aluminium-based ceramics, and the low-cost deposition method is used to improve corrosion resistance and thermal stability in industrial applications. As there are few comparative studies available on the effect of Al_2O_3–40%TiO_2 coating thermal conductivity and corrosion rate, the current research aims to investigate these properties of the thermally flame-sprayed Al_2O_3–40%TiO_2 composite coating on a stainless-steel substrate (grade 304).

2. Material and methods

2.1 *Coating deposition*

The study aimed to investigate the thermal conductivity and corrosion rate of a ceramic powder, specifically the commercially available Al_2O_3–40%TiO_2, for deposition on a base material sample made of SS 304 with a diameter of 100 mm and a thickness of 3 mm. The chemical composition of the substrate SS 304 is given in table 1. Before coating, the exposed surface of the sample was grit blasted with alumina grit to increase surface roughness and improve coating adhesion. Table 2 provides the coating parameters used for depositing the composite powders. Eight passes are performed to apply coatings using a manipulator speed of 1 millimetre per second. The deposition of coating occurs with an approximate step size of 20 microns for each pass. Figure 1 depicts schematic diagrams of the thermally flame spray process.

To analyse the morphology of the coating powder used for deposition, a finite electron scanning electron microscope (FESEM) from Carl Zeiss at the Indian Instrumentation Centre in the Indian Institute of Technology,

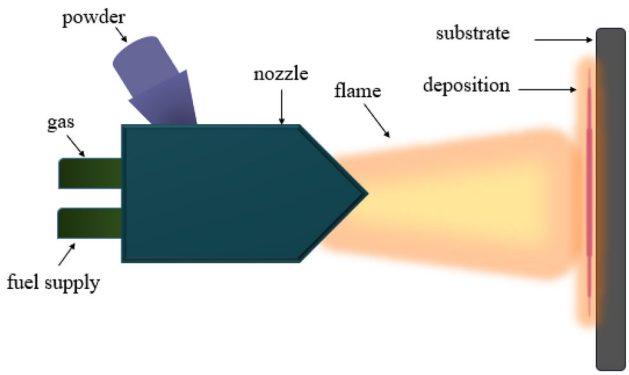


Figure 1. Flame spray process.

Table 1. The chemical composition of the base material (steel of grade 304 wt.%) [23].

Grade	Cr	Ni	Mn	N max.	S max.	C max.	Si max.	P max.	Fe
304	18–20	8–10	2	.10	.03	.08	.75	.045	Balanced

Table 2. Coating parameter used for flame spray deposition.

Flow-rate (Oxygen)	Flow-rate (Acetylene)	Spraying distance	Nitrogen carrier flow rate	Oxygen pressure	Acetylene pressure	Nitrogen carrier pressure
50 NLPM*	25 NLPM	70 mm	7 NPLM	2.5 bar	2.2 bar	4.5 bar

*NLPM-Normal litre per minute

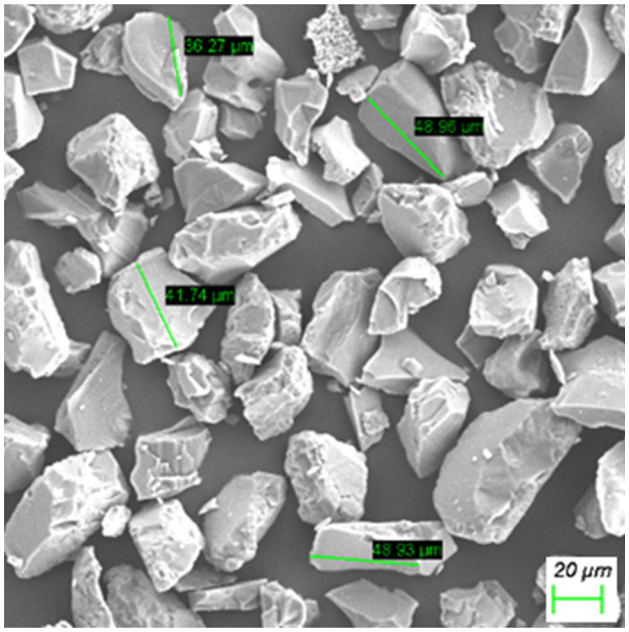


Figure 2. FESEM micrograph of Al_2O_3–40%(TiO_2) coating powder.

Roorkee, India, was employed. The electron dispersive X-ray spectroscopy (EDS) associated with FESEM was also conducted to confirm the presence of the desired element in the composite powders. The FESEM micrographs of Al_2O_3–40%(TiO_2) powder were presented in figure 2, and it depicts that the particles of the coating powder have sharp elongated edges. The size of particles was in the range of 15–50 µm. EDS analysis of the coating powder shown in figure 3 confirms the presence of Al, O_2 and Ti elements on the SS 304 substrate.

2.2 *Experimentation setup*

To determine the thermal conductivity of the coated substrate, a hot plate guarded setup was utilised at the Precision and Manufacturing Lab, New Delhi Technological University, New Delhi, India. The setup consisted of a heater, a sample plate, a top lid, and electrical components. The sample was placed on the heater, and the top lid was secured to prevent any heat loss. The entire system was enclosed in an insulated chamber to protect it from external factors. A schematic diagram of the enclosed chamber

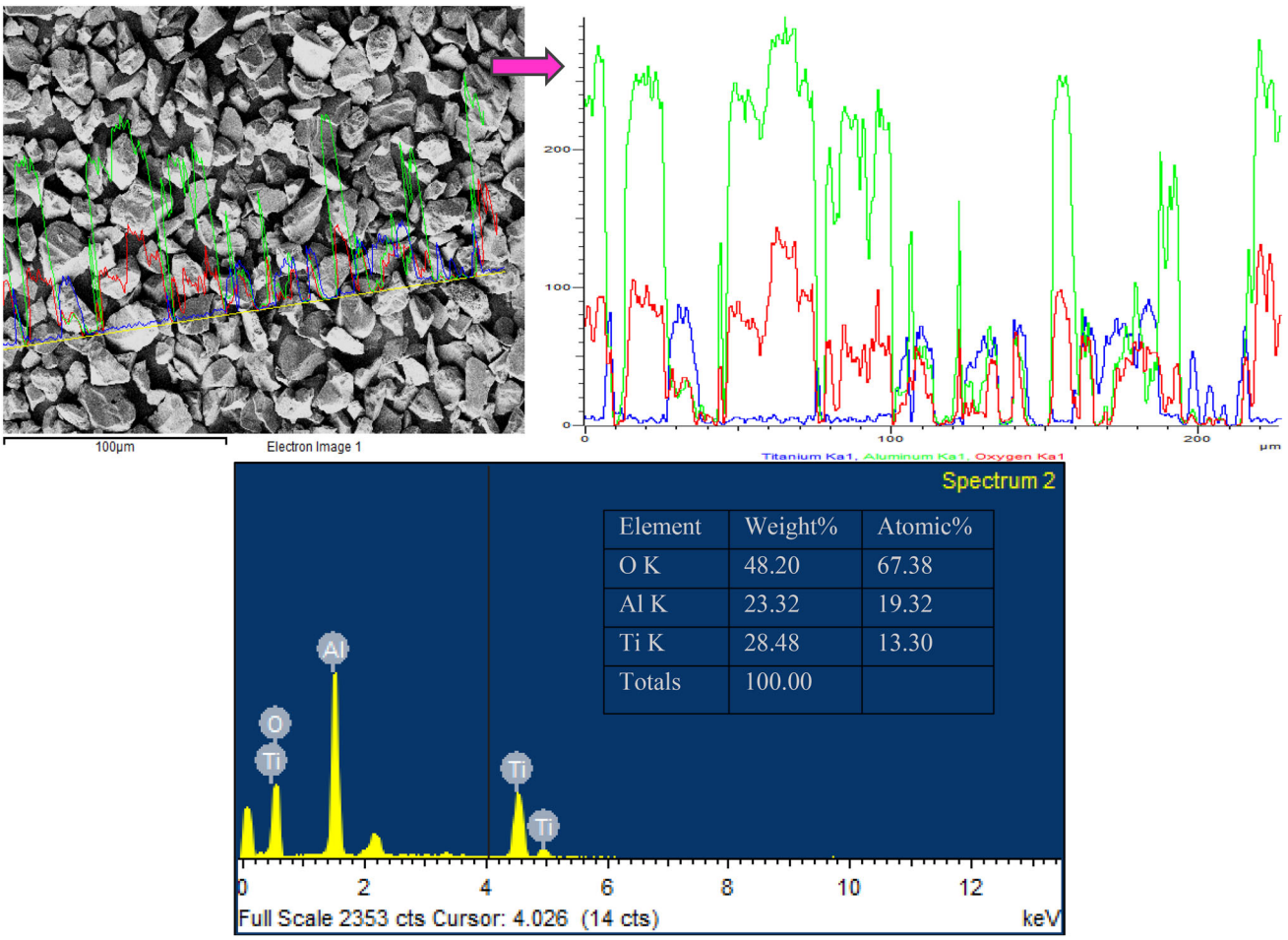


Figure 3. EDS analysis of Al_2O_3–40%(TiO_2) coating powder.

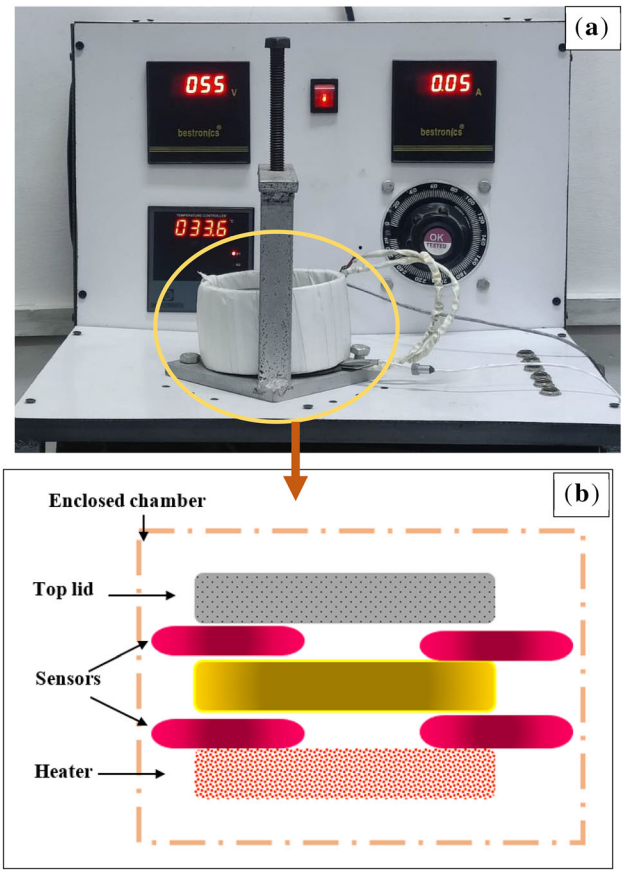


Figure 4. (**a**) Experimental hot plate guarded setup and (**b**) Schematic diagram of an enclosed chamber.

Table 3. The obtained data from the setup for the temperature difference.

Sl. Nos.	Voltage (V)	Current (A)	Temperature difference (K)
1	90	0.12	0.48
2	100	0.14	0.6
3	110	0.15	0.64

thermal conductivity setup used in this study is provided in figure 4.

To perform the experiment, the current supply of the setup was switched on, and the voltage was adjusted using a variable autotransformer. This caused a fixed value of current to flow through the ammeter, and the test piece began to heat up slowly. During this initial stage, the temperature of the test piece was in an unsteady state, and it took approximately one hour to reach a steady state. Once the steady state was achieved, the temperature values were recorded from sensors placed at the top and bottom of the substrate. The obtained data from the setup can be found in table 3.

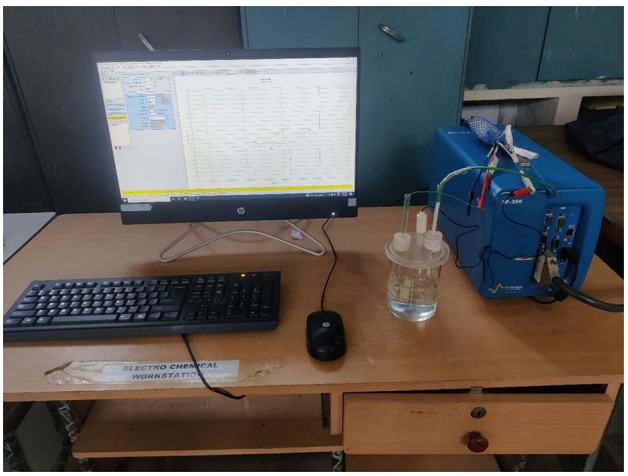


Figure 5. Electrochemical corrosion setup.

To ensure the accuracy and consistency of the data collected from the setup, the experiment was conducted three times. The thermal conductivity data gathered from this experiment can provide insights into the effectiveness of the ceramic coating in reducing heat transfer and improving thermal insulation.

The electrochemical corrosion behaviour of Al_2O_3–40%(TiO_2) composite coating was studied using Biologic Potentiostat Electrochemical Corrosion (Model SP 100) set up in 3.5% NaCl aqueous solution available at Particulate Metallurgy Laboratory facility PEC Chandigarh, India as shown in figure 5. The corrosion behaviour of the coating samples was investigated using a potentiodynamic polarisation technique employing three electrode configurations. The configured setup consists of a reference electrode considered saturated calomel, a counter electrode considered platinum mess (1 cm^2), and a working electrode considered a coating sample. Tafel plots were generated and analysed for corrosion rate using potentiostat by EC lab software. The test was repeated three times to ensure the reliability of the data obtained from the setup.

3. Results and discussions

3.1 *Coatings characterisation*

The coating was characterised using FESEM micrographs, which are shown in figure 6. The micrograph reveals that the melting of the powder was almost perfect, with only a few instances of partially melted powder. The surface of a thermally sprayed coating is made up of lamellae that are formed from molten and partially molten materials, and many lamellae are piled up on top of each other. During the coating deposition process, a few micro-cracks and pores were also observed. By examining the FESEM micrograph,

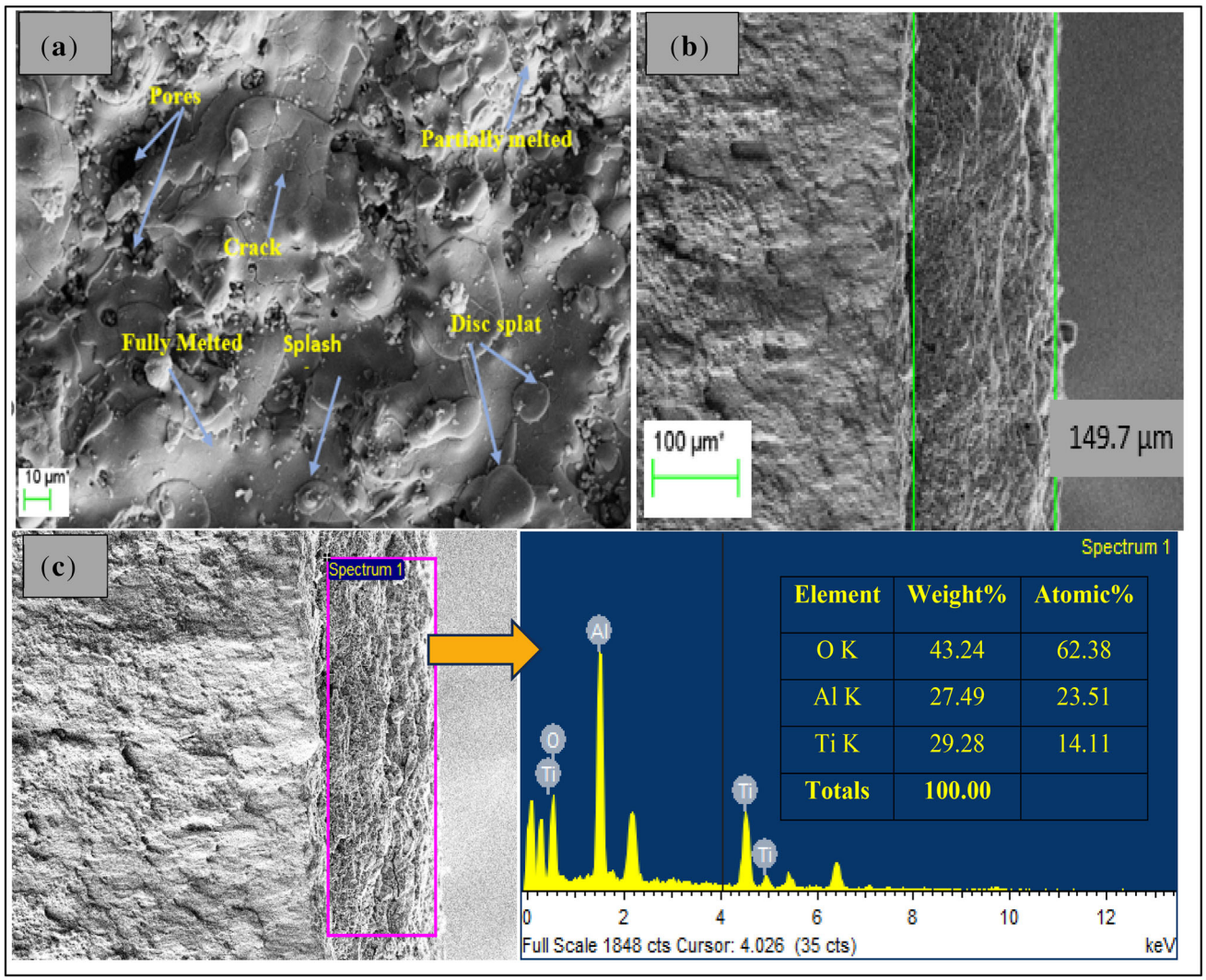


Figure 6. FESEM micrograph: (**a**) deposited coating, (**b**) coating thickness and (**c**) EDAX analysis of the coating thickness.

different types of splats were identified on the coated surfaces, including as-sprayed splat, splash splat, and disc splat. These splats indicate that the coating was properly applied and that the sprayed material bonded uniformly. The micrograph also shows flattened and solidified droplets, as well as cracks caused by tensile quenching stress resulting from the large temperature drop during coating.

The microstructure of the coating indicates that the splat of the sprayed material did not form a continuous layer, but when examined at the cross-section, the coating was observed to be more homogeneous and regular. The spraying torch moves with respect to the substrate, and the coating grows more slowly. In fact, the number of lamellae deposited during one torch pass depends on the velocity at which the torch moves over the substrate. Figure 6c illustrates the EDAX analysis of the deposited coating thickness, which authenticates the presence of the deposited element in the coating thickness.

Figure 7 displays the elemental mapping results of the deposited coating on the substrate. The maps illustrate a uniform distribution of elements across the entire surface, as shown by the presence of different colours. Red represents oxygen distribution, green represents aluminium, and yellow represents titanium. These results suggest that the deposition process effectively covered the substrate with the ceramic coating, resulting in a homogenous distribution of elements throughout the surface. To further investigate the surface characteristics of the coated substrate, a 3-D profilometer was used to obtain topographical data. Figure 8 shows 3D micrographs obtained from three different locations, and the data extracted from the micrographs reveal that the average surface roughness of the deposited coating was 1.8 µm.

The results of the elemental mapping and surface roughness measurements are important in determining the quality and effectiveness of the coating. A uniform

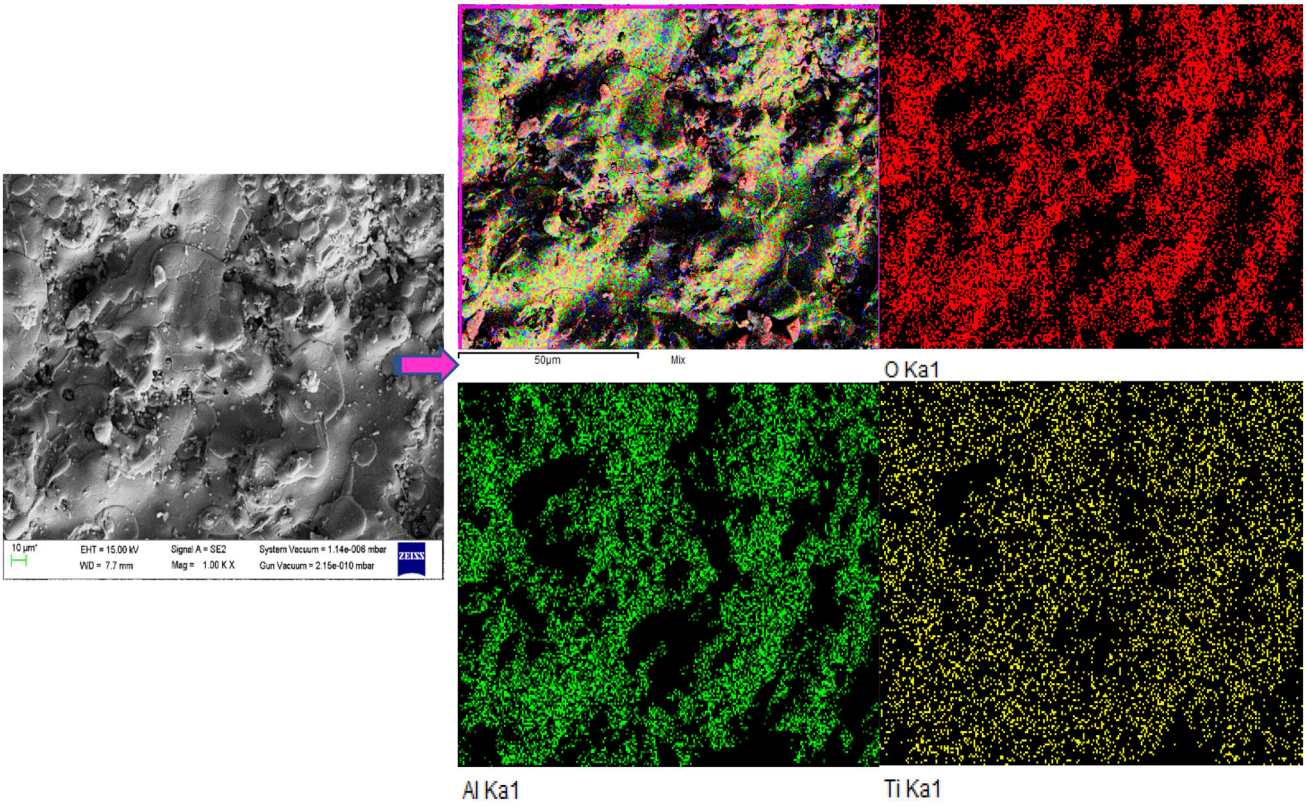


Figure 7. Elemental mapping of deposited coating.

distribution of elements and a low surface roughness indicate a well-formed and adherent coating, which is crucial for enhancing the corrosion resistance and mechanical durability of the substrate [24]. Additionally, surface roughness can influence the heat transfer behaviour of the coated substrate, making it important to measure in applications where thermal conductivity is a critical factor [25].

3.2 *Effect of coating on thermal conductivity*

To analyse the heat transfer in a disc, longitudinal heat flow is assumed to be the dominant mechanism, while radial and tangential heat loss is neglected due to the disc's small thickness. The Fourier law, described by Equation (1), is employed to determine the amount of heat (Q) transferred, which is proportional to the thermal conductivity (K) and the temperature gradient (dT/dx) in the flow direction.

$$Q = KA(dT/dx) \tag{1}$$

Equation (1) can be simplified to Equation (2),

$$Q = \Delta T/R \tag{2}$$

The thermal resistance (R) to heat flow in a plate is determined by the ratio of its thickness (L) to the product of its cross-sectional area perpendicular to heat flow (A) and the thermal conductivity (K). The temperature difference (ΔT) between the two surfaces of the plate is also taken into account, and convective and radiative heat losses are ignored. For a composite material consisting of a substrate and a coating, the thermal resistance of the composite is equal to the sum of the thermal resistances of the substrate and the coating. To calculate the thermal resistance of the composite, an electrical model is utilised that provides the rate of electrical energy produced by multiplying the instantaneous potential difference and current values measured by the setup. The required temperature difference is achieved by measuring the temperature of the top and bottom surfaces of the coating. The thickness of the substrate was measured to be 3 mm, and the coating thickness was determined as 149.7 µm using FESEM (as shown in figure 6b). The average temperature difference between the two surfaces of the composite was calculated as 0.57 K. With the help of Equation (2), the overall thermal resistance (i.e., the resistance of composite, Rc) was found to be 0.0437 K/W. Consequently, the thermal conductivity of the coated substrate was determined as 10.031 W/mK. Figure 9 shows a graph comparing the measured thermal conductivities of the substrate and the coated substrate.

The reduction in thermal conductivity observed upon applying a ceramic coating on the SS 304 substrate has important implications for the performance and durability of

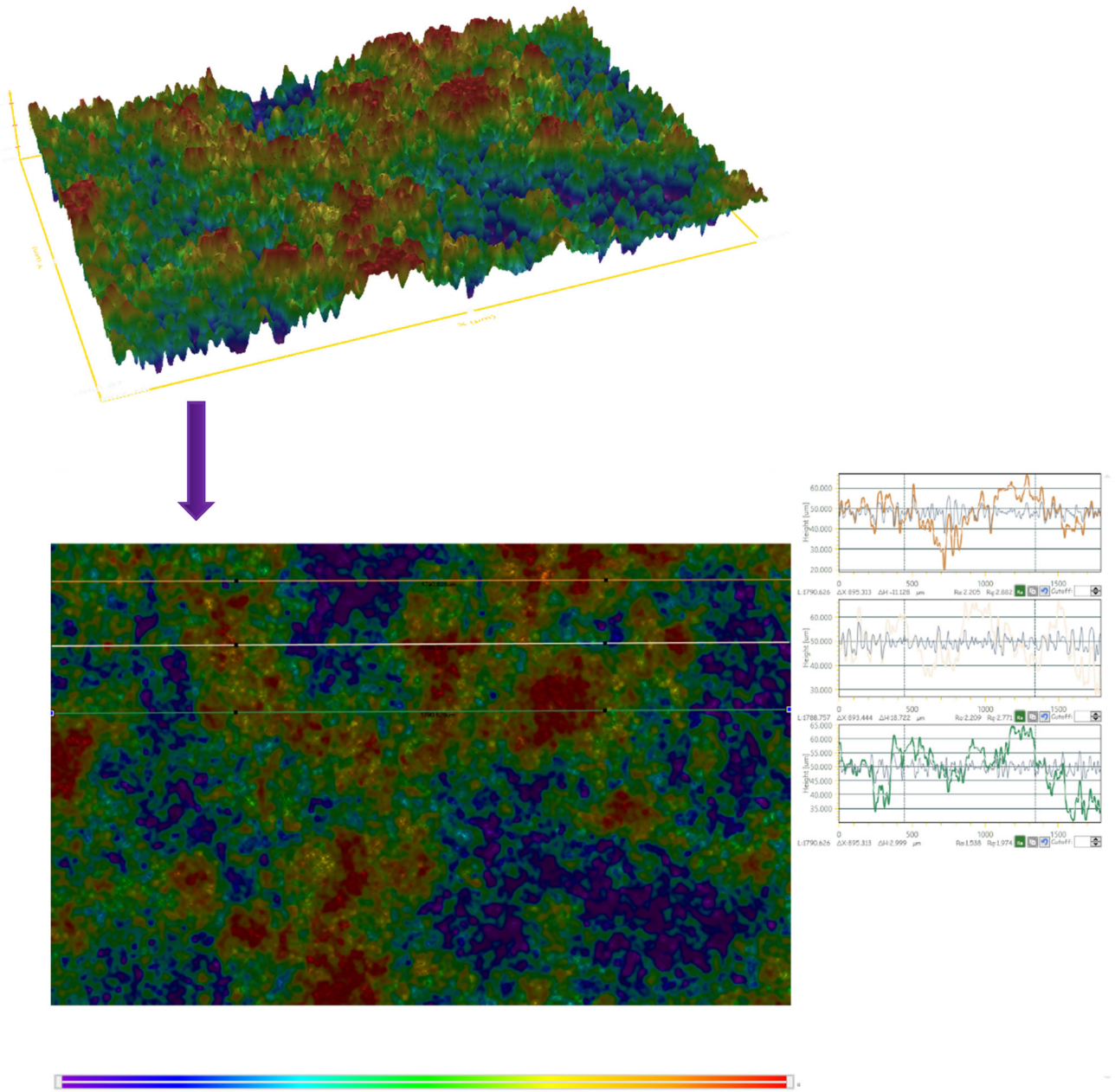


Figure 8. 3D profilometer analysis of deposited coating.

high-temperature devices. The reported thermal conductivity of the SS 304 substrate, which is widely used in various industrial applications, is 16.3 W/mK [26]. On the other hand, the measured thermal conductivity of the coated substrate was found to be 10.031 W/mK, which is a significant reduction of 38% compared to the substrate without the coating.

This decrease in thermal conductivity can be attributed to the presence of fine pores in the ceramic coating, which create a porous structure within the coatings. These fine pores act as insulating voids, interrupting the flow of heat through the sample. The insulating effect of porosity becomes more pronounced as the pore volume fraction increases [27, 28]. This effect can be particularly advantageous in high-temperature devices, such as steam pipes and the top of IC engine pistons, where reducing heat loss and improving thermal resistance are critical for enhancing efficiency and durability. Moreover, the benefits of the ceramic coating can be further enhanced if the thickness of the coating exceeds the critical thickness, as this can lead to even more excellent thermal insulation and protection against corrosion and other forms of degradation.

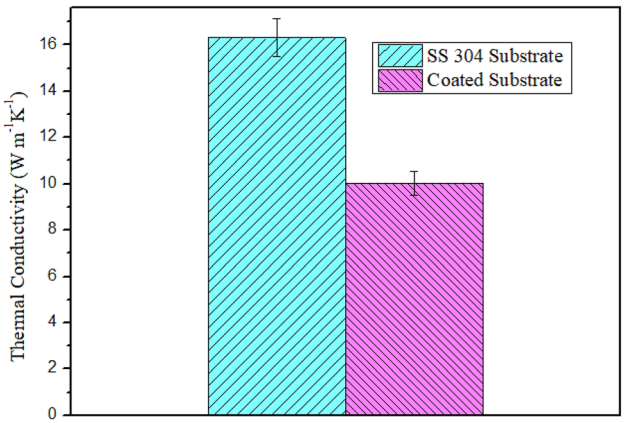


Figure 9. Thermal conductivity.

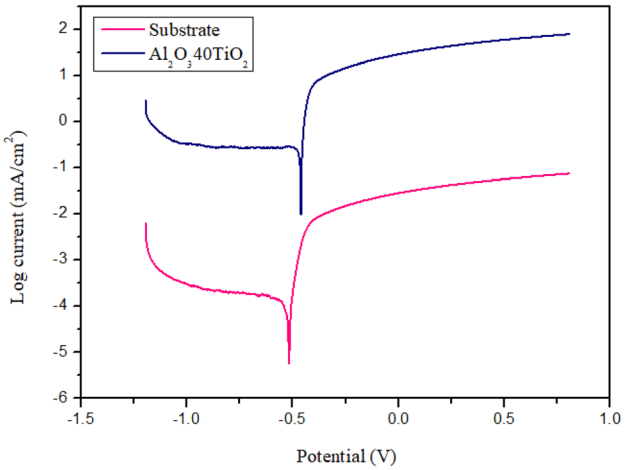


Figure 10. Tafel plot.

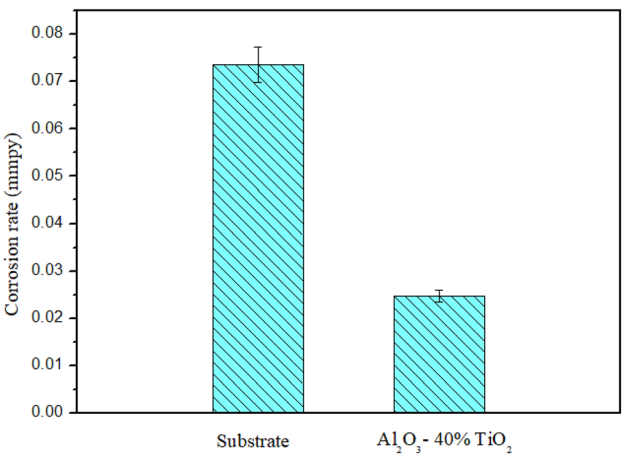


Figure 11. Corrosion rate.

3.3 *Effect of coating on electrochemical corrosion rate*

The electrochemical corrosion test was performed on the coated sample. The coated sample and electrodes were placed in a glass cell with the sample and electrodes sufficiently immersed in the solution, facing each other at the same level. Only the coated surface side of the sample has been faced towards the electrodes for analysing the electrochemical corrosion test. Initially, the open circuit potential (OCP) plot was recorded for 30 minutes to allow the sample to stabilise in the provided corrosion medium. Then, a Tafel plot was generated at a scan rate of 1 mV/s. The Tafel plot was fitted using Tafel fit in EC Lab software, and the corrosion potential (Ecorr.) and corrosion current (Icorr.) were calculated by filtering the polarisation curve. These electrochemical parameters were then used in Equation (3) to determine the corrosion rate of the coated sample. Each test study was performed at least three times with a fresh 300 ml electrolyte solution prepared for each study in a 3.5 wt.% NaCl solution to ensure reproducibility of results. Here, Icorr. represents the corrosion current density ($\mu A/cm^2$), EW denotes the equivalent weight (g), and d represents the density of the sample (g/cm^3).

$$\mathrm{CR(mmpy)} = (0.00327 * \mathrm{Icorr.} * \mathrm{EW})/\mathrm{d} \qquad (3)$$

The electrochemical corrosion test was conducted on the deposited coatings on SS substrates using a 3.5 wt.% NaCl solution. Tafel plots were generated for each coating, as shown in figure 10, and were then used to evaluate the corrosion rate in millimetres per year (mmpy) of each coating by tafel fitting in EC lab software. The measured corrosion rate of the deposited coating was found to be 0.0385 mmpy, while that of the substrate was 0.082 mmpy, as shown in figure 11. It is evident that after deposition, the corrosion rate decreased by 44% with respect to the substrate.

In figure 10, the tafel plot demonstrates the effect of deposited coatings on the corrosion behaviour of the coating. It is observed that the corrosion current density (Icorr.) decreases as compared to the substrate. Moreover, the Al_2O_3–40%TiO_2 coating has a low corrosion rate (mmpy), indicating that the coating with 40%TiO_2 exhibits more corrosion-resistant characteristics, which further resists the movement of ions through itself and provides passivation. This passive layer formation over the surface acts as a protective layer that resists the ions' loss and provides the desired corrosion resistance characteristics to the substrate.

These findings are consistent with previous studies that have reported improved corrosion resistance in SS substrates by applying ceramic coatings [29, 30]. Therefore, the results of this study suggest that the Al_2O_3–40%TiO_2 coating can serve as an effective protective barrier against corrosion in various applications.

4. Conclusion

This paper uses the flame spray method to describe the application of an Al_2O_3–40 wt.% TiO_2 composite coating onto an SS 304 grade steel substrate. The resulting coating exhibited good adhesion and density, with a thickness of 149.7 μm. The microstructure of the coating was observed to be composed of stacked lamellae with few micropores. The element maps depict the perfect distribution of elements over the surface of the substrate. The average surface roughness of the deposited coating was 1.8 μm.

The thermal conductivity of the coated substrate was 43% lower than that of the SS 304 substrate, with a calculated value of 9.18 W/mK. The value of thermal conductivity of the coated substrate becomes lower because of the fine pores of the coating being porous, which makes heat flow inhibit and gives thermal insulation.

The corrosion rate of the coating was found to be 0.0385 mmyp, compared to 0.082 mmyp for the substrate. Tafel plot analysis demonstrated that the corrosion current density (Icorr.) decreased when compared to the substrate, indicating a superior corrosion resistance for the Al_2O_3–40%TiO_2 coating. The presence of 40% TiO_2 in the coating led to the formation of a passive layer, which further inhibited the movement of ions through the coating and provided additional corrosion resistance to the substrate.

The Al_2O_3–40% TiO_2 coating demonstrates corrosion resistance properties and low thermal conductivity, making it suitable for both corrosion protection and thermal insulation applications. This combination of properties allows it to provide effective insulation against heat transfer while protecting the underlying substrate from wear and corrosion. Therefore, it is an ideal choice for applications where both thermal insulation and protective coating properties are required, such as in high-temperature environments or equipment exposed to harsh corrosive conditions, such as exhaust systems and engine components in the automotive industry, pipelines and storage tanks in oil and gas industry, power generation facilities like thermal power plants, nuclear reactor and boilers.

References

[1] Hardwicke C U and Lau Y C 2013 Advances in thermal spray coatings for gas turbines and energy generation: a review. *J. Therm. Spray Technol.* 22(5): 564–576

[2] Zavareh M A, Sarhan A A D M, Razak B B A and Basirun W J 2014 Plasma thermal spray of ceramic oxide coating on carbon steel with enhanced wear and corrosion resistance for oil and gas applications. *Ceram. Int.* 40(9): 14267–14277

[3] Popoola L T, Grema A S, Latinwo G K, Gutti B and Balogun A S 2013 Corrosion problems during oil and gas production and its mitigation. *Int. J. Ind. Chem.* 4(1): 1–15

[4] Yao H, Wang L, Zhou Z, Wang B, Tan Z, He D and Xue Y 2020 Thermal transport property correlated with microstructural evolution of Fe-based amorphous alloy. *Acta Mater.* 200: 793–802

[5] Li G R, Wang L S and Yang G J A 2019 Novel composite-layered coating enabling self-enhancing thermal barrier performance. *Scr. Mater.* 163: 142–147

[6] Xie X, Hosni B, Chen C, Wu H, Li Y, Chen Z, Verdy C, Kedim OE, Zhong Q, Addad A, Coddet C, Ji G and Liao H 2020 Corrosion behavior of cold sprayed 7075Al composite coating reinforced with TiB2 nanoparticles. *Surf. Coatings Technol.* 404: 126460

[7] Schütze M, Malessa M, Rohr V and Weber T 2006 Development of coatings for protection in specific high temperature environments. *Surf. Coat. Technol.* 201(7): 3872–3879

[8] Berndt M L and Berndt C C 2003 Thermal spray coatings. *Corros. Fundam. Test. Prot.* 13: 803–813

[9] Vargas F, Ageorges H, Fournier P, Fauchais P and López M E 2010 Mechanical and tribological performance of Al_2O_3–TiO_2 coatings elaborated by flame and plasma spraying. *Surf Coat. Technol.* 205(4): 1132–1136

[10] Łatka L, Szala M, Nowakowska M, Walczak M, Kiełczawa T and Sokołowski P 2023 The effect of microstructure and mechanical properties on sliding wear and cavitation erosion of plasma coatings sprayed from Al_2O_3 + 40 wt% TiO_2 agglomerated powders. *Surf. Coat. Technol.* 455: 129180

[11] Heisig L M, Wulf R and Fieback T M 2023 Investigation and optimization of the hot disk method for thermal conductivity measurements up to 750 °C. *Int. J. Thermophys.* 44(6): 1–21

[12] ASTM C1044-12 - Standard Practice for Using a Guarded-Hot-Plate Apparatus or Thin-Heater Apparatus in the Single-Sided Mode. Accessed September 24, 2023. https://webstore.ansi.org/standards/astm/astmc104412

[13] Wyczółkowski R, Bagdasaryan V, Gała M and Król P A 2022 The review of chosen methods used to investigate heat transfer in a steel porous charge. *Energies* 15(6): 2266

[14] Tan Y, Longtin J P and Sampath S 2006 Modeling thermal conductivity of thermal spray coatings: comparing predictions to experiments. *Proc. Int. Therm. Spray Conf.* 15: 545–552

[15] Eyob W Z L and Kim J 2017 Thermal conductivity and mechanical properties of thermoplastic polyurethane-/silane-modified Al_2O_3 composite fabricated via melt compounding eyob. *Polymers* 11(1103): 1–12

[16] Deshpande M, Vagge S T, Narayan M S V S, Kale H and Gondil R 2019 Studies on electro-chemical corrosion of carbon fiber reinforced aluminium alloy AA7075. *Sadhana Acad. Proc. Eng. Sci.* 44(11): 1–10

[17] Abdo H S, Samad U A, Abdo M S, Alkhammash H I and Aijaz M O 2021 Electrochemical behavior of inductively sintered Al/TiO_2 nanocomposites reinforced by electrospun ceramic nanofibers. *Polymers* 13(24): 4319

[18] Ikubanni P, Oki M, Adeleke A, Adesina O, Omoniyi P and Akinlabi E 2022 Electrochemical studies of the corrosion behavior of Al/SiC/PKSA hybrid composites in 3.5% NaCl solution. *J. Compos. Sci.* 6(10): 1–13

[19] Ruhi G, Modi O P, Sinha A S K and Singh I B 2008 Effect of sintering temperatures on corrosion and wear properties of sol–gel alumina coatings on surface pre-treated mild steel. *Corros. Sci.* 50(3): 639–649

[20] Shen G X, Chen Y C and Lin C J 2005 Corrosion protection of 316 L stainless steel by a TiO_2 nanoparticle coating prepared by sol–gel method. *Thin Solid Films.* 489(1): 130–136

[21] Kwon H, Yoo Y W, Park Y, Nam U H and Byon E 2023 Effect of TiO_2 on mechanical and thermal properties of Al_2O_3-based coating via atmospheric plasma spraying. *J. Asian Ceram. Soc.* 11(2): 282–290

[22] Singh G, Kumar S and Kumar R 2020 Comparative study of hot corrosion behavior of thermal sprayed alumina and titanium oxide reinforced alumina coatings on boiler steel. *Mater. Res. Express*. 7(2): 026527

[23] Kumar D, Murtaza Q, Walia R S and Singh P 2022 Comparative investigation on thermally sprayed Al_2O_3, Al_2O_3-13%(TiO_2) and Al_2O_3–40%(TiO_2) composite coatings from room to 400 °C temperature. *Surf. Topogr.: Metrol. Prop.* 10(1): 015043

[24] Ramanan G, Suresh S, Prabha N R and Dhas J E 2019 Characterization, wear surface roughness and tensile failure analysis of Al7075-TiC-MoS2 hybrid composites using online acoustic emission. *Mater. Res. Express* 6(6): 066544

[25] Zhai W, Bai L, Zhou R, Fan X, Kang G, Liu Y and Zhou K 2021 Recent Progress on Wear-Resistant Materials: Designs, Properties, and Applications. *Adv. Sci*. 8(11): 1–29

[26] Yadav R, Sah R K, Mann P, Kumar D and Singh P 2022 Study of thermal conductivity of NiCrFeSi based ceramic composite coating. *Optim. Ind. Syst.* 42: 545–553

[27] Shukla R K, Kumar A, Kumar R, Singh D and Kumar A 2019 Numerical study of pore formation in thermal spray coating process by investigating dynamics of air entrapment. *Surf. Coat. Technol.* 378: 124972

[28] Kumar D, Walia R S, Singh P and Murtaza Q 2023 Tribological behaviour of thermally sprayed ceramic coating at high temperature. *Int. J. Mater. Res.* 114(9): 773–782

[29] Sreekumar Rajesh T and Venkata Rao R 2018 Experimental investigation and parameter optimisation of Al_2O_3–40% TiO_2 atmospheric plasma spray coating on mild steel substrate. *Int. J. Mater. Eng. Innov.* 9(4): 257–278

[30] Mishra N K, Rai A K, Mishra S B and Kumar R 2014 Hot corrosion behaviour of detonation gun sprayed stellite-6 and stellite-21 coating on boiler steel SAE 431 at 900°C. *Int. J. Corros.* 2014: 1–4

Indian Journal of Chemical Technology
Vol. 30, November 2023, pp. 753-764
DOI: 10.56042/ijct.v30i6.1325

Synthesis, characterization and application of Lignosulphonate-g-poly(sodium acrylate) hydrogel

Manu, Rajinder K. Gupta* & Devendra Kumar

Department of Applied Chemistry, Delhi Technological University, Delhi, India

*E-mail: rkg67ap@yahoo.com

Received 9 May 2023; accepted 9 September 2023

Natural polymer-based hydrogels are of great interest to research community owing to their inherent characters of environment friendliness and biodegradability. Current work aims to synthesize lignosulfonate grafted sodium acrylate hydrogel (LS-g-SAH) and investigate its application in urea release behaviour. The hydrogel has been characterized by different techniques. The release kinetics has been analyzed by using a UV-visible spectrophotometer. The optimized composition of lignosulfonate, KPS, and N,N'-MBA has shown the highest water absorbency of 560 g g^{-1} in distilled water. The equilibrium swollen LS-g-SAH 12 hydrogel has slowly released 60% of loaded urea in 24 h and followed first-order release kinetics. Soil treatment with hydrogel has shown a significant effect in reducing the water evaporation rate. It also improved the seed germination and average height of wheatgrass. The synthesized LS-g-SAH is, thus, expected to have potential application in modern sustainable agriculture.

Keywords: Hydrogel, Lignosulphonate, Release kinetics, Sustainable agriculture, Urea release

As world's seventh largest country, India has approximately 157 million hectares of arable land, accounting for 53.2% of the total area. The land is continuously degraded by different means, thus reducing the total production potential to less than 20 percent. Therefore immediate attention is required in upgrading soil fertility, which caused less productivity due to low water retention capability, poor structure, high infiltration rate, low clay and humus content, and loss of agrochemicals through deep percolation or leaching[1].

A super absorbent crosslinked polymer known as hydrogel is a potential material to solve soil-fertility problems. Hydrogel is a 3D network structure composed of chemically or physically crosslinked polymers. It can hold a massive volume of water due to its swollen state's surface tension and capillary forces. The rate of water absorption by hydrogel is affected by the functional group and the density of the crosslinking network[2]. Natural biodegradable polymers should be used to develop economical agricultural water-absorbent materials. Nowadays, researchers are paying great attention to lignocellulosic polymeric material as a favourable, sustainable, and large-scale asset to be used as a chemical component in various syntheses. In this regard, lignin is a renewable feedstock readily available in an adequate amounts[3]. Lignin accounts for 16-33% of the biomass and is predominantly present in the secondary cell walls of terrestrial plants[4]. Lignin comprises three types of phenylpropane monomers known as monolignols. These monolignols are a) coniferyl alcohol, b) sinapyl alcohol, and c) p-coumaryl alcohol[5]. This inexpensive waste of pulp manufacturing process is the only scalable renewable raw material that consists of aromatic moieties[6]. It is also easily accessible, affordable, and underutilized today [7].

Also, a lignin hydrogel based study by Bukhard et al. showed that the synthesized hydrogels could degrade up to 14% in just 60 days as compared with acrylamide-based hydrogels, indicating degradation of just 2% in 120 days[8]. Lignin has antimicrobial, antioxidant, and stabilizer properties[9]. It has also been used for metal elimination. For example, graphene hydrogel based on lignosulfonate, manufactured in a column system, rapidly eliminated Pb^{2+} from the aqueous solution (Thakur et al., 2017). A free-standing lignin-based film with superior electrochemical performance was manufactured by using poly(N-vinyl imidazole)-co-poly(poly(ethylene glycol) methyl ether methacrylate)[6]. However, there has been little exploration of using lignin-based hydrogel for controlled-release applications in

agriculture. Song et al. in 2020 synthesized novel lignin, konjac flour, and sodium alginate hydrogels, which proved to be efficient when tested for the growth of tobacco plants in drought-like conditions. They exhibited high water retention capacity and reduced soil nutrient leaching [11]. Urea is one of the most commonly used nitrogen fertilizers globally. But due to its leaching losses, its use efficiency is relatively low (on average 30-35%). It is a waste of resources that also results in the pollution of the environment. Slow-release urea through hydrogel is an effective way to address this issue[12]. In this work, Lignosulfonate grafted poly(sodium acrylate) hydrogels (LS-g-SAH) have been synthesized and used for agricultural applications. The sample LS-g-SAH 12 was selected for the detailed investigation, and characterized using solid-state ^{13}C CP-MAS NMR, scanning electron microscopy (SEM), Fourier-transform infrared (FTIR) spectroscopy, X-ray diffraction (XRD), thermo-gravimetric analysis (TGA) and rheology, and urea release kinetics. Additionally, a control with a composition comparable to LS-g-SAH 12 but lacking lignosulfonate was compared to the synthesized hydrogels.

Experimental Section

Materials

Lignosulfonic acid sodium salt (LS) (Mol wt .~ 52,000) was purchased from Sigma Aldrich. Acrylic acid (AA) and N'N-methylenebisacrylamide (N,N'-MBA) were bought from CDH Pvt Ltd. Potassium persulfate (KPS) were purchased from SD fine chem Ltd. Urea and NaOH were purchased from CDH Pvt Ltd, and Qualikems Fine Chem Pvt. Ltd., respectively. All necessary water-based polymerization solutions were prepared using distilled water.

Synthesis of LS-g-SAH

The free radical co-polymerization technique was used to synthesize LS-g-SAH, where KPS and N,N'-MBA were utilized as an initiator and crosslinker, respectively. At room temperature, NaOH was added to neutralize acrylic acid. The desired amount of LS, AA, NaOH, N,N'-MBA, and KPS was stirred for 2 h continuously. The reaction product was then heated in a water bath preset at 60°C for 2 h. After co-polymerization, the hydrogel was cut into thin slices and immersed in distilled water for 24 h to remove the unreacted chemicals. The hydrogel was then oven-dried at 50°C until constant weight[13]. Fig. 1 Shows the visual representation of the synthesis of LS-g-SAH and Table 1 summarizes the different formulations of synthesized hydrogels.

Characterization of LS-g-SAH

The FTIR spectra in the range of 4000 to 650 cm^{-1} were recorded using Perkin Elmer 2000 FT-IR spectrometer. The solid-state ^{13}C-NMR analysis of samples was done on a Bruker Ascend-400 spectrometer operating at 400 MHz. The surface morphology was determined through SEM on EVO 18 Research, Zeiss, instrument. TGA was conducted on Perkin Elmer, Thermogravimetric Analyzer, TGA 4000 in Nitrogen environment with a uniform heating rate of 10 °C/min. XRD measurements were recorded through RIGAKU with source Cu-Kα performed at a voltage of 50 kV. The rheological analysis was carried out using Anton Par Rheometer (Model: Anton Par, Modular Compact Rheometer,

Table 1 — Various compositions used in synthesis with their swelling index

Formulation	LS (g)	NaOH (mol/L)	AAA (mL)	KPS (g)	NN-MBA (g)	Swelling index (g/g)	Gel %
LS-g-SAH 1	0.1	8.07	7.1	0.1	0.05	190.3	62.11
LS-g-SAH 2	0.1	8.07	7.1	0.1	0.07	172.82	63.38
LS-g-SAH 3	0.1	8.07	7.1	0.1	0.1	111.6	73.26
LS-g-SAH 4	0.1	8.07	7.1	0.1	0.12	60.22	77.59
LS-g-SAH 5	0.1	8.07	7.1	0.1	0.15	39.68	92.16
LS-g-SAH 7	0.1	8.07	7.1	0.04	0.08	89.31	85.31
LS-g-SAH 8	0.1	8.07	7.1	0.05	0.08	172.92	92.73
LS-g-SAH 9	0.1	8.07	7.1	0.08	0.08	252.83	96.90
LS-g-SAH 10	0.1	8.07	7.1	0.12	0.08	120.92	92.50
LS-g-SAH 11	0.1	8.07	7.1	0.14	0.05	560.24	95.28
LS-g-SAH 12	0.2	8.07	7.1	0.14	0.05	397.22	91.20
LS-g-SAH 13	0.3	8.07	7.1	0.14	0.05	336.35	85.99
LS-g-SAH 14	0.4	8.07	7.1	0.14	0.05	320.63	75.33
LS-g-SAH 15	0.05	8.07	7.1	0.14	0.05	248.32	72.04
LS-g-SAH 16	0.08	8.07	7.1	0.14	0.05	461	89.36

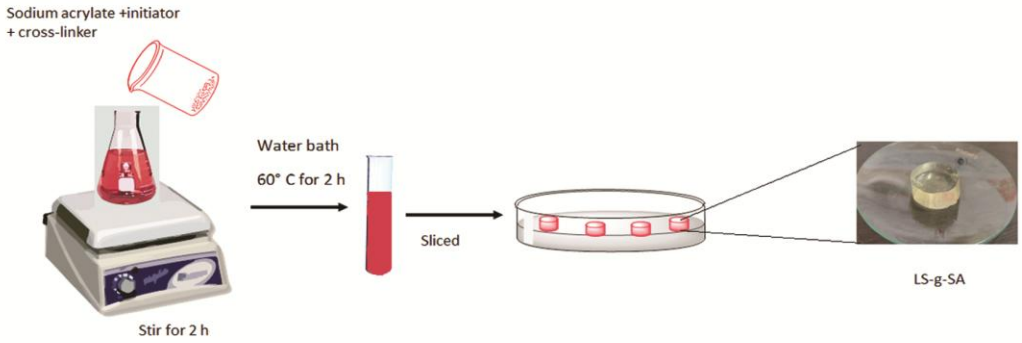


Fig. 1 — Visual representation of the synthesis of LS-g-SAH

MCR 302) with PP-25 parallel plate and 2.5 mm inter-platen gap.

Swelling experiment

The swelling studies for LS-g-AA hydrogels were performed in pH 4.0, 7.0, and 9.2 buffer and distilled water at 37°C. Dried hydrogel discs were weighed and placed in various pH media. These were removed after a fixed time interval and weighed. Filter paper has been used to remove additional surface water[14]. The samples were observed in triplicate. The swelling index (SI) (g g^{-1}) was calculated using the given Eq. (1):

$$SI = \frac{M_1 - M_2}{M_2} \quad \ldots (1)$$

Where, M_1 and M_2 represent the weight of equilibrium swollen hydrogel discs and wt. of dried hydrogel discs, respectively.

Gel content determination

The fraction of insoluble weight is known as gel content. The soluble part of the gel is extracted using the Soxhlet apparatus in boiling water for 24 h to obtain the gel content. The swollen hydrogels were removed from the apparatus and oven dried at 50°C [14]. The gel content of hydrogel was determined from Eq. 2.

$$\text{Gel content (\%)} = SI = \frac{M_e}{M_i} * 100 \quad \ldots (2)$$

Where, M_e and M_i are the respective weights of dry hydrogel before and after extraction.

Measurement of the highest water-retention capacity (WH%) of LS-g-SAH treated soil

This study was carried out using the garden soil of DTU, Delhi. Up to 1% of the hydrogel was added to 50 g of dry soil (less than 30 mesh size) and placed in a PVC tube (4.5 cm diameter). Its bottom was secured with a nylon cloth and weighed to obtain W_1. The treated soil samples were slowly saturated by adding tap water until they started to seep out from the bottom. The tubes were then re-weighed and marked W_2. A control experiment without LS-g-SAH was also conducted[15]. The largest water-retention capacity of treated soil was determined using Eq. 3.

$$\text{WH\%} = \frac{W_2 - W_1}{50} * 100 \quad \ldots (3)$$

Water Evaporation rate of LS-g-SAH and soil

A water evaporation test was used to investigate LS-g-SAH-treated soil's ability to retain water within its system structure for 55 days. Oven-dried soil (at 60°C for 48 h) was used to prepare different sets of hydrogel-treated soil along with a control set. Each mixture was fed with 50 mL distilled water and weighed (marked as M_i). The total water, soil, and hydrogel weight was kept as 'M_T.' Then, these were placed at room temperature and weighed daily (M_t) [16]. Here, the effect of LS-g-SAH in different amounts on soil was studied by considering the water evaporation loss value from blank soil as a control. The water evaporation ratio (WER) was determined using Eq. 4.

$$\text{WER} = \frac{M_i - M_t}{M_T} \times 100 \ \ldots (4)$$

Loading and release of urea by LS-g-SAH

Urea was loaded onto the hydrogel by immersing a pre-weighed dry LS-g-SAH into a urea solution (0.034 mol/L) for 24 h. After that, the swollen gels were removed from the water and dried till a constant weight was achieved.

The amount of urea released by LS-g-SAH was evaluated by soaking the dried urea-loaded samples in

a beaker with 100 mL of distilled water at ambient temperature. Later, 1 mL water was drawn from the medium at fixed intervals to observe the released urea, and the same amount of distilled water was added to the beaker to keep the volume constant [17].

A standard method from "The Specialized Standard of the People's Republic of China (SN/T 1004-2001)" was used to determine the concentration of urea residue. At 430 nm, a spectrophotometer detected a complex compound of urea and p-dimethylamino benzaldehyde (DMAB). The general reaction showing the formation of chromogen that follows Beer-Lambert law and absorbs energy at 430 nm in the UV-visible spectrum is shown below in Scheme 1. Then, the standard calibration curve was used to formulate the below-mentioned formula for the determination of urea concentration[12].

$$C\ (ppm) = \frac{A-0.1735}{0.0002} \quad \ldots (5)$$

Where, A is the UV absorbance of solution of unknown concentration.

Plant Growth Performance study

Three sets of 300 g of soil were placed in plastic pots with filter paper placed below with a small hole; (a) only urea, (b) control hydrogel-loaded urea, and (c) LS-g-SAH loaded with urea (at a depth of 5 cm). A plastic pot with direct urea in soil without hydrogel was kept as a check. Fifteen wheat seeds were sown in all three plastic pots with the same amount of hydrogel. Pots were exposed to environmental conditions, and irrigation was kept constant from day one for all containers. Growth profiles for all wheat grasses were observed at various time intervals. Germination was recorded by counting germinated seeds over two weeks. Just the appearance of the radicle was considered an indicator of seed germination. The plant shoots and roots were removed carefully from the soil. The shoot and root lengths and their fresh weight were recorded. The plant material was dried for 24 h at 70°C to determine the dry mass[18].

Results and Discussion

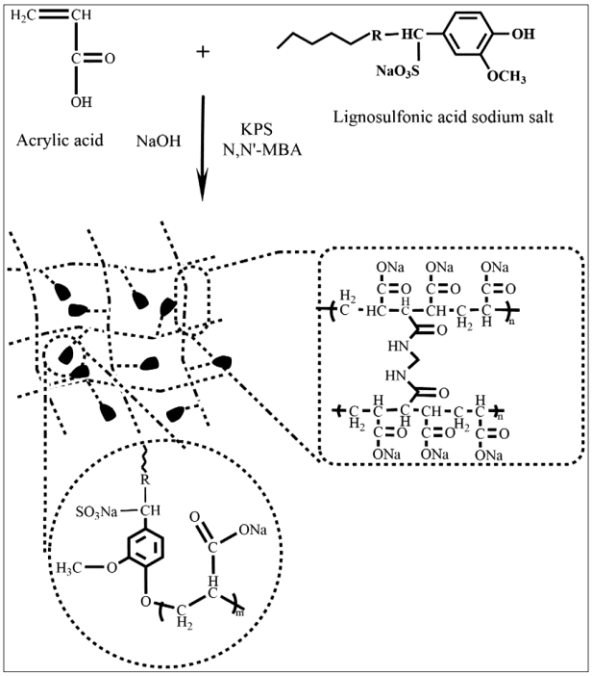
Mechanism of the formation of lignosulfonate-g-sodium acrylate hydrogels

The KPS used as an initiator generated anion radicals which further helped to generate phenoxy radicals on LS. The active radicals also reacted with the vinyl groups of sodium acrylate, leading to the propagation of chain-forming poly sodium acrylate. This poly(sodium acrylate) propagated until crosslinked and terminated by MBA to form lignosulfonate-g-acrylic acid hydrogels, as shown in Scheme 2. The electrostatic repulsion created by the negatively charged carboxylate ions in the poly(sodium acrylate) polymer chains will lead to a network expansion. This will improve the hydrogel's capacity to absorb water [15].

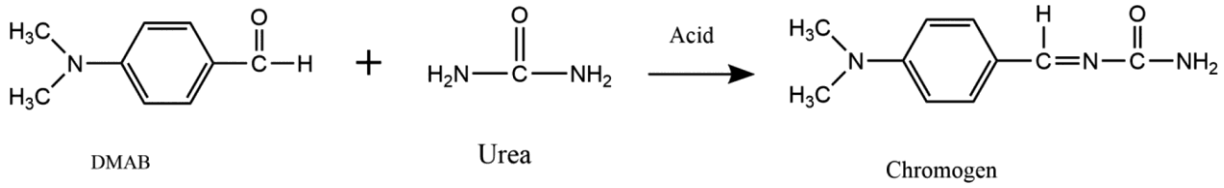
Characterization of hydrogel

Fourier-transform infrared (FTIR) spectroscopy

The IR spectra of LS, ctrl, and LS-g-SAH are shown in Fig. 2. The -OH stretching vibration of LS is observed as a broad band around 3,303 cm^{-1}. LS, ctrl



Scheme 2 — Synthesis of LS-g-SAH



Scheme 1 — DMAB-urea reaction

and LS-g-SAH show a band around 2,920 cm^{-1} due to C–H stretching in $–CH_2$ or $– CH_3$ groups. The vibration of the aromatic skeleton is observed from the two firm peaks at 1,572 and 1,417 cm^{-1}. [4] The strong peak at 1,046 cm^{-1} arises due to C–O–S stretching vibration, and the peak at 1,114 cm^{-1} is assigned to O=S=O antisymmetric stretching vibration in the LS.[4] Both these characteristic peaks confirm the presence of LS in LS-g-SAH. Further, new peaks were found when LS was compared to LS-g-SAH. The strong bands at 1404 and 1554 cm^{-1} are due to the asymmetric stretching in the carboxylate anion. The broad band and a peak at 1317 cm^{-1} and 627 cm^{-1}, respectively, correspond to the C-N and O=C–N groups of N,N'-MBA[19].

SS ^{13}C CPMAS

Solid state ^{13}C NMR spectroscopy is a powerful tool for the structural characterization of polymers. The spectra of LS, LS-g-SAH, and ctrl hydrogel prepared by graft co-polymerization are presented in Fig. 3. Peaks around 45 ppm and 185 ppm of LS-g-SAH and ctrl were assigned to saturated aliphatic chains and C from

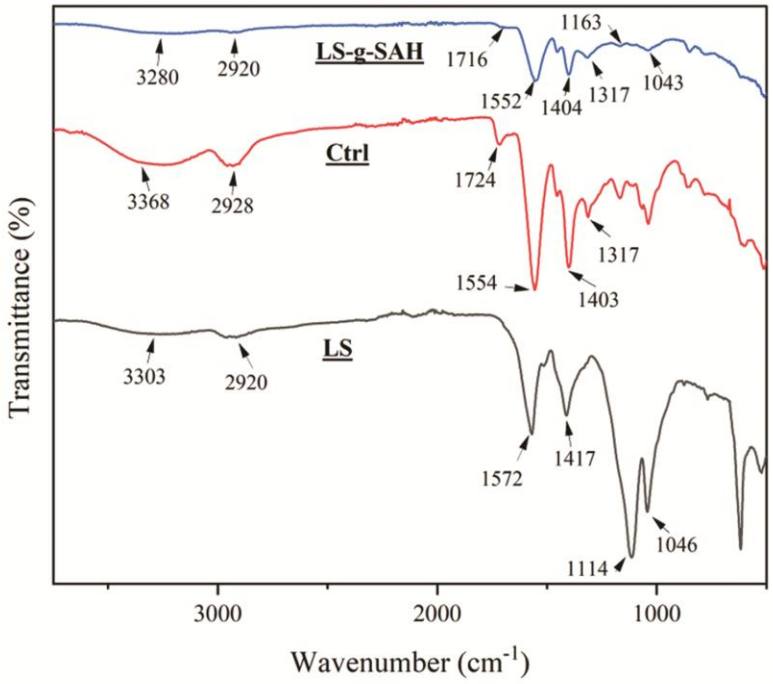


Fig. 2 — FTIR spectra of (a) LS, (b) Ctrl and (c) LS-g-SAH

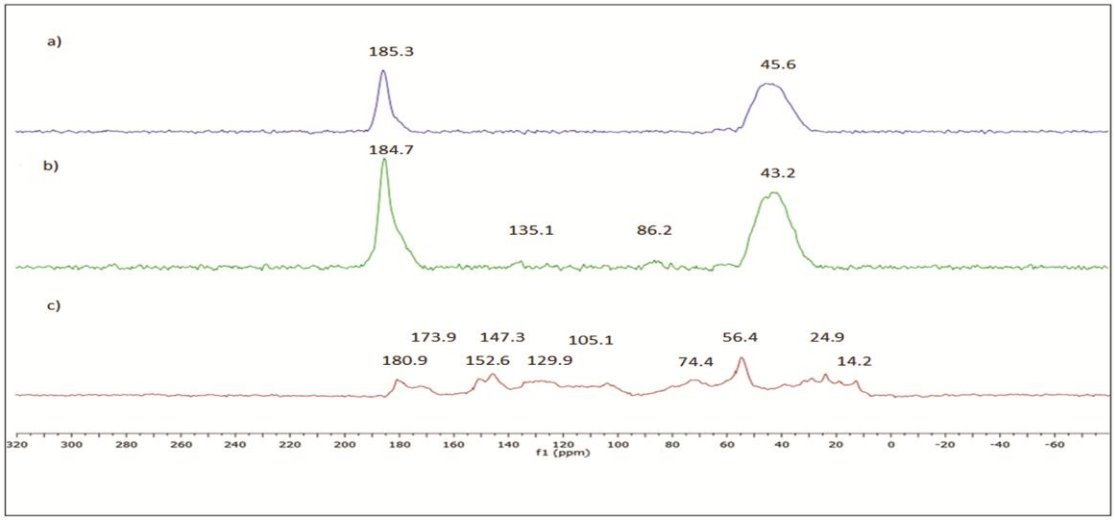


Fig. 3 — SS ^{13}C CP-MAS NMR spectrum of (a) ctrl, (b) LS-g-SAH, and (c) sodium LS

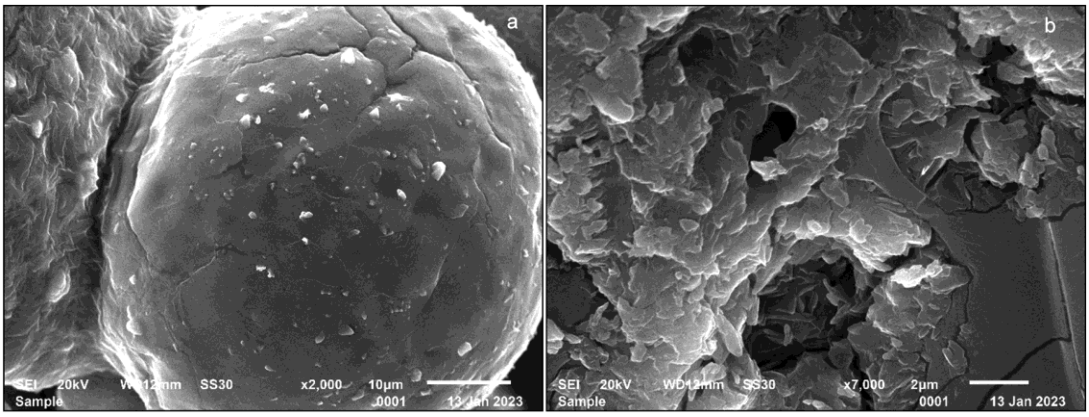


Fig. 4 — The SEM images of (a) LS and (b) LS-g-SAH

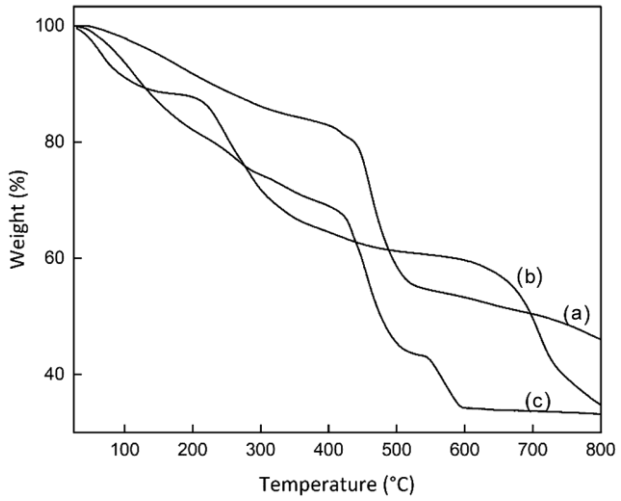


Fig. 5 — TGA plots of (a)-LS-g-SAH, (b) LS and (c) Ctrl

carbonyl ester (COONa), COOH, respectively. Characteristic peaks of basic aromatic and aromatic rings with branched C-C bonds can be observed in the range of 140-110 ppm and 75-10 ppm. The appearance of signals around 50-10 ppm in LS represents aliphatic C chain[20-23]. Most of the peaks of LS merged inside LS-g-SAH due to their low intensity. However, the presence of peaks at 135.1 ppm and 86.2 ppm confirms the presence of LS in LS-g-SAH. The results support the observations of the FT-IR spectra. The presence of free COOH groups and crosslinks probably increased the intensity of the peak at 184.7 ppm.

Scanning electron microscopy (SEM)

The SEM images of the LS and LS-g-SAH samples at the different magnifications are presented in Fig. 4. The surface of LS is even and smooth in granular form. On the contrary, the LS-g-SAH is rough, uneven, and porous. This could be due to the grafting of LS onto poly(acrylate) chain, further facilitating the adsorption of more water.

Thermal analysis

As shown in the thermograms (Fig. 5) of both LS and LS-g-SAH samples three weight loss ranges have been observed in the range of 29-802°C. The initial 16.3% weight loss at the stage (70-380°C) in LS-g-SAH is due to moisture loss from the polymer. Between 436-514°C, the weight loss percent of 23.4 is mainly due to the beginning of decomposition. The third stage, from 660-781°C, is the decomposition of the main polymeric backbone with a weight loss of 19.7% [6]. For LS, up to 150°C, the first decomposition is associated with removing volatile compounds and moisture, like acetaldehyde, acetone, and methanol. With the increase in temperature from 150°C to 250°C, LS began to decompose, known as vitrification transformation. The primary decomposition occurred between 200 and 400°C. Weight loss of LS above 500°C could be attributed to the CO from degrading C=O, C–O–C, and degeneration of aromatic rings [24,25]. The results from ^{13}C NMR and LS FTIR also support high amounts of C=O groups. Moreover, LS likely reacted with carbon from a sodium-containing salt that served as an activator for the dry pyrolysis process, further creating a mass loss between 650 and 800°C [26]. The point of the highest rate change is found at approximately 500°C for LS-g-SAH, 450°C, and 270°C for ctrl and LS, respectively. LS-g-SAH had a higher residual weight of about 83% up to 400°C compared to 69% and 64% of ctrl and LS. Similarly, at 800°C, the LS-g-SAH resulted in more than 46% of residual mass when compared to just 34% and 33% of the residual mass of LS and ctrl hydrogels. This observation indicated that grafting of LS has improved the thermal stability of LS-g-SAH.

X-ray diffraction

The XRD patterns observed in the 2θ range of 20°-80° of lignin showed several peaks at 23.12°,

25.52°, 27.32°, 28.34°, 31.64°, 33.60°, 40.48°, 45.34°, 56.36° and 75.16° (Fig. 6). The LS-g-SAH and ctrl showed two broad complex peaks in the range with their maxima at 22° and 32°. The observed maxima are characteristic of an amorphous polymer structure[27]. The data revealed that lignin's crystallinity is destroyed after grafting onto poly(sodium acrylate). Such a similar result has been observed by grafting carboxymethyl cellulose on sodium acrylate [28].

Rheology

The knowledge of rheology can easily explain the relationship between chemical structure and the macroscopic behaviour of hydrogels [29]. Rheological properties of LS-g-SAH are compared with ctrl to investigate the effect of the presence of lignin in the hydrogel, as shown in Fig. 7. During the amplitude sweep test, the shear stress amplitude was changed at a rate of 10 rad/s. At low deformation, the loss modulus (G″) and storage modulus (G′) are consistent, indicating that the hydrogel structure remained unaffected. It is called a linear-viscoelastic region (LVE). The disruption of structure with decreased modulus suggests the termination of the LVE region. The ctrl and LS-g-SAH showed viscoelastic solid behaviour up to a high strain of 402 and 511%, representing that LS-g-SAH may bear a higher shear strain than ctrl.

The frequency varied from 0.1 to 500 rad/s during the frequency sweep test while keeping the shear strain % within the LVE region constant. It was observed that the loss modulus was found to be prevalent at higher frequencies while the storage modulus was prevalent at lower frequencies representing the viscoelastic solid behaviour. Tan θ (dimensionless) generally varies from 0 to 1 for viscoelastic solutions. For an ideal viscous behaviour, tan θ is infinity, i.e., θ is 90° and for a perfect elastic behaviour, tan θ is 0, i.e., θ is 0°.[30]

It was observed from Table 2 that the cross-over point for LS-g-SAH falls in the low-frequency region (126 rad/s), indicating that its chain-chain interaction breaks at a lower frequency in comparison to ctrl. LS-g-SAH and ctrl had a loss factor of 0.04 and 0.018, representing the dominancy of elastic behavior at low frequency and viscous behavior at high frequency, which agrees with the previously reported literature [31]. However, ctrl hydrogel shows more elastic behaviour

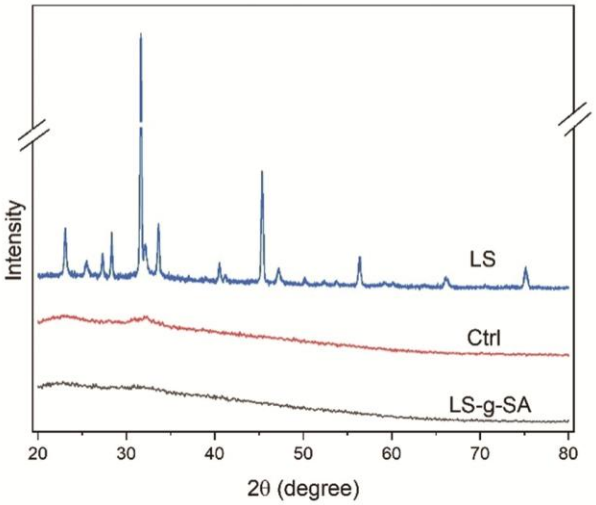


Fig. 6 — XRD plots of LS, Ctrl, and LS-g-SAH

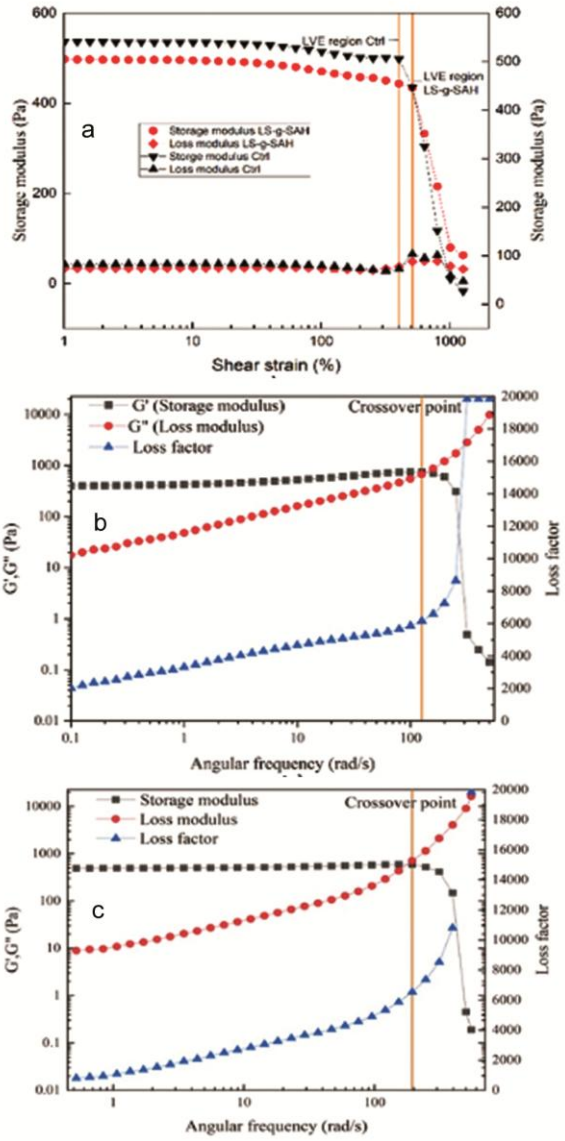


Fig. 7 — Rheological analysis: (a) amplitude sweep, (b) Frequency sweep of LS-g-SAH and (c) Frequency sweep of Ctrl hydrogel

Table 2 — Frequency sweep data of LS-g-SAH and Ctrl hydrogels

Sample	Low frequency			Cross-over point	High frequency		
	Storage modulus G' (Pa)	Loss modulus G'' (Pa)	Loss factor Tan θ	Angular frequency (rad/s)	Storage modulus G' (Pa)	Loss modulus G'' (Pa)	Loss factor Tan θ
LS-g-SAH	400	17.58	0.04	126	0.14	1722	infinite
Ctrl	491	8.89	0.018	194	0.19	2107	infinite

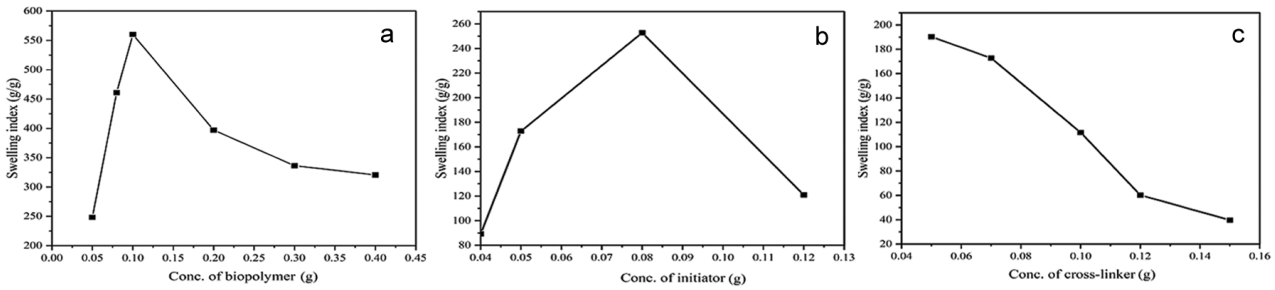


Fig. 8 — Effect of variation of (a) biopolymer, (b) initiator and (c) crosslinker on the swelling index

at low frequencies. Frequency sweep data of LS-g-SAH and ctrl hydrogels rae shown in Table 2.

Gel content (%)

The gel content percentage of LS-g-SAH present in Table 1 was calculated using Eq.(2). The gel percentage of hydrogels varied according to the concentration of crosslinker, biopolymer, and initiator. The gel content reduced as the concentration of the crosslinker and biopolymer increased. In contrast, the initiator initially enhanced the gel content but decreased when added in amounts greater than 0.08 g. Gel contents of synthesized hydrogels were found in the 64 - 97 % range.

Effect of variation of framework parameters on the swelling index

The relationship between framework parameters and water absorbency was investigated by Flory, who concluded that hydrogel absorption depends on several parameters[32]. Therefore, the variables affecting the reaction, such as the concentration of biopolymer, initiator and crosslinker are studied to obtain the optimum reaction conditions.

Effect of biopolymer

The concentration of hydrophilic biopolymer is one of the influencing factors of the swelling property of hydrogel. Fig. 8(a) shows that LS was varied from 0.05 g to 0.4 g in hydrogel to find the optimum concentration. Adding LS provided a more significant number of diverse hydrophilic groups to the hydrogel, resulting in an increment in the SI. After the LS content reached 0.1 g, the amount of ionic hydrophilic groups in the hydrogel, such as sulfonic acid and carboxyl groups, could have become higher than the non-ionic hydrophilic groups such as amide group and hydroxyl group, which further led to the declination of the SI. [33]

Effect of initiator

The reaction initiator creates active sites on the polymer. Thus, it plays a crucial role in the preparation of hydrogel. The SI of hydrogel increased with an increase in initiator concentration from 0.04 g to 0.08 g. Afterward, it decreased with the increasing amount of the initiator, according to Fig. 8(b). The maximum water absorbance was obtained at 0.08 g of the initiator. An increase in the amount of initiator up to 0.08 g generated more free radicals, resulting in larger chain ends in the network while shortening the average length of the kinetic chain. Accordingly, the SI would increase. However, an initiator concentration greater than 0.08 g may have increased the reaction velocity and decreased the network space of the polymer, which resulted in a decrease in the SI. [34]

Effect of concentration of crosslinker

It is evident from Fig. 8(c) that the SI sharply decreased with the increase in the crosslinker concentration of hydrogel. The graft reaction did not occur effectively with a concentration of MBA lower than 0.03 g resulting in a jelly-like product with poor dimensional stability. Beyond this amount, adding

more crosslinker to hydrogel resulted in more crosslink junctions in the polymeric chains. It increases the extent of crosslinking and reduces the free available volume inside the polymeric network[34,35]. This results in a less swelling tendency of the hydrogel.

Swelling studies

In general, the swelling index of different formulations in distilled water has been reported in Table 1. The swelling performance of L10 was investigated in various pH, i.e., pH 4.0, 7.0, and 9.2, and distilled water. Fig. 9 shows the enhancement in equilibrium SI from 51 to 73 with an increase in pH from 4, 7, and 9.2. Lower swelling of hydrogel in an acidic medium may be attributed to the improper ionization of the crosslinked network. Hence, at a lower pH, a collapsed state was observed. However, in alkaline conditions, electrostatic repulsions may have been produced by the carboxylate anions (COO^-). This might have caused a widening of space present in hydrogel networks, leading to an increase in the SI of the hydrogel[35,36].

Agricultural applications of LS-g-SAH

Highest water retention capacity

Water availability in the soil is of utmost importance for better plant growth. As a result, the impact of LS-g-SA hydrogel on soil's maximum water retention capacity (WH%) was investigated. The amount of absorbed water and the corresponding concentration of incorporated hydrogel is shown in Fig. 10. It proved that the synthesized hydrogel exhibited excellent water-absorbing capacity and could be used to improve the soil water-retention capacity and reduce the frequent irrigation requirement.

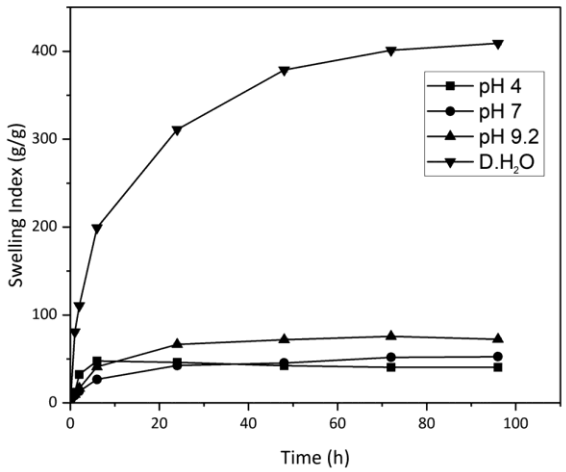


Fig. 9 — Swelling study of hydrogel

Evaporation study

Evaporation refers to the water loss from the soil directly to the atmosphere through water vapour. The water-holding capacity of soil affects the water evaporation rate of the soil. The control's water evaporation ratio from the study was higher than soil amended with LS-g-SAH, probably due to the interaction between water and polymer molecules and the macromolecular network hindrance [16]. Fig. 11 shows that adding just 0.6 g LS-g-SAH to the soil reduced the evaporation rate from 99 to 76.69% of the total added water on the 54th day.

Release kinetics of urea

The power of law equation is used to understand the mechanism involved in releasing urea.

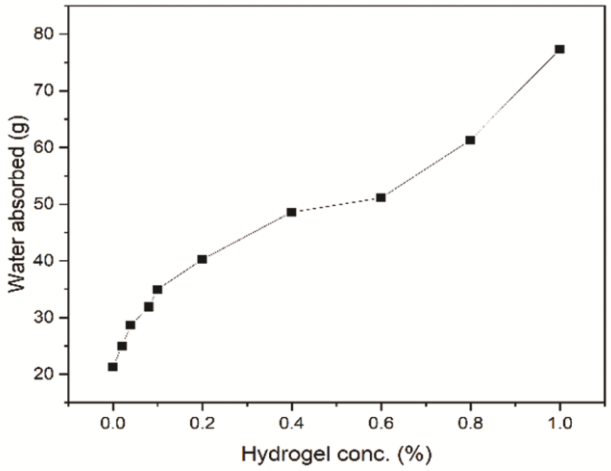


Fig. 10 — Water retention capacity plot of soil amended with LS-g-SAH

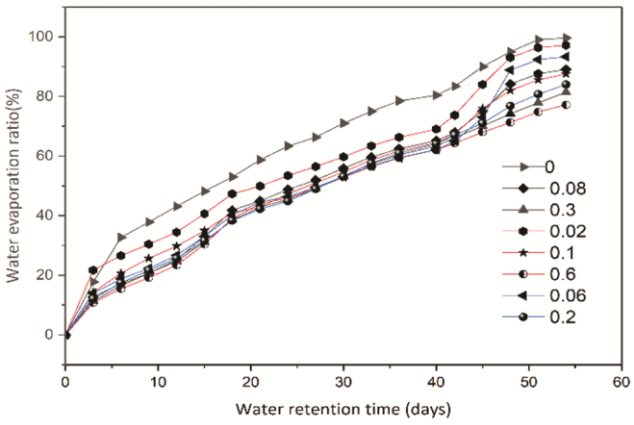


Fig. 11 — Water evaporation study of soil amended with different concentrations of LS-g-SAH

$$\frac{Mt}{M\infty} = k_{kP}t^{n} \quad \text{... (6)}$$

where M_t and M_∞ Represents the urea released at time t and the equilibrium point, respectively. Whereas the ratio of M_t to M_∞ denotes the fraction of total urea released at the time 't', while 'k' is a polymer structure constant. The coefficient 'n' describes the type of diffusion involved. Only the section of the release curve where the ratio $\frac{M_t}{M_\infty} < 0 \cdot 6$ should be used to determine the exponent 'n' [37,38].

Diffusion-controlled release is indicated by a Fickian diffusion mechanism with n≥ 0.45, whereas 0.45<*n*<0.89 corresponds to non-Fickian transport, and *n*=0.89 corresponds to case-II relaxational transport, which specifies swelling-controlled release [38].

Urea release kinetics studied using different models such as a) first order, b) Korsmeyer–Peppas, c) Higuchi model in distilled water are shown in Fig. 12, The slow-release process of urea through LS-g-SAH occurs in the following manner; firstly, the water tries to enter the hydrogel. Afterward, the hydrogel starts to swell by absorbing water until a dynamic exchange is developed between the free water of the hydrogel and the water present in the soil. Meanwhile, the urea in the hydrogel matrix continues to dissolve and diffuse. It gets released into the water due to the osmosis differences and dynamic exchange of free water [16].

From Table 3, "n" is 0.8193, representing a non-Fickian diffusional release of urea from the synthesized hydrogel. The coefficient of determination values (R^2) showed its maximum value to be 0.9932 for the first-order kinetics; hence, the release followed the First order kinetics model.

Plant Growth Performance

The plant growth performance study using the LS-g-SAH is shown in Fig. 13. The wheatgrass growth can be observed by the changed total dry and fresh weight, the average number of germinated seeds, the average wheatgrass height, and the average shoot height of all samples. All these are mentioned in Table 4. The height of the plants increased with the advancing age of the crops. On the 14th day, significantly tallest grasses (the tallest grass was 27 cm) were recorded in LS-g-SAH, followed by sample B, i.e., ctrl, though the shortest plants were recorded

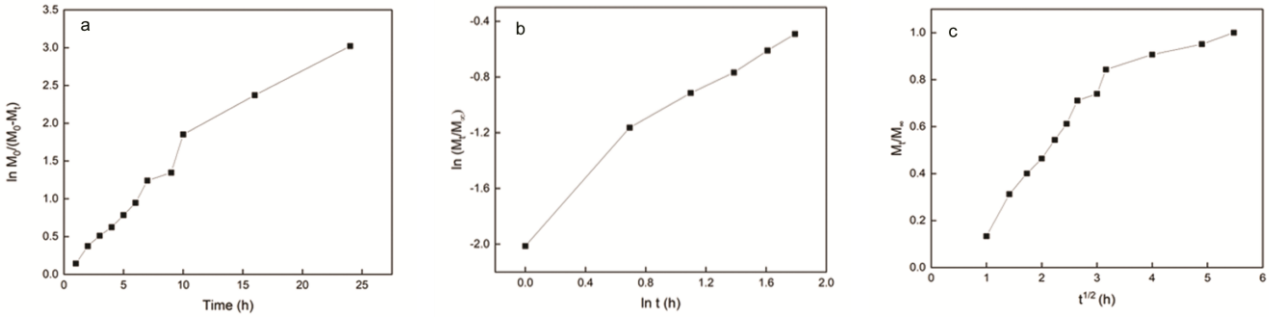


Fig. 12 — Urea release kinetics using (a) first order, (b) Korsmeyer–Peppas and (c) Higuchi model in distilled water

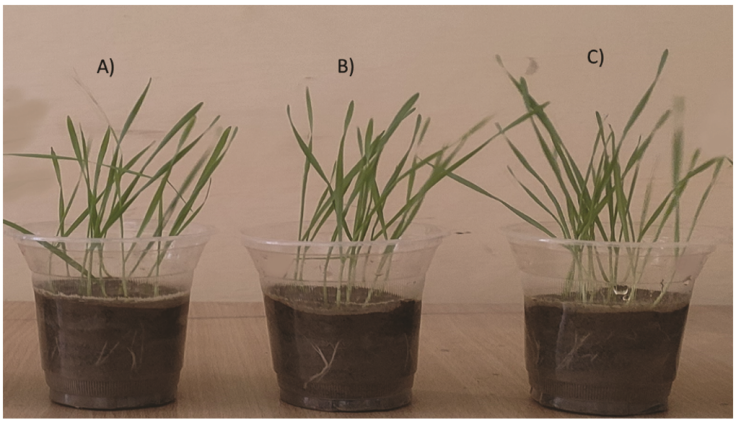


Fig. 13 — Effect of hydrogel treatment on wheatgrass after 14 days using (A) only urea, (B) control hydrogel-loaded urea, and (C) LS-g-SAH loaded with urea

Table 3 — Mathematical models used for the study of release kinetics from LS-g-SAH

Model	Equation	Parameter		R sq.
First-Order	$f = 100[1 - e^{-k_1 t}]$	k_1	0.174	0.9932
	f = amount of the released drug k_1 = rate constant, t = time			
Higuchi	$f = k_H t^{0.5}$	k_H	22.134	0.8242
	f = amount of the released drug k_H = Higuchi dissolution constant t = time			
Korsmeyer– Peppas	$f = k_{kp} t^n$	k_{KP}	0.4099	0.9656
	k_{KP} = constant depicting the experimental parameters f = amount of the drug released n = release exponent	n	0.8193	

Table 4 — Wheatgrass growth parameters on the 14th day: (A) only urea, (B) control hydrogel-loaded urea, and (C) LS-g-SAH loaded with urea

Sample	Total fresh weight of wheatgrass (Average) (g)	Total dry Weight of wheatgrass (Average) (g)	Average number of germinated seeds (per pot) at 14th day	Average wheatgrass height (cm) at 14th day	Average shoot height (cm) on the 14th day
A	0.7927	0.2860	13	17.2	12.4
B	0.8102	0.3002	12	17.7	12.8
C	0.8993	0.3004	15	19.7	14.4

under sample A. The germination and growth of seeds are known to be improved with the prolonged availability of urea. LS-g-SAH showed 100% germination followed by control hydrogel (86.6%), i.e., sample B and untreated soil (80%) urea release, respectively.

Conclusion

Novel LS-g-SAH was synthesized and loaded with urea using the gravitational equilibrium method. When observed in different media, the maximum water absorbency was 560 g/g in distilled water. Using synthesized hydrogel remarkably raised the soil's water-holding capacity from 21.27 to 77.3 g. Also, the water evaporation rate reduced from 99 to 76.69% of the total added water. The release kinetics study of urea in water revealed that LS-g-SAH showed a slow-release pattern with 60% release in 24 h following first-order release kinetics. The release pattern followed the Non-Fickian release mechanism. The increased height of wheatgrass also confirmed the beneficial effects of LS-g-SAH. Thus, the synthesized poly(sodium acrylate) hydrogel based on lignin can serve as a water retention agent and has the potential to be used for the gradual release of various agrochemicals.

Acknowledgment

The authors would like to acknowledge the financial support of the Department of Applied Chemistry, Delhi Technological University, Delhi.

References

1 Rajakumar R & Sankar J, *Int J Appl Pure Sci Agric*, 2 (2016) 163.
2 Wang Q, Guo J, Lu X, Ma X, Cao S, Pan X & Ni Y, *Int J Biol Macromol*, 181 (2021) 45.
3 Mandlekar N, Cayla A, Rault F, Giraud S, Salaün F, Malucelli G & Guan J P, *Lignin - Trends Applications* In Tech, (2018).
4 Yu C, Wang F, Zhang C, Fu S & Lucia L A, *React Funct Polym*, 106 (2016) 137.
5 Ludmila H, Michal J, Andrea Š & Aleš H, *Wood Res*, 60 (2015) 973.
6 Meng Y, Lu J, Cheng Y, Li Q & Wang H, *Int J Biol Macromol*, 135 (2019) 1006.
7 Fernández-Pérez M, Flores-Céspedes F, Daza-Fernández I, Vidal-Peña F & Villafranca-Sánchez M, *Ind Eng Chem Res*, 53 (2014) 13557.
8 Wilske B, Bai M, Lindenstruth B, Bach M, Rezaie Z, Frede H G & Breuer L, *Environ Sci Pollut Res*, 21 (2014) 9453.
9 Thakur V K & Thakur M K, *Int J Biol Macromol*, 72 (2015) 834.
10 Thakur S, Govender P P, Mamo M A, Tamulevicius S, Mishra Y K & Thakur V K, *Vacuum*, 146 (2017) 342.
11 Song B, Liang H, Sun R, Peng P, Jiang Y & She D, *Int J Biol Macromol*, 144 (2020) 219.
12 Zheng T, Liang Y, Ye S & He Z, *Biosyst Eng*, 102 (2009) 44.
13 Gupta A P & Warkar S G, *Int J Pharm Bio Sci*, 6 (2015) 516.
14 Akalin G O & Pulat M, *J Polym Res*, 27 (2020) 1.
15 Ni B, Liu M & Lü S, *Chem Eng J*, 155 (2009) 892.
16 Rabat N E, Hashim S & Majid R A, *Procedia Eng*, 148 (2016) 201.
17 Raafat A I, Eid M & El-Arnaouty M B, *Nucl Instrum Methods Phys Res Sect B Beam Interact Mater Atoms*, 283 (2012) 71.

18 Akhter J, Mahmood K, Mali K A, Mardan A & Ahmad M, *Plant Soil Environ*, 10 (2004) 463.
19 Reddy V & Rao R, *Indian J Pure Appl Phys*, 46 (2008) 611.
20 Sun Y, Ma Y, Fang G, Li S & Fu Y, *Bioresour J*, 11 (2016) 5731.
21 Ma Y, Sun Y, Fu Y, Fang G, Yan X & Guo Z, *Chemosphere*, 163 (2016) 610.
22 Tomar R S, Gupta I, Singhal R & Nagpal A K, *Polym Plast Technol Eng*, 46 (2007) 481.
23 Mali K K, Dhawale S C & Dias R J, *Int J Biol Macromol*, 105 (2017) 463.
24 He M K, He Y L, Li Z Q, Zhao L N, Zhang S Q, Liu H M & Qin Z, *Int J Biol Macromol*, 209 (2022) 258.
25 Hemmilä V, Hosseinpourpia R, Adamopoulos S & Eceiza A, *Waste Biomass Valori*, 11 (2020) 5835.
26 Zhang J, Zhao L, Yu F & Zang L, *Energy Fuels*, 33 (2019) 8812.
27 Adnadjevic B & Jovanovic J, *J Appl Polym Sci*, 107 (2008) 3579.
28 Kumar B, Priyadarshi R, Sauraj, Deeba F, Kulshreshtha A, Gaikwad K K, Kim J, Kumar A & Negi Y S, *Gels*, 6 (2020) 49.
29 Stojkov G, Niyazov Z, Picchioni F & Bose R K, *Gels*, 7 (20201) 255.
30 Yadav R & Purwar R, *Polym Test*, 93 (2021) 106916.
31 Sani M I, Teo Y Y & Misran M, *Polym Bull*, 78 (2021) 3399.
32 Flory P J, *Principles Polymer Chemistry*, Cornell University Press, Ithaca, NY, (1953) 5.
33 Wang X, Wang Y, He S, Hou H & Hao C, *Ultrason Sonochem*, 40 (2018) 221.
34 Li W, Wang J, Zou L & Zhu S, *J Polym Res*, 15 (2008) 435.
35 Khushbu, Warkar S G & Kumar A, *Polymer*, 182 (2019) 121823.
36 Singh B & Sharma V, *Polymer*, 91 (2016) 50.
37 Costa P & Sousa L J M, *Drug Dev Ind Pharm*, 29 (2003) 89.
38 Alam M A, Takafuji M & Ihara H, *Polym J*, 46 (2014) 293.

J Mater Sci: Mater Electron **(2023) 34:2059**

Thermally stable multi-color emitting Dy^{3+}/Eu^{3+} co-doped $BaO–ZnO–Li_2O–P_2O_5$ glasses for w-LEDs

Kartika Maheshwari[1,2], Aman Prasad[3], Ravita Pilania[4], Shaik. Mahamuda[5], Yasha Tayal[1,2], and A. S. Rao[1,*]

[1] *Department of Applied Physics, Delhi Technological University, Bawana Road, Delhi 110042, India*
[2] *ABES Engineering College, Ghaziabad 201009, India*
[3] *Department of Physics and Computer Science, Dayalbagh Educational Institute (DEI), Agra 282002, India*
[4] *Department of Physics, Chaudhary Bansilal University, Bhiwani 127021, India*
[5] *Department of Engineering Physics, College of Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522302, India*

Received: 10 July 2023
Accepted: 25 September 2023
Published online: 31 October 2023

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

ABSTRACT

Dy^{3+} and Dy^{3+}/Eu^{3+} co-doped $BaO–ZnO–Li_2O–P_2O_5$ (BZLP) glasses have been prepared and studied for their luminescence properties. Dy^{3+}doped BZLP glass emits red ($^4F_{9/2} \rightarrow {}^6H_{13/2}$), green ($^4F_{9/2} \rightarrow {}^6H_{15/2}$) and blue ($^4F_{9/2} \rightarrow {}^6H_{11/2}$) emission under 350 nm excitation. Doping of Eu^{3+} tunes the emission from Dy^{3+} in the BZLP glasses. The cool white emission shifts towards the warm white region at 350 nm excitation wavelength owing to the energy transfer between sensitizer Dy^{3+} and activator Eu^{3+} ions. Also, Dy^{3+}/Eu^{3+} co-doped BZLP glass gives intense red emission under 393 nm excitation. Dexter's theory was used to ascertain the energy transfer mechanism among the sensitizer Dy^{3+} and activator Eu^{3+} ions via Reisfeld's approximation. The glasses have significant thermal stability and high activation energy. These Dy^{3+}/Eu^{3+} co-doped BZLP glasses can be utilized in the fabrication of cool/warm w-LEDs and red emitting devices applications.

1 Introduction

Demand for energy is ever increasing. So, energy saving has become a very important research field being considered significantly [1, 2]. In artificial lighting, white light-emitting diodes (w-LEDs) lighting sources have become particularly important for reducing the world's energy consumption due to advantages like huge amount of energy saving, superior efficacy, compact size, environmental friendliness, etc. related to fluorescent lamps [3–5]. A commercial w-LED is based on a blue LED and YAG: Ce^{3+} phosphor [6, 7]. These w-LEDs suffer from low color rendering index, inappropriate color temperature and halo effect [8]. Glasses doped with transition ions have attracted considerable interest because of their memory and photoconducting properties. They also find potential applications in the development of new tunable solid-state lasers, solar-energy converters and fiber-optic communication devices [9]. To improve these shortcomings, the development of rare earth-activated glasses as a worthy substitute for phosphors is being investigated [8]. Furthermore, glass has many distinctive characteristics such as an

Address correspondence to E-mail: drsrallam@gmail.com

https://doi.org/10.1007/s10854-023-11394-7

easy cost-effective production process, good thermal and chemical stability along with a high value of rare earth solubility [10, 11].

The choice of a good glass host activated with rare earth ions is still a difficult task in assessing optical and lasing characteristics efficiently [12]. Numerous glass hosts like borates, tellurites, silicates, fluorides, phosphates etc. have been prepared and inspected for innumerable optical features. Phosphate is one of the appropriate glass hosts having distinguished properties like low melting temperature, highly thermally stability, high rare earth (RE) solubility, low dispersion and clear visibility in the wide-ranging spectrum [13, 14]. Phosphate glasses have found usage in the area of photonics but their hygroscopic nature and poor chemical stability limit their utility [15]. So, network modifiers (BaO, Li_2O) and an intermediate compound (ZnO) are added to the host thereby enhancing chemical & thermal stability and reducing thermal expansions. Additionally, ZnO helps reduce the hygroscopic affinity of the glass and increases rare earth ion solubility in the glass matrix. RE doped phosphate glasses improve the luminescent characteristics of materials in various fields such as optoelectronic devices and lasers due to the emission from the ultraviolet region to the infra-red region [16–21].

Trivalent lanthanide ions are considered the most appropriate activator in many crystalline or amorphous inorganic host lattices. Among the lanthanides, the trivalent Dy^{3+} is an important ion, which emits in the blue, yellow and red regions under numerous excitations. Through combination of these color, Dy^{3+} ion doped luminescent material is able to produce white lights. The white light is tuned by a proper yellow to blue ratio that can be altered by varying Dy^{3+} ions doping concentration [22, 23]. The bluish white light emitted from the Dy^{3+} ions doped glasses lacks a red constituent and also has high co-related color temperature (CCT). This cool bluish light can be tuned to warm white light via co-doping of appropriate rare earth ions that can emit in the red region. Various RE ions such as Dy^{3+}/Ce^{3+}, Dy^{3+}/Eu^{3+}, Dy^{3+}/Tm^{3+} etc. are being used to obtain proper artificial white light [24–26]. We have introduced red emitting Eu^{3+} ion as a co-activator in Dy^{3+} doped glass compositions so as to improve the CCT and generate warm white light. This was undertaken due to the overlapping of Dy^{3+} emission profile with Eu^{3+} ions excitation profile. This fact shows that some energy transfer can occur between both ions in the proposed phosphate glass composition [27].

In the current investigation, Dy^{3+}/Eu^{3+} activated BZLP glasses have been prepared to achieve the optimal energy transfer with thermal stability for warm white light generation and usage in other solid-state artificial lighting devices. Structural analysis along with the study of optical properties of the BZLP glasses has been undertaken in this research work.

2 Materials and methods

2.1 Preparation of Dy^{3+} and Dy^{3+}/Eu^{3+} induced BZLP glasses

Transparent RE ions doped glasses have been prepared via melt quenching routes with the help of highly pure reagents using the following chemical equations:

$$(15-x)\ \mathrm{BaO}-15\mathrm{ZnO}-10\mathrm{Li_2O}-60\mathrm{P_2O_5}\ :\ x\mathrm{Dy_2O_3}\left(x\text{varies from 0.1 to 2.0 mol\% concentration of } \mathrm{Dy^{3+}}\text{ions}\right)$$

$$(15-y)\ \mathrm{BaO}-15\mathrm{ZnO}-10\mathrm{Li_2O}-60\mathrm{P_2O_5}\ :\ y\mathrm{Eu_2O_3}\left(\mathrm{y}\ =\ 0.1\ \text{mol\% concentration of } \mathrm{Eu^{3+}}\text{ions}\right)$$

Besides the aforementioned glasses, a series of Dy_2O_3 and Eu_2O_3 co-doped glasses was also prepared:

(15-x–y) BaO-15ZnO-10Li_2O-60P_2O_5:$x$$Dy_2O_3$:$y$$Eu_2O_3$ (x = 1.5 mol% concentration of Dy^{3+} ions and y = 0.0, 1.0, 2.0, 3.0, 4.0 and 5.0 mol% of Eu^{3+} ions). The prepared Dy^{3+}/Eu^{3+} co doped glass samples of size 3 cm and uniform thickness 2 mm were named as BZLP:DE0, BZLP:DE1, BZLP:DE2, BZLP:DE3, BZLP:DE4 and BZLP:DE5 and shown in the Fig. 1.

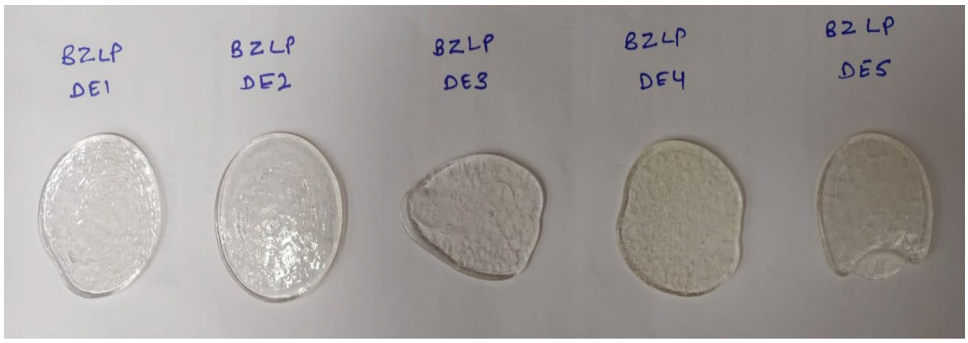


Fig. 1 Dy^{3+}/Eu^{3+}co-doped ions BZLP glass samples

Appropriate stoichiometric amounts of $BaCO_3$, ZnO, Li_2CO_3, $NH_4H_2PO_4$ and Dy_2O_3 were taken and ground for 1 h and the blended mixture was further heated in a furnace for melting at 1200 °C. After a while, the powder mixture turned into a liquid. This liquid mix was then immediately quenched on a preheated brass plate to give the final glass product of uniform thickness. Further, the prepared glasses were annealed at 350 °C for 1 h to remove thermal shock and air bubbles strain. These glasses were polished with emery paper with 40 grit and also used red oxide powder so that the smoothness of the glass does not be affected.

2.2 Instrumentation

XRD data of Dy^{3+}/Eu^{3+} co-doped BZLP glass was recorded using a D8 advance Bruker X-ray diffractometer in the $10° \leq 2\theta \leq 80°$ range. The absorption profile was recorded via JASCO V 670 spectrophotometer. Photoluminescence (PL) studies were carried out in a JASCO 8300FL Spectro fluorophotometer at room temperature. The decay profile co-doped BZLP glass has been characterized using an Edinburgh FLS980 spectrophotometer. Temperature-dependent PL (TD-PL) profiles were recorded using ocean optics spectrophotometer (FLAME- S-XR1-ES).

3 Results and discussions

3.1 Structural analysis

The amorphous nature of the prepared Dy^{3+}/Eu^{3+} co-doped BZLP glass has been analysed using the recorded XRD profile as shown in Fig. 2. The profile contains a wide band with the nonexistence of any

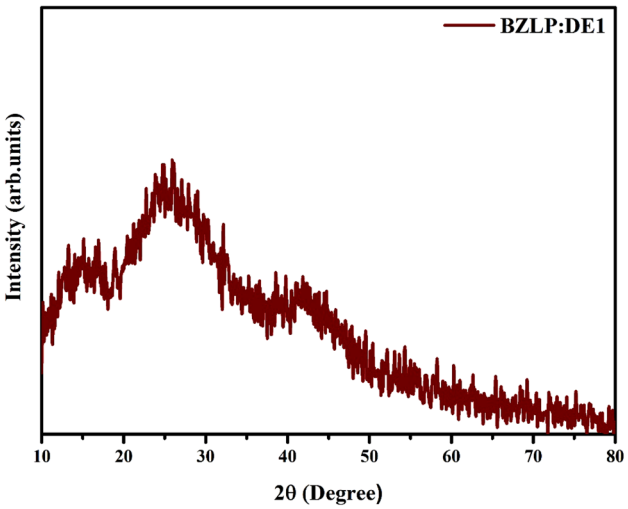


Fig. 2 XRD pattern of Dy^{3+}/Eu^{3+}co-doped BZLP glass

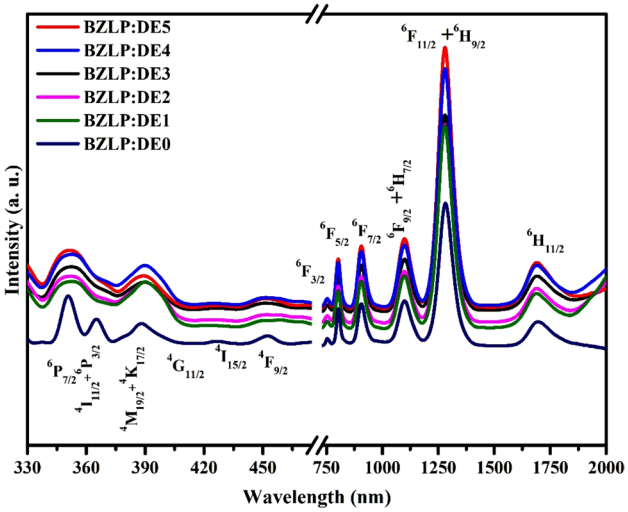


Fig. 3 Absorption profile of Dy^{3+} doped and Dy^{3+}/Eu^{3+}co-doped BZLP glasses

sharp peaks, confirming the prepared Dy^{3+}/Eu^{3+} co-doped BZLP glasses having glassy and amorphous in nature.

3.2 Absorption characteristics

The absorption profiles of Dy^{3+} doped BZLP glasses in 330–2000 nm range have been shown in Fig. 3. Peaks pertaining to transitions from $^6H_{15/2}$ state to many higher energy levels at $^6P_{7/2}$,$^4I_{11/2}$ + $^6P_{3/2}$,$^4M_{19/2}$ + $^4K_{17/2}$, $^4G_{11/2}$,$^4I_{15/2}$, $^4F_{9/2}$,$^6F_{3/2}$, $^6F_{5/2}$, $^6F_{7/2}$,$^6F_{9/2}$ + $^6H_{7/2}$,$^6F_{11/2}$ + $^6H_{9/2}$ and $^6H_{11/2}$ of Dy^{3+} ions [20]. The absorption peak due to $^6H_{15/2}$ to $^6F_{11/2}$ + $^6H_{9/2}$ transition is most intense and follows $|\Delta S| = 0$, $|\Delta L| \leq 2$ and $|\Delta J| \leq 2$ [28].

The absorption characteristics of Dy^{3+}/Eu^{3+} co-doped BZLP glasses were also illustrated in Fig. 3. The bands are mainly due to Dy^{3+} ions and broaden in the UV region, when co-doped with Eu^{3+} ions as depicted in Fig. 3. The absorption peaks observed for Dy^{3+} doped BZLP glass in infrared region were in good agreement with the literature [6, 7, 15].

On the basis of absorption characteristics, optical band gap of all prepared glasses was evaluated via Davis and Mott relation as follows [29]:

$$\alpha h\nu = B\left(h\nu - E_{opt}\right)^{n}$$

Here E_{opt} is indirect energy bandgap, α represents the absorption coefficient, B is the band trailing parameters and $h\nu$ is incident photon energy. The value of n is 2 for indirect allowed transitions [30]. As per the plot presented in Fig. 4, estimated indirect E_{opt} values are 4.21, 4.04, 3.96, 3.91, 3.88 and 3.86 eV for BZLP:DE0, BZLP:DE1, BZLP:DE2, BZLP:DE3, BZLP:DE4 and BZLP:DE5 glasses, respectively.

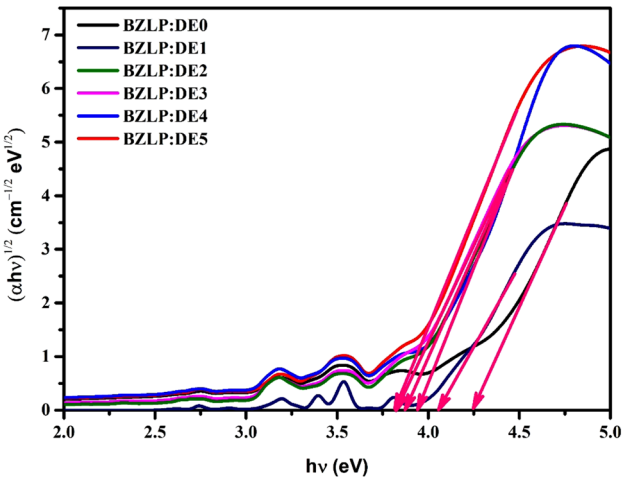


Fig. 4 Indirect bandgap plot of Dy^{3+} doped and Dy^{3+}/Eu^{3+}co-doped BZLP glasses

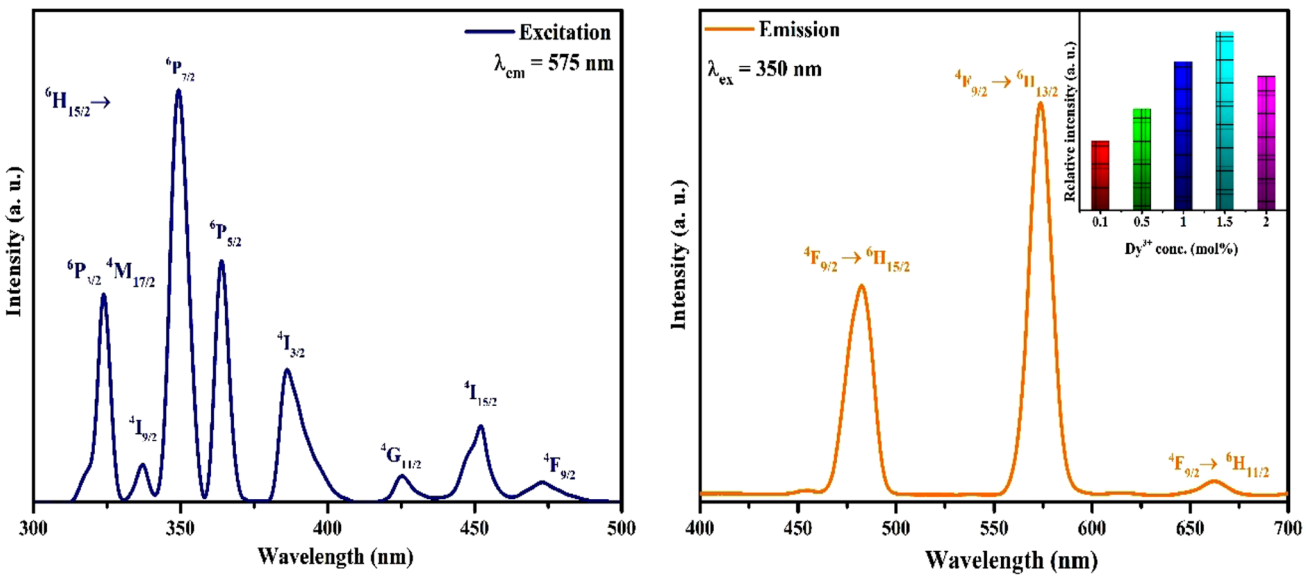


Fig. 5 Excitation and emission spectrum of Dy^{3+} doped BZLP glass at λ_{em} = 575 and λ_{ex} = 350 nm respectively. The inset plot represents the change in emission intensity with Dy^{3+} ions concentration in BZLP glasses

3.3 PL study of singly doped Dy^{3+} ions in BZLP glasses

PL characteristics of 0.1 mol% Dy^{3+} doped BZLP glass was presented in Fig. 5. PL excitation spectrum with monitoring the 575 nm emission wavelength contains various peaks in 300–500 nm range. The excitation peaks were observed at 322, 337, 350, 362, 386, 425, 449 and 473 nm corresponding transitions are mentioned in Fig. 5. Amongst the excitation peaks, the dominant peak at 350 nm ($^6H_{15/2} \rightarrow {}^6P_{7/2}$) transition was chosen as the excitation wavelength to record the emission characteristics [31, 32].

Figure 5 shows the emission spectrum of 0.1 mol% Dy^{3+} doped BZLP glass under the excitation of 350 nm wavelength. The emission spectrum exhibits the blue, yellow and red peaks at 483 nm ($^4F_{9/2} \rightarrow {}^6H_{15/2}$), 575 nm ($^4F_{9/2} \rightarrow {}^6H_{13/2}$) and 661 nm ($^4F_{9/2} \rightarrow {}^6H_{11/2}$). The nature of transitions $^4F_{9/2} \rightarrow {}^6H_{15/2}$ and $^4F_{9/2} \rightarrow {}^6H_{13/2}$ is magnetic dipole and forced electric dipole respectively. The $^4F_{9/2} \rightarrow {}^6H_{11/2}$ transition, on the other hand, is both magnetic as well as in forced electric in nature [33]. The ($^4F_{9/2} \rightarrow {}^6H_{13/2}$) transition is the leading the all-emission peaks. Further, the emission intensity can depend on the doping concentrations. Hence the doping concentration of Dy^{3+} ion was optimized by preparing a series of glass with 0.1 to 2.0 mol% of Dy^{3+} ions. The inset shows the variation in emission intensity with increasing Dy^{3+} concentration. The emission intensity upsurges till 1.5 mol% of Dy^{3+} and then declines [34]. This was due to mutual interactions and resonant energy transfer among the doping ions in BZLP i.e., concentration quenching effect [34]. On the basis of the emission profile, the 1.5 mol% is the optimized concentration of Dy^{3+} ions in BZLP glasses.

3.4 PL study of singly doped Eu^{3+} ions in BZLP glasses

Figure 6 shows the luminescence profile of Eu^{3+} doped BZLP glass at $\lambda_{em} = 612$ and $\lambda_{ex} = 393$ nm respectively. The excitation profile shows a number of peaks in 300–500 nm range owing to the ($^7F_0 \rightarrow {}^5H_6$), ($^7F_0 \rightarrow {}^5D_4$), ($^7F_0 \rightarrow {}^5L_7$), ($^7F_0 \rightarrow {}^5L_6$), ($^7F_0 \rightarrow {}^5D_3$) and ($^7F_0 \rightarrow {}^5D_2$), Eu^{3+} ions transitions. Amongst the several excitations, the dominant peak was observed at 393 nm corresponding to $^7F_0 \rightarrow {}^5L_6$ transition. Emission profile is illustrated in 500–750 nm range at $\lambda_{ex} = 393$ nm in the same figure. The emission profile shows numerous sharp peaks attributed to ($^5D_0 \rightarrow {}^7F_0$), ($^5D_0 \rightarrow {}^7F_1$), ($^5D_0 \rightarrow {}^7F_2$), ($^5D_0 \rightarrow {}^7F_3$), and ($^5D_0 \rightarrow {}^7F_4$) Eu^{3+} ions transitions. The transitions $^5D_0 \rightarrow {}^7F_1$ and $^5D_0 \rightarrow {}^7F_2$ of Eu^{3+} ions can be characterised as magnetic and forced electric dipole in nature, respectively. The dominant emission peak was situated at 612 nm under 393 nm excitation [35].

3.5 PL study of co-doped BZLP:Dy^{3+}/Eu^{3+} glasses

The excitation spectra of Dy^{3+}/Eu^{3+} co-doped BZLP glass (BZLP:DE1) were recorded by monitoring the emission wavelengths at 575 and 612 nm as

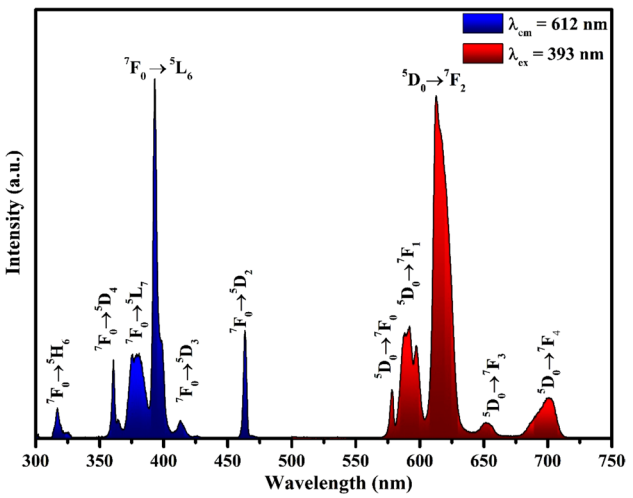


Fig. 6 Excitation and emission spectrum of 1.0 mol% Eu^{3+} doped BZLP glass at $\lambda_{em} = 612$ and $\lambda_{ex} = 393$ nm respectively

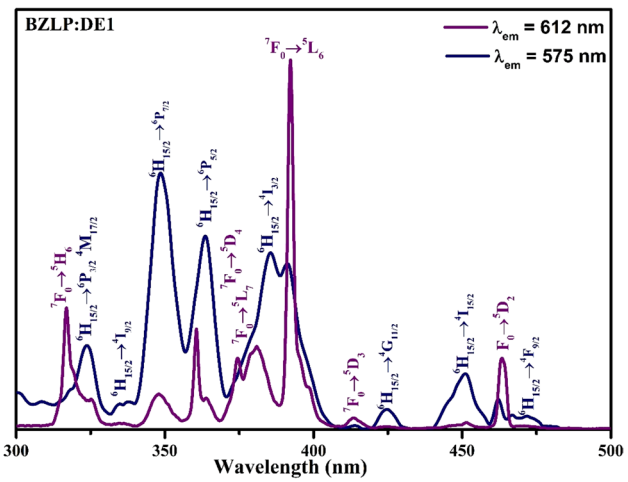


Fig. 7 Excitation spectra of BZLP:DE1 glass at $\lambda_{em} = 575$ and 612 nm

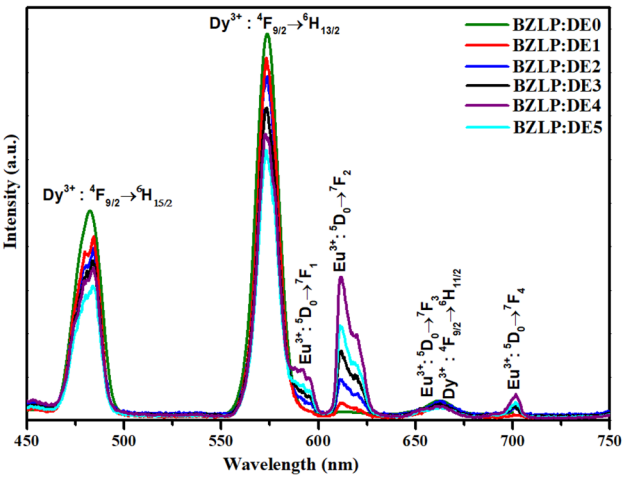


Fig. 8 Emission spectra of Dy^{3+}/Eu^{3+}co-doped BZLP glasses at $\lambda_{ex} = 350$ nm

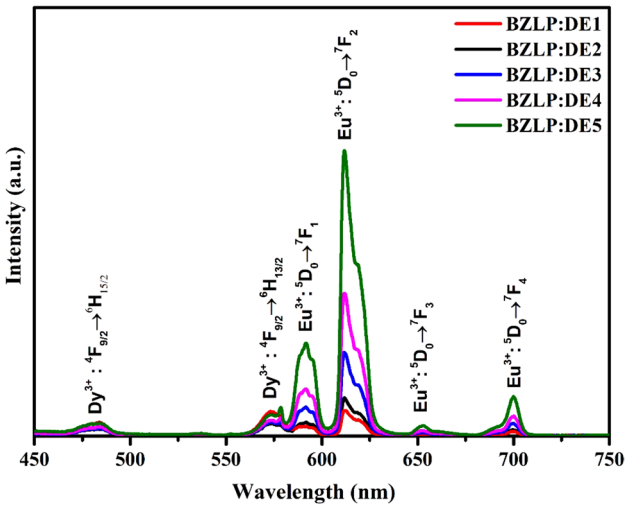


Fig. 9 Emission spectra of Dy^{3+}/Eu^{3+}co-doped BZLP glasses at $\lambda_{ex} = 393$ nm

illustrated in Fig. 7. The excitation profile by monitoring the emission wavelength at 575 nm exhibits a number of excitation peaks in 300–500 nm range, owing to the transition from $^6H_{15/2}$ to $^6P_{3/2}$, $^4M_{17/2}$, $^4I_{9/2}$, $^6P_{7/2}$, $^6P_{5/2}$, $^4I_{3/2}$, $^4G_{11/2}$, $^4I_{15/2}$ and $^4F_{9/2}$ of Dy^{3+} ions. Also, some additional peaks were detected at 393 and 465 due to the $^7F_0 \rightarrow {}^5L_6$ and $^5D_2$ transitions of Eu^{3+} ions. The occurrence of excitation peaks related to Eu^{3+} ions while monitoring the Dy^{3+} ions emission peaks suggested that the Dy^{3+} plays an important role in energy transfer to the Eu^{3+} ion. The excitation profile while monitoring 612 nm emission exhibits excitation peaks due to the $^7F_0 \rightarrow {}^5H_6$, $^5D_4$, $^5L_7$, $^5L_6$, $^5D_3$ and $^5D_2$, transitions of Eu^{3+} ions. The overlap of the excitation spectra of BZLP: DE1 confirms the ability of the glass to be excited by n-UV radiation.

Dy^{3+}/Eu^{3+} co-doped glasses were excited by 350 nm wavelength and the corresponding emission spectra are shown in Fig. 8. The emission profile changes with increasing content of Eu^{3+} ions in the glasses. The peak intensity related to $^4F_{9/2} \rightarrow {}^6H_{15/2}$, $^6H_{13/2}$ and $^6H_{11/2}$ transitions of Dy^{3+} reduces with increasing Eu^{3+} ions content whereas new peaks attributed to $^5D_0 \rightarrow {}^7F_1$, $^7F_2$, $^7F_3$, and $^7F_4$ transitions of Eu^{3+} ions are observed in the spectra. The reduction in Dy^{3+} emission peak intensity and simultaneous enrichment in the Eu^{3+} ions emission intensity signifies energy transfer between sensitizer Dy^{3+} and activator Eu^{3+} ions [36].

Furthermore, the PL spectra of Dy^{3+}/Eu^{3+} co-doped BZLP glasses were recorded under 393 nm excitation as revealed in Fig. 9. Numerous emission peaks situated at 591($^5D_0 \rightarrow {}^7F_1$), 612 ($^5D_0 \rightarrow {}^7F_2$), 651($^5D_0 \rightarrow {}^7F_3$) and 700 ($^5D_0 \rightarrow {}^7F_4$) nm of Eu^{3+} ion can be seen. Apart from this, two peaks were also detected at 483 and 575 nm corresponding to $^4F_{9/2} \rightarrow {}^6H_{15/2}$ and $^6H_{13/2}$ transitions of Dy^{3+} ions, respectively. Peaks corresponding to Eu^{3+} ion transitions are more intense than those of Dy^{3+} ion transitions. The intensity of $^5D_0 \rightarrow {}^7F_1$,$^7F_2$, $^7F_3$,$^7F_4$ transitions enhances with increase in Eu^{3+} ions, whereas minute changes are observed in emission related with to $^4F_{9/2} \rightarrow {}^6H_{15/2}$, $^6H_{13/2}$ transitions of Dy^{3+} ions. Thus, Dy^{3+} ions sensitize Eu^{3+} ions as is evident from PL studies and transfer a part of absorbed energy which is converted to visible emission [37, 38].

A partial energy level diagram of Dy^{3+} and Eu^{3+} ions is shown in Fig. 10 to realize the sensitizer and activator ions energy transfer process. Dy^{3+} and Eu^{3+} ions get excited to upper energy levels by n-UV radiation and then fall to lower energy states via non-radiative decay. Subsequently, the Dy^{3+} ions give emissions at 483, 575 and 661 nm ascribed to the $^4F_{9/2} \rightarrow {}^6H_{15/2}$, $6H_{13/2}$ and $^6H_{11/2}$ transitions whereas Eu^{3+} emits owing to $^5D_0 \rightarrow {}^7F_1$, $^7F_2$, $^7F_3$, and $^7F_4$ transitions. Some energy is also shifted from Dy^{3+} to Eu^{3+} ions along with the individual emissions. The energy shifting takes place as the energy level $^4I_{15/2}$ of Dy^{3+} ion is above the $^5D_2$, $^5D_1$ and $^5D_0$ energy levels of Eu^{3+}

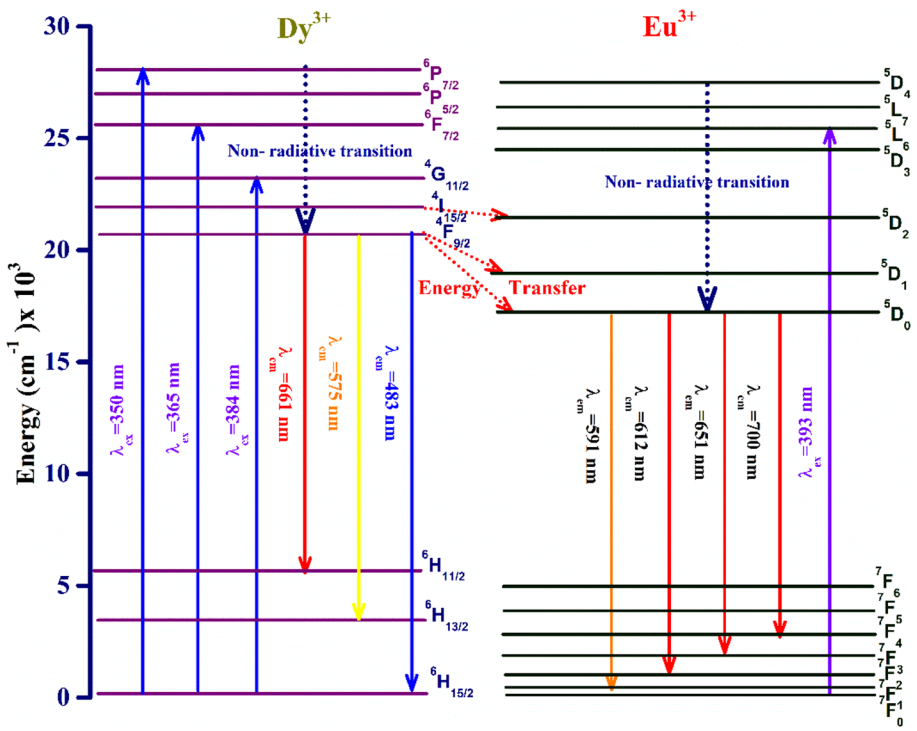


Fig. 10 Partial energy level diagram of Dy^{3+} and Eu^{3+} ions in BZLP glasses

ion. Also, the $^4F_{9/2}$ level of Dy^{3+} lies above the $^5D_1$ and $^5D_0$ levels of Eu^{3+} ions. Owing to the less energy difference between these levels, a part of absorbed energy from Dy^{3+} ions migrate to the Eu^{3+} ions in BZLP glasses as seen in PL spectra.

Mainly multipolar exchange interaction type of energy transfer takes places in rare earth doped glasses, which can be categorized into three ways of interaction i.e., quadrupole–quadrupole, dipole–dipole and dipole-quadrupole. The types of energy transfer among the nearest doped ions were analysed via Dexter's multipolar exchange interaction expression with Reisfeld's approximation as follows [39]:

$$\frac{\eta_0}{\eta} \propto C^{n/3} \tag{1}$$

η_0 and η symbolize the radiant quantum efficiency of sensitizer Dy^{3+} ions in the absence of an activator and quantum efficiency of sensitizer Dy^{3+} ions with Eu^{3+} (activator), respectively. C signifies the sum of concentration of dopant ions. The nature of multipolar exchange interaction is indicated by n in the equation above. Depending whether the interaction is dipole–dipole, dipole-quadrupole and quadrupole–quadrupole, the value of n can be 6, 8, or 10 respectively. The radiation intensities are linked to the ratio between η_0 and η as shown below [40]:

$$\frac{I_{SO}}{I_s} \propto C^{n/3} \tag{2}$$

In the above equation, I_{SO} signifies emission intensity of Dy^{3+} doped glass and I_S represent the emission intensity of Dy^{3+}/Eu^{3+}co-doped glass. The graphs of I_{SO}/I_S versus $C^{n/3}$ taking n = 6, 8, 10 can be seen in Fig. 11. The non-radiative, dipole-to-dipole character of the interaction and energy transfer between the rare earth ions was confirmed by the best linear fit, which was found for n = 6.

3.6 Lifetime and energy transfer analysis

The lifetime curves of Dy^{3+}/ Eu^{3+}co-doped BZLP glasses were recorded at λ_{em} = 572 nm under excitation λ_{ex} = 350 nm and shown in Fig. 12. The lifetime curves are exponential in nature and can be fitted with bi-exponential function as given below [41]:

$$I = I_0 + A_1 exp\left(-\frac{t}{\tau_1}\right) + A_2 exp\left(-\frac{t}{\tau_2}\right) \tag{3}$$

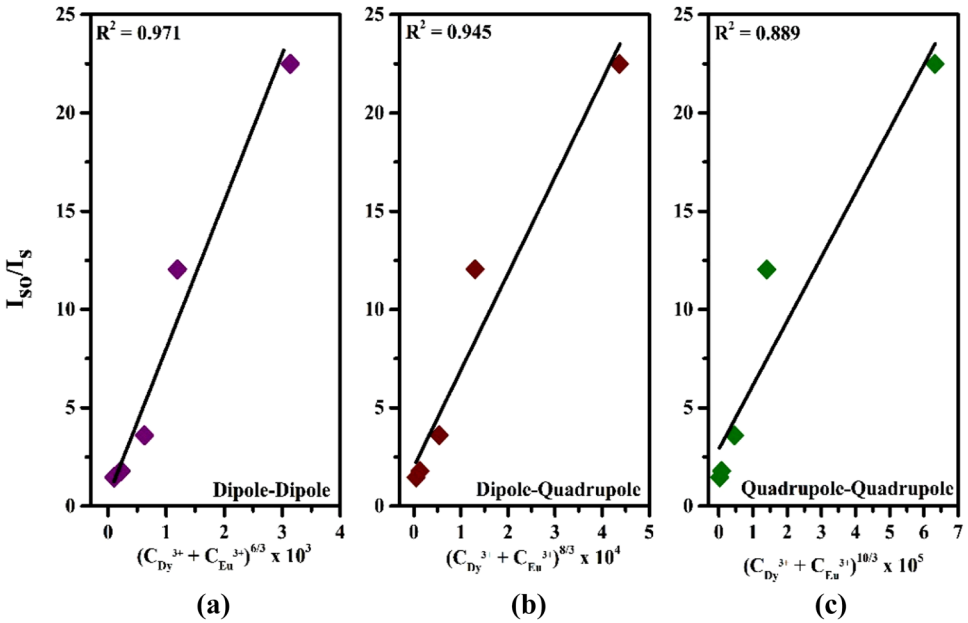


Fig. 11 Plot between I_{SO}/I_S versus (a) $(C_{Dy+Eu})^{6/3}$, (b)$(C_{Dy+Eu})^{8/3}$ and (c)$(C_{Dy+Eu})^{10/3}$

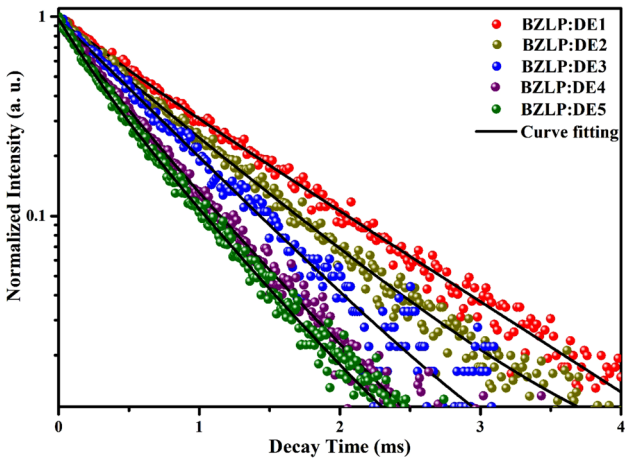


Fig. 12 Decay curves of Dy^{3+}/Eu^{3+}co-doped BZLP glasses at $\lambda_{ex}=350$ & $\lambda_{em}=575$ nm

Here I and I_0 denote the emission intensity at any instant t and t = 0, respectively and τ_1 and τ_2 denote the slow and fast decay components. A_1 and A_2 denote the fitting constants. On the basis of these parameters, the average lifetime (τ_{avg}) were calculated using the formula as mentioned [42]:

$$\tau_{avg} = \frac{A_1\tau_1^2 + A_2\tau_2^2}{A_1\tau_1 + A_2\tau_2} \tag{4}$$

The evaluated τ_{avg} values for the prepared Dy^{3+}/Eu^{3+}co-doped BZLP glasses are tabulated in Table 1. The occurrence of energy transfer between the dopant ions is confirmed by the decrease in the τ_{avg} values with increasing Eu^{3+} concentration. η_T is the energy transfer efficiency probability given as [43]:

$$\eta_T = 1 - \frac{\tau_d}{\tau_{d0}} \tag{5}$$

Table 1 CIE color coordinates (*x, y*), CCT (K), lifetime, Dy^{3+} to Eu^{3+} energy transfer efficiency of prepared co-doped BZLP glasses at $\lambda_{ex}=350$ and $\lambda_{em}=575$ nm

Sample name	CIE coordinates (*x, y*)	CCT (K)	Decay time (μs)	Energy transfer efficiency η_T(%)
BZLP:DE0	(0.372, 0.424)	4474	589	–
BZLP:DE1	(0.383, 0.410)	4125	571	3.05
BZLP:DE2	(0.376, 0.416)	4340	543	7.80
BZLP:DE3	(0.378, 0.405)	4185	524	11.03
BZLP:DE4	(0.383, 0.394)	3432	504	14.43
BZLP:DE5	(0.402, 0.391)	3180	496	15.78

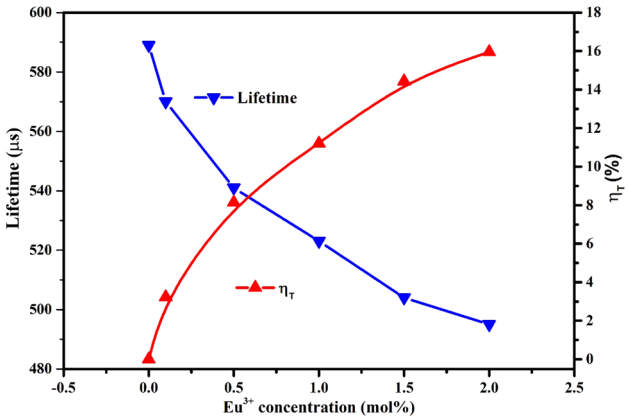


Fig. 13 Variation in lifetime and ET efficiency with Eu^{3+} ions concentration in Dy^{3+}/Eu^{3+}co-doped BZLP glasses

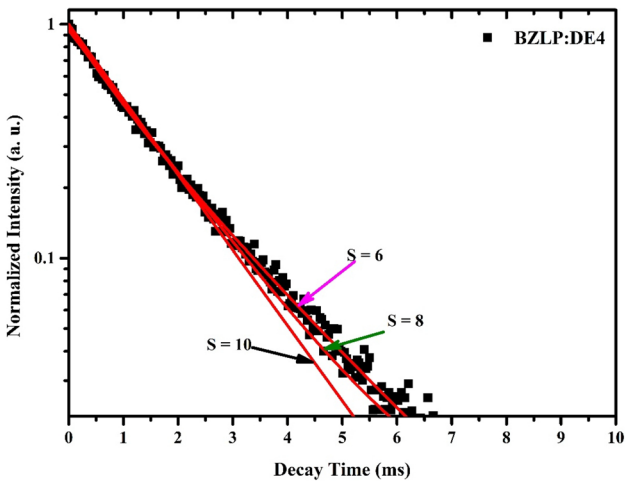


Fig. 14 I-H fitted decay curves of BZLP:DE4 glass

Here τ_d and τ_{d0} denote the lifetime of the co-doped and singly doped glasses respectively. Table 1 provides a tabulation of the energy transfer probability (η_T) parameter as well. It can be seen that when concentration of Eu^{3+} ions increases, η_T increases as well. The likelihood of energy transfer seems to increase by five times for an increase in Eu^{3+} concentration from 0.0 mol% to 5.0 mol%. Figure 13 demonstrates how lifetime and energy transfer efficiency in co-doped BZLP glasses change when the Eu^{3+} ion concentration in the glass lattice increases.

Furthermore, on the basis of PL decay curve of BZLP:DE4, the type of energy transfer interaction has been analysed using Inokuti- Hirayama (I-H) model. As per the I-H model, the equation which relates the type of interaction as follows [44]:

$$I_t = I_0 \exp\left\{\frac{-t}{\tau_o} - Q\left(\frac{t}{\tau_o}\right)^{\frac{3}{s}}\right\} \quad (6)$$

Here I_t is intensity of luminescence at time t, τ_o signifies the decay time, Q signifies the energy transfer parameter and the value of S designates the type of interaction having values 10, 8 and 6 for quadrupole–quadrupole, dipole-quadrupole and dipole–dipole interaction, respectively [44]. The best fit is observed for $S = 6$ value as presented in Fig. 14. Thus, the I-H models also confirm the occurrence of dipole–dipole interaction among the sensitizer and activator ions in BZLP glasses.

3.7 Colorimetric analysis

Table 1 shows the CIE coordinates of the as prepared glasses. The color coordinates depend upon the incident excitation wavelength and also on the amount of dopants in the glass. Varying Eu^{3+} concentration leads to the modulation of luminescence hues and is helpful for a variety of applications. CIE chromaticity coordinates of Dy^{3+}/Eu^{3+}co-doped BZLP glasses at $\lambda_{ex} = 350$ nm correspond to the white region as seen in Fig. 15a. The coordinates for BZLP:DE5 glass at $\lambda_{ex} = 393$ nm were also plotted on the CIE diagram and shown in Fig. 15b. CIE color coordinates were situated in red region, under the $\lambda_{ex} = 393$ nm, which shows the prepared glasses can be a potential red constituent for many photonic devices.

Correlated colour temperature (CCT) is also a critical parameter which allows the researchers to determine the specific temperatures for which the CIE coordinates fall closer to the whole light region. It can be calculated using the McCamy relation [45]. The CCT values for Dy^{3+}/Eu^{3+} doped BZLP glass have been presented in Table 1 as well. The values signify emission tuning from cool to warm white light regions with the addition of Eu^{3+} ions. According to the aforementioned findings, CIE colour co-ordinates and CCT values may be effectively modified by the amount of doped RE ions BZLP glasses. In turn, this demonstrates that as-prepared Dy^{3+}/Eu^{3+} co-doped BZLP glasses are appropriate for n-UV pumped white illumination systems.

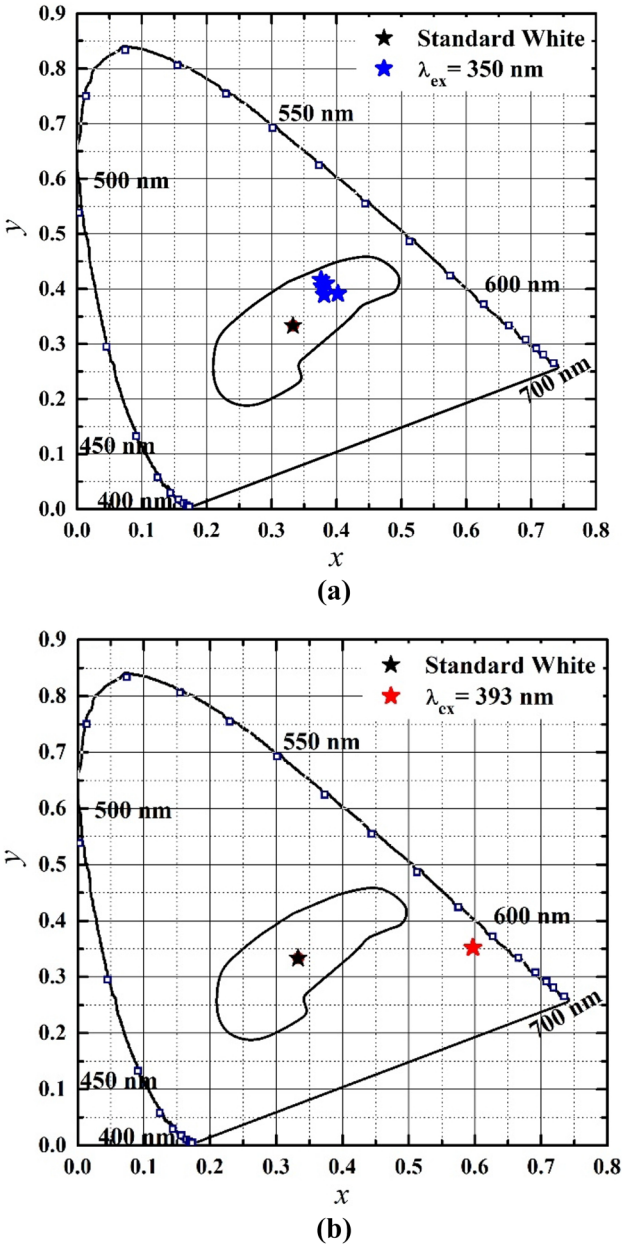


Fig. 15 CIE chromaticity coordinates of **a** Dy^{3+}/Eu^{3+}co-doped BZLP glasses at λ_{ex} = 350 nm. **b** BZLP:DE5 glass at λ_{ex} = 393 nm

3.8 Temperature dependent PL study

PL spectra are affected by variation of temperature. Hence, the temperature-dependent emission characteristics for Dy^{3+}/Eu^{3+} co-doped BZLP glass were recorded at λ_{ex} = 350 & 393 nm and shown in Fig. 16a and b. With increase in temperature from 25 to 200 °C, the PL intensity continuously decreases as can be seen in the figure. The intensity is 93.40% of maximum at 150 °C and 91.25% of maximum at 200 °C, which shows that the Dy^{3+}/Eu^{3+} co-doped BZLP glass has very good thermal stability. Furthermore, the activation energy (ΔE_a) can be accessed by using the Arrhenius equation as follows [46]:

$$I_T = \frac{I_0}{1 + Cexp\left(-\frac{\Delta E}{K_B T}\right)}$$

Here I_0 and I_T indicate the emission intensity at 25 °C and at a particular temperature T respectively whereas CandK_B signify an arbitrary constant and Boltzmann constant respectively. On the basis of temperature-dependent emission characteristics, the $ln((I_0/I_T)\text{-}1)$ versus $1/K_BT$ plots were analysed and fitted with linear equation for the both emission characteristics. The plots are shown in Fig. 17a and b. The activation energy for Dy^{3+}/Eu^{3+} co-doped BZLP glass was assessed to be 0.242 and 0.241 eV for emission spectra under λ_{ex} = 350 & 393 nm respectively which have relatively higher activation energy (0.196 eV) of reported glass in literature [3]. Thus, temperature dependent emission characteristics show that the Dy^{3+}/Eu^{3+} co-doped BZLP glass can produce thermally stable luminescence and can be used as red emitting component in w-LEDs and other luminescent devices.

4 Conclusions

Transparent Dy^{3+}/Eu^{3+} co-doped BZLP glasses were effectively synthesized via melt quench technique. The structural and luminous properties were examined to gain understanding of their applicability in photonic devices. The single Dy^{3+} doped BZLP glasses emits blue, yellow and red emission under n-UV excitation pertaining to the $^4F_{9/2} \rightarrow {}^6H_{15/2}$, $^6H_{13/2}$ and $^6H_{11/2}$ transitions respectively. Dy^{3+}/Eu^{3+} co-doped BZLP glasses give emission in visible region owing $^4F_{9/2} \rightarrow {}^6H_{15/2}$, $^6H_{13/2}$ and $^6H_{11/2}$ transitions of Dy $^{3+}$ along with $^5D_0 \rightarrow {}^7F_1$, $^5D_0 \rightarrow {}^7F_2$, $^5D_0 \rightarrow {}^7F_3$ and $^5D_0 \rightarrow {}^7F_4$ transitions of Eu $^{3+}$ ion under 350 nm and 393 nm.With an increase in Eu^{3+} ion concentration, energy transfer takes place from Dy^{3+} ions to Eu^{3+} ions resulting in the decrement of intensity of PL peaks of Dy^{3+} ions. The cool white emission shifts towards the warm white emission under 350 nm and 393 nm excitation for the co-doped glasses as can be seen in the CIE diagram.

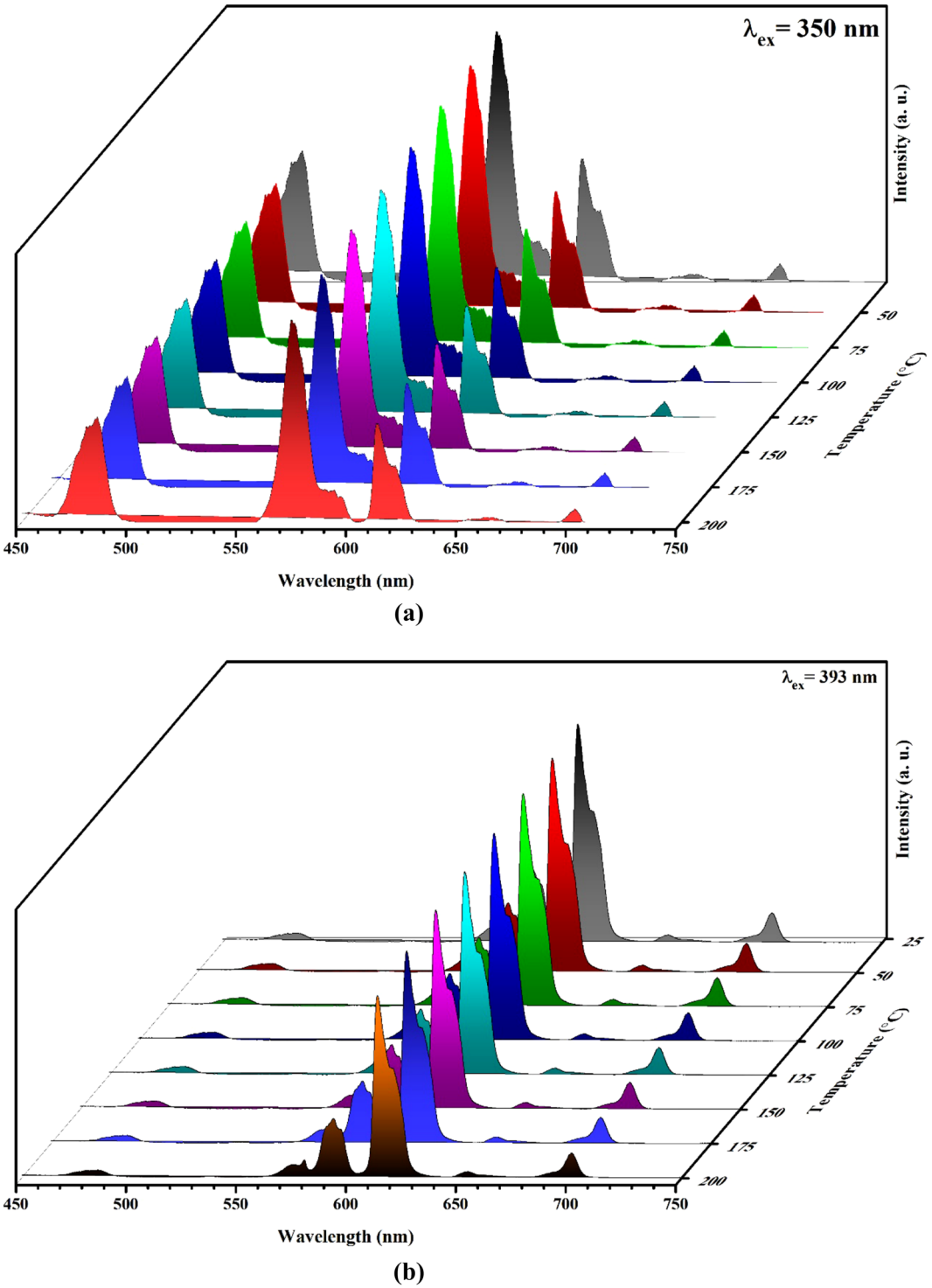


Fig. 16 Temperature dependent emission spectra of **a** BZLP:DE4 glass at $\lambda_{ex} = 350$ nm. **b** BZLP:DE5 glass at $\lambda_{ex} = 393$ nm

The dipole–dipole type of energy transfer sensitizer and activator ions in the BZLP glasses was verified using Dexter's ET formula, Reisfeld's approximation and I–H model theories. The PL lifetimes for the co-doped glasses reduced whereas the energy transfer efficiency η_T enflamed with upsurge in Eu^{3+} ions concentration from $y = 0.0$ to 5.0 mol%. The temperature-dependent PL spectra under 350 nm and 393 nm

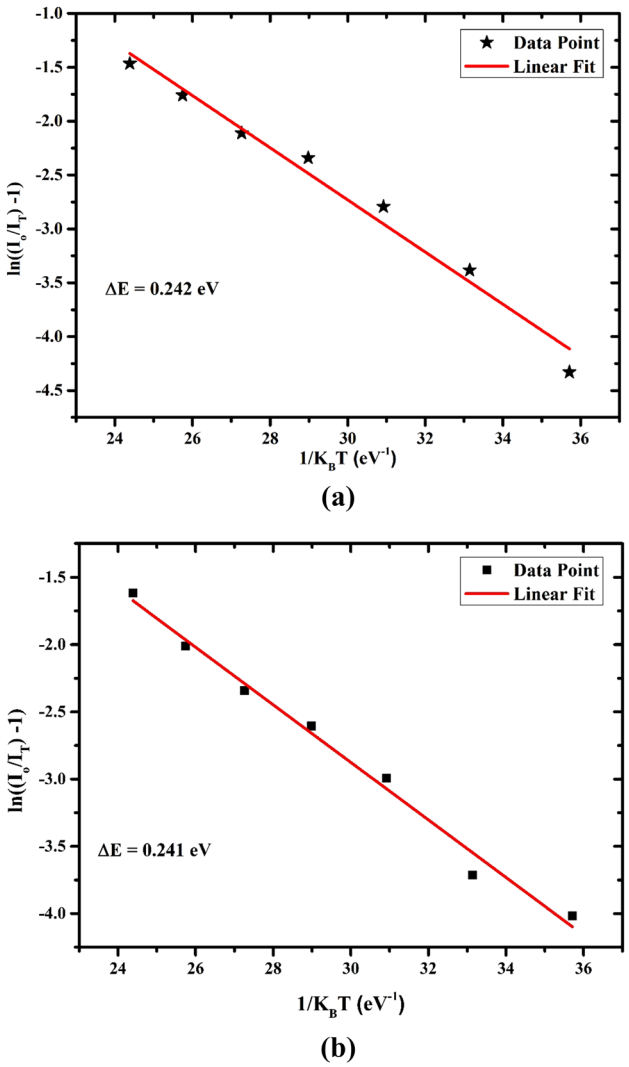


Fig. 17 $\ln[(I_0/I_T)-1]$ versus $1/K_BT$ plot for **a** BZLP:DE4 glass at $\lambda_{ex} = 350$ nm. **b** BZLP:DE5 glass at $\lambda_{ex} = 393$ nm

excitation were thermally stable with values of activation energy being 0.242 eV and 0.241 eV, respectively. The aforementioned results for Dy^{3+}/Eu^{3+} co-doped BZLP glasses show that these glasses can work is a red light emitting component in w-LEDs and for warm white light generation in various luminescent applications.

Author contributions

All authors contributed to the study, conception and design of the manuscript. The author credits are as follows: KM: Data collection, methodology, synthesis, result analysis, writing. AP: Result analysis, review, writing. RP: Result analysis, writing. SM: Result analysis. YT: Data collection, synthesis. ASR: Methodology, review, editing, supervision.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

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

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.

Research involving human and animal rights The authors declare that no animal or animal data was used in any study or experiment mentioned in this manuscript.

References

1. Y. Yan, Z. Chen, X. Jia, S. Li, Investigations on the effects of the Stark splitting on the fluorescence behaviors in Yb^{3+}-doped silicate, tellurite, germanate, and phosphate glasses. Opt. Mater. **75**, 465–470 (2018). https://doi.org/10.1016/j.optmat.2017.10.008
2. L. He, D. Wang, M. Chen, R. Liu, Z. Su, C. Li, H. Lin, S. Li, Y. Zhou, L. Liu, F. Zeng, Luminescence properties of Tb^{3+}/Eu^{3+} co-doped glass based on GSBR system. Opt. Mater. **122**, 111754 (2021). https://doi.org/10.1016/j.optmat.2021.111754
3. A.S. Rao, Effective energy transfer from Dy^{3+} to Tb^{3+} ions in thermally stable KZABS glasses for intense green emitting device applications. J. Lumin. **239**, 118325 (2021). https://doi.org/10.1016/j.jlumin.2021.118325

4. C. Zhang, H. Liang, S. Zhang, C. Liu, D. Hou, L. Zhou, G. Zhang, J. Shi, Efficient Sensitization of Eu^{3+} Emission by Tb^{3+} in $Ba_3La(PO_4)_3$ under VUV-UV excitation: energy transfer and tunable emission. J. Phys. Chem. C **116**, 15932–15937 (2012). https://doi.org/10.1021/jp304717z
5. N. Baig, R.S. Yadav, N.S. Dhoble, V.L. Barai, S.J. Dhoble, Near UV excited multi-color photoluminescence in RE^{3+} (RE=Tb, Sm, Dy and Eu) doped $Ca_2Pb_3(PO_4)_3Cl$ phosphors. J. Lumin. **215**, 116645 (2019). https://doi.org/10.1016/j.jlumin.2019.116645
6. C. Sumalatha, R. Doddoji, M. Venkateswarlu, P.R. Rani, K. Swapna, S. Mahamuda, A.S. Rao, White light emission from Dy^{3+}-doped ZnO+Bi_2O_3+BaF_2+B2O_3+TeO_2 glasses: structural and spectroscopic properties. Spectro. Acta Part A: Mol. Biomol. Spectrosc. **240**, 118568 (2020). https://doi.org/10.1016/j.saa.2020.118568
7. P.R. Rani, M. Venkateswarlu, S. Mahamuda, K. Swapna, N. Deopa, A.S. Rao, Spectroscopic studies of Dy^{3+} ions doped barium lead alumina fluoro borate glasses. J. Alloys Compd. **787**, 503–518 (2019). https://doi.org/10.1016/j.jallcom.2019.02.088
8. H. Bouchouicha, G. Panczer, D. De Ligny, Y. Guyot, M.L. Baesso, L.H. Andrade, S.M. Lima, R. Ternane, Synthesis and luminescent properties of Eu^{3+}/Eu^{2+} co-doped calcium aluminosilicate glass–ceramics. J. Lumin. **169**, 528–533 (2016). https://doi.org/10.1016/j.jlumin.2014.11.054
9. A.S. Rao, R.R. Reddy, T.R. Rao, J.L. Rao, Electron paramagnetic resonance and optical absorption spectra of Fe^{3+} ions in alkali cadmium borosulphate glasses. Solid State. Comm. **96**, 701–705 (1995). https://doi.org/10.1016/0038-1098(95)00375-4
10. B.H. Babu, V.R.K. Kumar, White light generation in $Ce^{3+}Tb^{3+}$ Sm^{3+} codoped oxyfluoroborate glasses. J. Lumin **154**, 334–338 (2014). https://doi.org/10.1016/j.jlumin.2014.05.010
11. S. Ye, F. Xiao, Y.X. Pan, Y.Y. Ma, Q.Y. Zhang, Phosphors in phosphor-converted white light-emitting diodes: recent advances in materials, techniques and properties. Mater. Sci. Eng. **71**(1), 1–34 (2010). https://doi.org/10.1016/j.mser.2010.07.001
12. H. Masai, Y. Yamada, Y. Suzuki, K. Teramura, Y. Kanemitsu, T. Yoko, Narrow energy gap between triplet and singlet excited states of Sn21 in borate glass. Sci. Rep. **3**, 3541 (2013). https://doi.org/10.1038/srep03541
13. N. Deopa, S. Saini, S. Kaur, A. Prasad, A.S. Rao, Spectroscopic investigations on Dy^{3+} ions doped zinc lead alumino borate glasses for photonic device applications. J. Rare Earths **37**, 52–59 (2019). https://doi.org/10.1016/j.jre.2018.04.013
14. N. Deopa, A.S. Rao, S.K. Mahamuda, M. Gupta, M. Jayasimhadri, D. Haranath, G. Vijaya Prakash, Spectroscopic studies of Pr3+ doped lithium lead alumino borate glasses for visible reddish orange luminescent device applications. J. Alloys Compd. **708**, 911–921 (2017). https://doi.org/10.1016/j.jallcom.2017.03.020
15. G. Lakshminarayana, S. Buddhudu, Spectral analysis of Eu^{3+} and Tb^{3+}:B_2O_3–ZnO–PbO glasses. Mater. Chem. Phys. **102**, 181–186 (2007). https://doi.org/10.1016/j.matchemphys.2006.11.020
16. R.J. Amjad, M.R. Dousti, M.R. Sahar, Spectroscopic investigation and Judd-Ofelt analysis of silver nanoparticles embedded Er^{3+}-doped tellurite glass. Curr. Appl. Phys. **15**, 1–7 (2015). https://doi.org/10.1016/j.cap.2014.10.022
17. M. Higuchi, R. Sasaki, J. Takahashi, Float zone growth of Dy:$GdVO_4$ single crystals for potential use in solid-state yellow lasers. J. Cryst. Growth **311**, 4549–4552 (2009). https://doi.org/10.1016/j.jcrysgro.2009.08.028
18. K. Siva Rama Krishnareddy, K. Swapna, S.K. Mahamuda, M. Venkateswarlu, A.S. Rao, G. Vijayaprakash, Investigation on structural and luminescence features of Dy^{3+} ions doped alkaline-earth boro tellurite glasses for optoelectronic devices. Opt. Mater. **85**, 200–210 (2018). https://doi.org/10.1016/j.optmat.2018.08.057
19. P. Manasa, C.K. Jayasankar, Spectroscopic assessment of Dy^{3+} ions in lead fluorosilicate glass as a prospective material for solid state yellow laser, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. **212**, 315–321 (2019). https://doi.org/10.1016/j.saa.2019.01.015
20. Q. Luo, X. Qiao, X. Fan, H. Yang, X. Zhang, S. Cui, L. Wang, G. Wang, Luminescence behavior of Ce^{3+} and Dy^{3+} codoped oxyfluoride glasses and glass ceramics containing LaF_3 nanocrystals. J. Appl. Phys. **105**(043506), 1–5 (2009). https://doi.org/10.1063/1.3077266
21. N. Deopa, A.S. Rao, Spectroscopic studies of single near ultraviolet pumped Tb^{3+} doped Lithium Lead Alumino Borate glasses for green lasers and tricolour w-LEDs. J. Lumin. **194**, 56–63 (2018). https://doi.org/10.1016/j.jlumin.2017.09.057
22. A. Vidhi, Anu, AS Rao, Spectroscopic characterizations of Dy^{3+} ions doped phosphate glasses for epoxy-free white LED applications. Opt. Mater. **132**, 112863 (2022). https://doi.org/10.1016/j.optmat.2022.112863
23. Y. Tayal, A.S. Rao, Spectroscopic analysis of Dy^{3+} ions activated borosilicate glasses for photonic device applications. Opt. Mater. **117**, 111112 (2021). https://doi.org/10.1016/j.optmat.2021.111112
24. Z. Zhu, Y. Zhang, Y. Qiao, H. Liu, D. Liu, Luminescence properties of Ce^{3+}/Tb^{3+}/Sm^{3+} co-doped CaO–SiO_2–B_2O_3

glasses for white light emitting diodes. J. Lumin. **134**, 724–728 (2013). https://doi.org/10.1016/j.jlumin.2012.07.003

25. A. Ravita, P. Prasad, A. Rohilla, A.S. Shandilya, Rao, Photoluminescence and energy transfer studies on Tm^{3+}/Dy^{3+}/Eu^{3+}doped borosilicate glasses for color tunability and warm white light generation. J. Non-Cryst. Solids **606**, 122192 (2023). https://doi.org/10.1016/j.jnoncrysol.2023.122192
26. U. Caldino, A. Speghini, E. Alvarez, S. Berneschi, M. Bettinelli, M. Brenci, G.C. Righini, Spectroscopic characterization and optical waveguide fabrication in Ce^{3+}, Tb^{3+} and Ce^{3+}/Tb^{3+} doped zinc–sodium–aluminosilicate glasses. Opt. Mater. **33**(12), 1892–1897 (2011). https://doi.org/10.1016/j.optmat.2011.03.012
27. N. Deopa, M.K. Sahu, P.R. Rani, R. Punia, A.S. Rao, Realization of warm white light and energy transfer studies of Dy^{3+}/Eu^{3+} co-doped Li_2O-PbO-Al_2O_3-B_2O_3 glasses for lighting applications. J. Lumin. **222**, 117166 (2020). https://doi.org/10.1016/j.jlumin.2020.117166
28. G. Lakshminarayana, H.C. Yang, J.R. Qiu, White light emission from Tm^{3+}/Dy^{3+} co-doped oxyfluoride germanate glasses under UV light excitation. J. Solid State Chem. **182**, 669–676 (2009). https://doi.org/10.1016/j.jssc.2008.11.020
29. E.A. Davis, N.F. Mott, Conduction in non-crystalline systems V. Conductivity, optical absorption and photoconductivity in amorphous semiconductors. Philos. Mag. **22**, 903–922 (1970). https://doi.org/10.1080/14786437008221061
30. A. Agarwal, V.P. Seth, S. Sanghi, P. Gahlot, D.R. Goyal, Optical band gap studies and estimation of two photon absorption coefficient in alkali bismuth borate glasses. Radiat. Eff. Defects Solids **158**, 793–801 (2003). https://doi.org/10.1080/10420150310001618214
31. P. Borisut, S. Insiripong, P. Limkitjaroenporn, P. Meejitpaisan, E. Kaewnuam, W. Chaipaksa, R. Rajaramakrishna, H.J. Kim, J. Kaewkhao, Luminescence and Judd-Ofelt analysis of gallium aluminum gadolinium yttrium borate scintillating glass doped with Dy^{3+}. Radiat. Phys. Chem. **199**, 110284 (2022). https://doi.org/10.1016/j.radphyschem.2022.110284
32. A.R. Venugopal, R. Rajaramakrishna, K.M. Rajashekara, J. Rajaguguk, N.H. Ayachit, S. Kothan, J. Kaewkhao, Dy^{3+} doped B_2O_3 – Li_2O – CaO – CaF_2 glass for efficient white light emitting sources. J. Non-Cryst. Solids **554**, 120604 (2021). https://doi.org/10.1016/j.jnoncrysol.2020.120604
33. N. Wantana, Y. Ruangtaweep, E. Kaewnuam, S.C. Kang, H.J. Kim, S. Kothan, J. Kaewkhao, Development of WO_3-Gd_2O_3- B_2O_3 high density glasses doped with Dy^{3+} for photonics and scintillation materials application. Solid State Sci. **101**, 106135 (2020). https://doi.org/10.1016/j.solidstatesciences.2020.106135
34. K. Maheshwari, A.S. Rao, Photoluminescence downshifting studies of thermally stable Dy^{3+} ions doped phosphate glasses for photonic device applications. Opt. Mater. **129**, 112518 (2022). https://doi.org/10.1016/j.optmat.2022.112518
35. M. Seshadri, M.J.V. Bell, V. Anjos, Y. Messaddeq, Influence of silver ions in Eu^{3+} doped glass for efficient reddish-orange and white light generation. J. Alloy. Compd. **838**, 155548 (2020). https://doi.org/10.1016/j.jallcom.2020.155548
36. D. Rajesh, K. Brahmachary, Y.C. Ratnakaram, N. Kiran, A.P. Baker, G.G. Wang, Energy transfer based emission analysis of Dy^{3+}/Eu^{3+} co-doped ZANP glasses for white LED applications. J. Alloys Compd. **646**, 1096–1103 (2015). https://doi.org/10.1016/j.jallcom.2015.05.138
37. D.A. Rodríguez-Carvajal, A.N. Meza-Rocha, U. Caldino, R. Lozada-Morales, E. Alvarez, Reddish-orange, neutral and warm white emissions in Eu^{3+}, Dy^{3+} and Dy^{3+}/Eu^{3+} doped CdO-GeO_2-TeO_2 glasses. Ma. Solid State Sci. **61**, 70–76 (2016). https://doi.org/10.1016/j.solidstatesciences.2016.09.009
38. T.A. Lodi, M. Sandrini, A.N. Medina, M.J. Barboza, F. Pedrochi, A. Steimacher, Dy: Eu doped CaBAl glasses for white light applications. Opt. Mater. **76**, 231–236 (2018). https://doi.org/10.1016/j.optmat.2017.12.043
39. Y. Tayal, A.S. Rao, S. Kaur, Photoluminescence characteristics of Sm^{3+}/Eu^{3+} co-doped LPZABS glasses for solar cell applications. Solid State Sci. **125**, 106834 (2022). https://doi.org/10.1016/j.solidstatesciences.2022.106834
40. K. Li, X. Liu, Y. Zhang, X. Li, H. Lian, J. Lin, Host-sensitized luminescence properties in $CaNb_2O_6$:Ln^{3+} (Ln^{3+} = Eu^{3+}/Tb^{3+}/Dy^{3+}/Sm^{3+}) phosphors with abundant colors. Inorg. Chem. **54**(1), 323–333 (2015). https://doi.org/10.1021/ic502493c
41. F. Kang, Y. Zhang, M. Peng, Controlling the energy transfer via multi luminescent centers to achieve white light/tunable emissions in a single-phased X2-Type Y_2SiO_5:Eu^{3+}, Bi^{3+} phosphor for ultraviolet converted LEDs. Inorg. Chem. **54**, 1462–1473 (2015). https://doi.org/10.1021/ic502439k
42. L. Mishra, A. Sharma, A.K. Vishwakarma, K. Jha, M. Jayasimhadri, B.V. Ratnam, K. Jang, A.S. Rao, R.K. Sinha, White light emission and color tunability of dysprosium doped barium silicate glasses. J. Lumin. **169**, 121–127 (2016). https://doi.org/10.1016/j.jlumin.2015.08.063
43. X. Min, Z. Huang, M. Fang, Y. Liu, C. Tang, X. Wu, Energy transfer from Sm^{3+} to Eu^{3+} in red-emitting phosphor $LaMgAl_{11}O_{19}$:Sm^{3+}, Eu^{3+} for solar cells and

near-ultraviolet white light-emitting diodes. Inorg. Chem. **53**, 6060–6065 (2014). https://doi.org/10.1021/ic500412r

44. S. Kaur, N. Deopa, A. Prasad, R. Bajaj, A.S. Rao, Intense green emission from Tb^{3+} ions doped zinc lead alumino borate glasses for laser and w-LED applications. Opt. Mater. **84**, 318–323 (2018). https://doi.org/10.1016/j.optmat.2018.07.020
45. C.S. McCamy, Correlated color temperature as an explicit function of chromaticity coordinates. Color. Res. Appl. **17**, 142–144 (1992). https://doi.org/10.1002/col.5080170211
46. P. Li, M. Peng, X. Yin, Z. Ma, G. Dong, Q. Zhang, J. Qiu, Temperature dependent red luminescence from a distorted Mn^{4+} site in $CaAl_4O_7$:Mn4+. Express. **21**, 18943–18948 (2013). https://doi.org/10.1364/OE.21.018943

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 Material Science and Mechanical Engineering (JMSME)
Print ISSN: 2393-9095; Online ISSN: 2393-9109; Volume 2, Number 8; April-June, 2015 pp. 81-83
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/jmsme.html

Thermodynamic Analysis of Single-effect LiBr-H_2O Absorption Refrigeration System

Rajesh Kumar*[1] and Ravindra Kannojiya[2]

[1,2]*Department of Mechanical Engineering, Delhi Technological University (Government of NCT of Delhi), Bawana Road, Delhi-110042, India*
*E-mail: *dr.rajeshmits@gmail.com*

Abstract—*Thermodynamic analysis of single-effect LiBr-H_2O refrigeration system has been performed to investigate the effect of temperature variation in the major components of the cycle. The effect of governing parameters such as generator temperature, absorber temperature, condenser temperature and evaporator temperature on COP, ECOP and circulation ratio has been observed and accordingly suitable application for the system has been found in the real world scenario.*

Keywords*: Energy, single-effect, absorption, circulation ratio.*

1. INTRODUCTION

The real world application of absorption refrigeration through single-effect lithium bromide-water absorption refrigeration technique has been an area of interest for researchers since last sixty years. Analysis of this system shows that performance of the system is limited to only low-temperature applications. For efficient utilization of the low-temperature heat sources, absorption refrigeration cycle has been analyzed to determine the change in performance of the system by changing some parameters such as temperature. He and Chen (2007) showed that the absorption refrigeration system, driven by low-potential thermal power such as solar energy gives potential of energy saving. It also have the benefits of using ecofriendly working pair, H_2O/LiBr and NH3/H_2O. Vereda et al. (2012) said that the potential of the absorption cycle for low-temperature waste heat recovery has been of good extent. It has application to solar cooling which enables this for the benefit due to the appreciative combination between solar energy and air-conditioning demand. Single-effect absorption cycle can operate with low driving temperatures, which supports the use of flat plate collectors. But it should not be lower than about 70˚C for air-conditioning. Stitou et al. (2000) observed that increase in the driving temperature for the process provides a comparatively better exploitation of the thermal power of heat source. The amount of heat at the top temperature can be degraded many times before being rejected to the sink, shows the way for possibility of cooling or heating from the machine. Tozer and James (1998) have said that two single stage cycles with one operating at a higher driving temperature can be considered equivalent to double stage cycle. They also showed that heat rejected from condenser must be equal to the heat absorbed by the low temperature generator.

2. SYSTEM DESCRIPTION

Fig.1 shows the single-effect LiBr-H_2O absorption refrigeration cycle. It consists of an absorber (A) and generator (G) which form the part of solution circuit. Condenser (C) and evaporator (E) are the parts of refrigeration circuit which produces cooling. Strong solution of refrigerant from heat exchanger (HE) enters into the generator and refrigerant water vapour sent to the condenser. The remaining part of the solution strong in LiBr reaches to the absorber. The absorber and evaporator are at low pressure levels while generators and condenser are at higher pressure level.

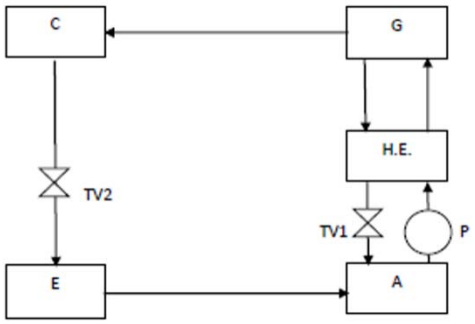


Fig. 1: Single-effect absorption cycle

Table 1: Main parameters considered for the analysis

Environment Temperature (K)	298
Environment pressure (MPa)	0.10135
AbsorberTemperature (K)	311
Heat exchanger effectiveness	0.60
Condenser temperature (K)	313
Evaporator temperature (K)	283
Generator temperature(K)	363
Mass flow rate of refrigerant vapour(kg/s)	1

3. THERMODYNAMIC ANALYSIS

The thermodynamic behavior of the absorption refrigeration system and its components have been studied. Part of the exergy supplied to an actual thermal system is destroyed due to irreversibility within the system. The exergy destruction is equal to the product of entropy generation within the system and the temperature of the reference environment.

The coefficient of performance of the cycle is given as:

$$COP = Q_e/Q_g\# \quad (1)$$

The exergetic coefficient of performance is given as:

$$ECOP = Q_e\{1-(To/T_e)\}/[Q_g\{1-(T_o/T_g)] \quad (2)$$

The circulation ratio of EARS is given as:

$$f = m_r/m_s\# \quad (3)$$

To find the steady-state performance of the system from these equations, the operating temperatures, weak and strong solution concentrations, effectiveness of heat exchanger and the refrigeration capacity. It is assumed that the solution at the exit of absorber and generator is at equilibrium so that P-T-ξ and h-T-ξ charts can be used for evaluating solution property data.

4. RESULTS AND DISCUSSION

A thermodynamic analysis has been done to observe the effect of temperature variation of various components of single-effect absorption system.

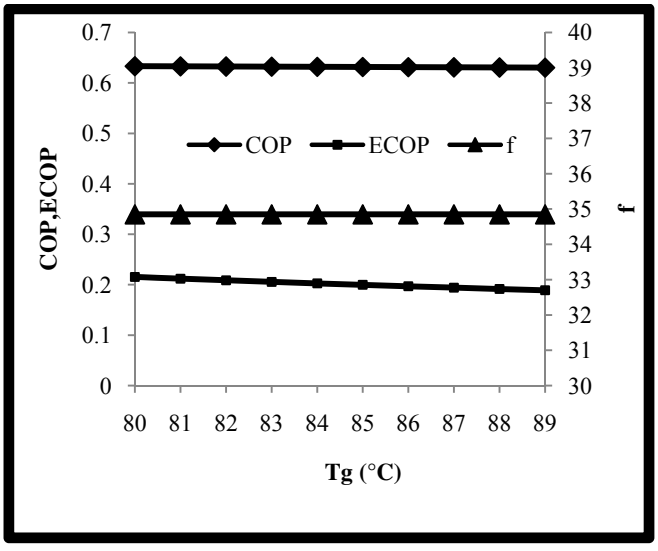


Fig. 2: Effect of generator temperature variation

Fig. 2, 3, 4 and 5 shows the effect on COP, ECOP and circulation ratio with temperature variation in generator, absorber, condenser and evaporator. The circulation ratio (f) remains the same as the temperature increases in case of condenser, evaporator and generator. It varies non-linearly in case of absorber. The COP increases in case of evaporator temperature but it decreases in other components like absorber, condenser and generator.

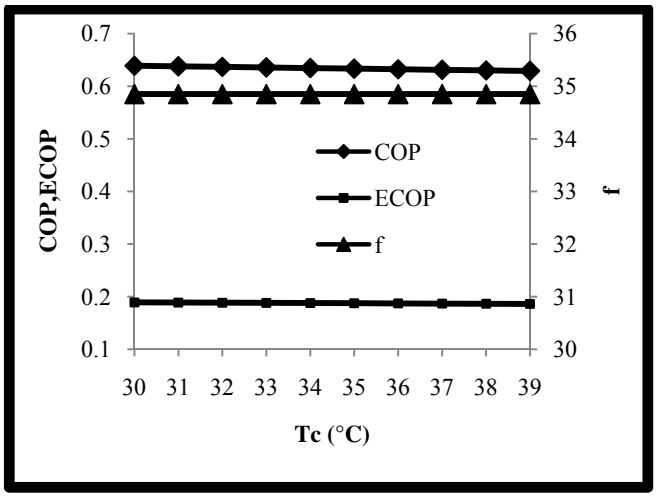


Fig. 3: Effect of condenser temperature variation

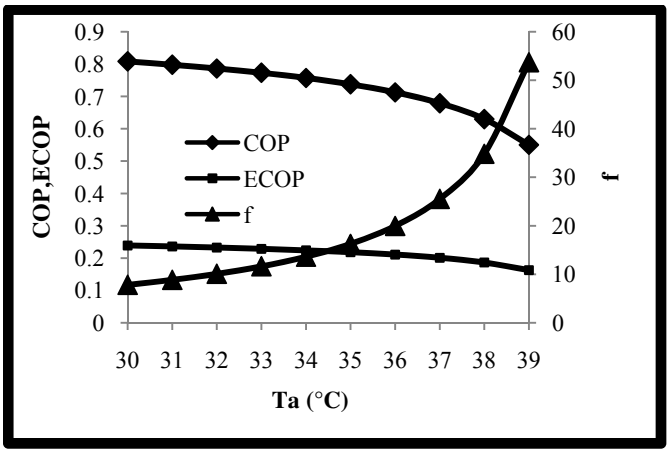


Fig. 4: Effect of absorber temperature variation

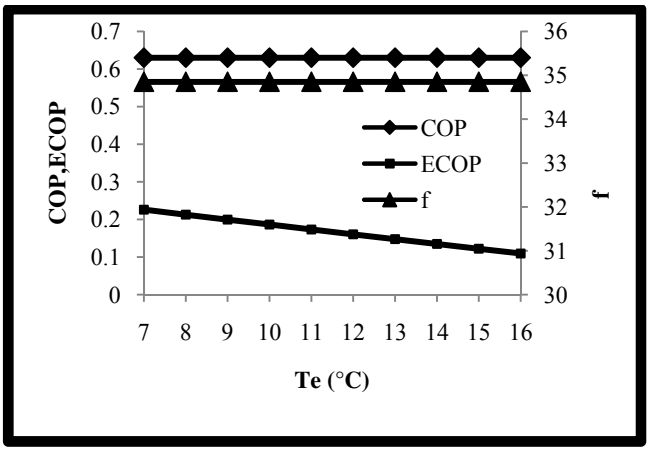


Fig. 5: Effect of evaporator temperature variation

The ECOP decreases with increase in the temperature of all these components. The results of this analysis also indicate that the designing of the absorption refrigeration system should be on the basis of thermodynamic point of view. In the residential and industrial locations, installation of the system is of major concern. The components of absorption refrigeration cycle like absorber, condenser, evaporator and generator are significant in the real world applications.

For commercial applications the above study can be considered a better tool and also for the cooling load and other parameters. From this discussion we see the essential and avoidable parameters of the cycle.

5. CONCLUSION

This study deals with the LiBr-H_2O absorption refrigeration system for cooling and air-conditioning. The effect of temperature variation has been observed on the performance with the help of first law and second law equations of the cycle.

From this useful discussion, it can be concluded that

- As the temperature varies in absorber, the circulation ratio increases while COP and ECOP decrease.
- As the temperature varies in evaporator, the circulation ratio remains same and ECOP decrease while COP increases.
- As the temperature varies in condenser, the circulation ratio remains the same, COP and ECOP decease.
- As the temperature varies in the generator, the circulation ratio remains, COP and ECOP decrease.

REFERENCES

[1] He Yijian, ChenGuangming, “Experimental study on an absorption refrigeration system at low temperatures”, International Journal of Thermal Sciences,46, 2007, pp. 294–299.

[2] Vereda C., Ventas R., Lecuona A., Venegas M., “Study of an ejector-absorption refrigeration cycle with an adaptable ejector nozzle for different working conditions”, Applied Energy, 97, 2012, pp. 305–312.

[3] Stitou D., Spinnera B., Satzger P., Ziegler F., “Development and comparison of advanced cascading cycles coupling a solid/gas thermochemical process and aliquid/gas absorption process”, Applied Thermal Engineering, 20, 2000,pp.1237-1269.

[4] Tozer Robert, James Ron W, “Heat powered refrigeration cycles”, Applied Thermal Engineering, 18, 1998, pp. 731-743.

Signal, Image and Video Processing
https://doi.org/10.1007/s11760-023-02774-x

ORIGINAL PAPER

Uncovering visual attention-based multi-level tampering traces for face forgery detection

Ankit Yadav[1] · **Dhruv Gupta**[1] · **Dinesh Kumar Vishwakarma**[1]

Received: 2 August 2023 / Revised: 26 August 2023 / Accepted: 4 September 2023
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

Abstract
With the rise of realistic face forgery techniques, the threat of identity fraud is more significant than ever. Several research works have focused on detecting such forgeries, but they extract forgery clues as a preprocessing step to the feature extraction phase of deep neural networks. A novel *DenseTrace-Net* architecture is designed in this manuscript to extract more comprehensive and refined face tampering traces locally and globally. Specifically, DenseTrace-Net extracts attentional multi-level tampering traces from facial images. A novel 'Local Attentional Tamper Trace Extractor' (LATTE) module extracts face tampering traces locally at the block level. A novel 'Global Attentional Tamper Trace Extractor' (GATTE) module aggregates multi-scale tampering traces globally. The LATTE and GATTE modules use visual depth attention to enhance their feature representation capability. Additionally, the proposed DenseTrace-Net is computationally lightweight with just 1.378 million parameters. DenseTrace-Net is evaluated on three benchmark datasets, the FF + + , CelebDF and DFDC datasets, achieving AUC scores of 0.9784, 0.9843 and 0.9916, respectively. These excellent scores allow the DenseTrace-Net to outperform the existing state-of-the-art face forgery detection methods comfortably.

Keywords DeepFake · Face tampering · Face manipulation · Face forgery · Detection · Classification

1 Introduction

Several realistic face tampering methods have been developed in recent years. Some examples of such facial tampering methodologies are DeepFakes [1], Face2Face [2], DeepFaceLab [3], etc. DeepFakes utilize artificial neural networks to modify facial features. FaceSwap implements computer graphics to manipulate or swap entire faces. DeepFaceLab is an extensive online repository of realistic facial tampering utilities, including modification of lip movements, replacement of faces, age reduction, etc.

Such realistic manipulation capabilities pose a serious threat to individuals or organizations. For example, identity-changing manipulations can be used to plant fake evidence against an innocent person, endangering his life [4–6].

Recent developments in computer vision have proposed several contributions toward face tampering detection. In the last decade of computer vision research, convolutional neural networks (CNNs) have been the primary driving force in deep learning-based face tampering detection. Several novel contributions have utilized a variety of CNN architectures to extract discriminative features from the spatial domain [7–10], textural information [11–13], optical flow maps [14], frequency features [15], etc., to detect facial tampering.

However, the recently proposed self-attention-based vision transformers (ViTs) [16] have demonstrated superior classification capabilities than CNNs on the ImageNet classification challenge. The main advantage of ViTs is their capacity to capture long-term dependencies that regular CNNs typically overlook.

Despite the growing popularity of ViTs, they have several limitations. ViTs are data-hungry and require complex training paradigms on huge datasets to show any noticeable improvements over CNNs [17]. ViTs perform worse when

✉ Dinesh Kumar Vishwakarma
dvishwakarma@gmail.com

Ankit Yadav
ankit4607@gmail.com

Dhruv Gupta
dhruvg029@gmail.com

[1] Department of Information Technology, Delhi Technological University, Bawana Road, Delhi 110042, India

Published online: 04 November 2023

trained for small-scale downstream tasks, unlike CNNs [18]. ViTs are usually massive networks with nearly 100 million trainable parameters, while CNNs can be tweaked to be lightweight.

Hence, going against the current trends of ViTs, a novel CNN-based approach is proposed in this paper. The proposed DenseTrace-Net model is computationally lightweight. Unlike [19], it captures tampering traces both locally and globally. To incorporate the long-term dependency-capturing capability of ViTs, it contains a novel multi-scale tamper trace extractor that captures long-term contextual information. It also employs depth attention to highlight important feature channels.

The significant contributions of this manuscript are:

1. A novel DeepFake detection model called *DenseTrace-Net* based on dual-order attentional tamper trace extraction is proposed.
2. DenseTrace-Net contains a novel 'Local Attentional Tamper Trace Extractor' (LATTE) module that suppresses the semantic content of facial images and highlights tampering traces locally at the block level. It also contains a novel 'Global Attentional Tamper Trace Extractor' (GATTE) module that aggregates multi-scale tamper residuals globally across multiple blocks to capture the long-term contextual relationship of tampering traces. The LATTE and GATTE modules use depth attention to boost the model performance.
3. DenseTrace-Net is computationally lightweight by design, with only 1.378 million trainable parameters.
4. DenseTrace-Net achieves excellent AUC scores of 0.9784 (average), 0.9843 and 0.9916 on the FF + + , CelebDF and DFDC datasets, respectively, beating the state of the art for DeepFake detection.

In this manuscript, Sect. 2 presents the related work regarding face tampering methods. Section 3 offers the details of the proposed architecture DenseTrace-Net. Section 4 explains the experimental setup, specifying the datasets used, classification metrics, hardware, choice of face extractor, preprocessing operations and hyperparameter specifications. Section 5 presents the performance of DenseTrace-Net on publicly available datasets FF + + , CelebDF and DFDC. Section 6 includes a comprehensive ablation investigation to validate the validity of DenseTrace-Net's distinct components. Section 7 specifies the conclusion and future work possible. Section 8 contains the references cited in this work.

2 Related work

This section presents an extensive overview of existing state-of-the-art methods for face tampering detection. Traditional face tampering detection methods developed handcrafted features such as [11]. However, the main disadvantage of handcrafted feature extraction approaches is their reliance on prior information and weak generalization ability.

Amerini et al. [20] use optical flow maps to CNNs to classify them as real or fake. Caldelli et al. [21] improve on this method by employing bidirectional optical flow maps that provide additional discriminative input information. Yang et al. [13] detect face manipulation by textual feature extraction. The authors propose a novel convolution equivalent operator to extract textual-based differences. Detecting facial manipulation from videos requires processing all frames, which is computationally expensive. Hu et al. [22] reduce this computational complexity by only considering video I-frames for forgery detection in facial images. Shang et al. [7] propose feature learning at multiple scales, i.e., patch and pixel levels. Xu et al. [8] train CNN on a set of frames aggregating multi-frame features from videos.

With the development of attention-based methods, several contributions have been used to boost model performance for face tampering detection. Dang et al. [23] propose a novel manipulation appearance model based on soft attention for tampering detection. Choi et al. [24] suggest using certainty-aware attention maps to emphasize critical video frames important for final prediction. Lu et al. [25] enhance the popular inception module by using attention to compute channel and spatiotemporal attention.

Some methods, like [19], intend to remove tampering traces and conceal face picture semantic content. However, the tampering trace extraction is designed as a preprocessing step that does not capture short-term inconsistencies and long-term contextual relationships of tampering residuals. The proposed DenseTrace-Net model is designed to extract tampering traces locally and globally.

3 Proposed architecture

This section provides a detailed description of the proposed model DenseTrace-Net.

Figure 1 shows the design of the proposed DenseTrace-Net model.

3.1 Feature reuse and improved information flow

Computer vision has witnessed a variety of deep learning architectures that have performed well in image classification. Traditional feed-forward networks are built on the generic idea that a network layer $\varnothing_L(.)$ takes input from the

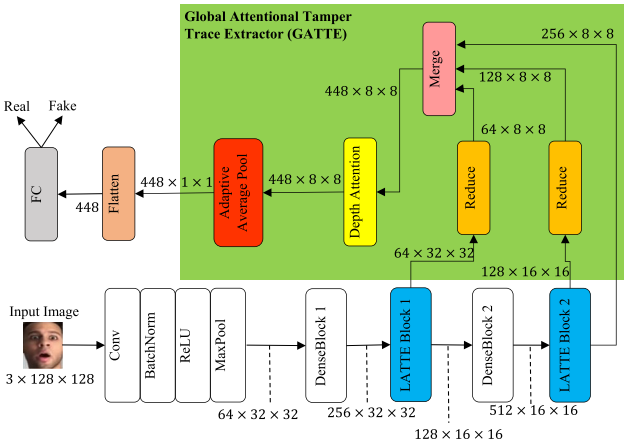


Fig. 1 Structure of DenseTrace-Net. The local attentional tamper trace extractor or LATTE blocks (blue) extract tamper traces at block level. The global attentional tamper trace extractor or GATTE block (green) aggregates multi-scale tamper traces globally. The predefined layers from baseline architecture [3] is shown in 'white' color while newly added novel blocks have been populated with multiple colors

previous layer $\mathfrak{y}_{L-1}$ and produces the next set of features $\mathfrak{y}_L$ as shown in Eq. (1):

$$\mathfrak{y}_L = \emptyset_L(\mathfrak{y}_{L-1}) \tag{1}$$

However, as networks have grown progressively deep, the vanishing gradient problem has restricted model performances. Famously, the ResNet [26] included the additive identity skip connection block to solve this problem, as shown in Eq. (2):

$$\mathfrak{y}_L = \emptyset_L(\mathfrak{y}_{L-1}) + \mathfrak{y}_{L-1} \tag{2}$$

However, ResNet suffers from several flaws. For example, not all layers contribute equally to the overall prediction, and some can be dropped from the architecture [27]. Another disadvantage is a large number of trainable parameters from multiple convolutional kernels in each layer, leading to feature maps with ever-increasing depth and increasing computational complexity. Also, the additive nature of skip connection via the summation operation interferes with efficient information flow.

The architecture from [28] is used as the baseline architecture in this experiment to overcome these drawbacks. It encourages feature reuse by letting a layer access feature maps from all previous layers, as shown in Eq. (3):

$$\mathfrak{y}_L = \emptyset_L\left(\left[\mathfrak{y}_0, \mathfrak{y}_1, \mathfrak{y}_2, \mathfrak{y}_3 \ldots\ldots \mathfrak{y}_{L-1}\right]\right) \tag{3}$$

This benefits the model, as each layer can access the learnings of all previous layers instead of just one. While a traditional deep network with $\mathbb{L}$ layers contains $\mathbb{L}$ connections, this architecture includes $\mathbb{L}(\mathbb{L}+1)/2$ connections for the same number of layers. The overall computational cost of this architecture is reduced counterintuitively by restricting the number of feature channels produced by a factor 'k' termed the growth rate. Restricting the growth parameter value helps limit the sudden growing depth of feature maps in deep models.

There are several distinct advantages to choosing this baseline. It improves the feature reusability since each layer has access to features of all preceding layers. The information flow is maximized since features are concatenated instead of the summation approach. Additionally, the dense connections help to regularize and prevent overfitting on the training data [28]. The growth rate parameter 'k' ensures reduced computation. The 121-layer variant of [28] is used in this experiment.

3.2 Dual-order attentional tamper trace extractor

It makes sense to suppress the semantic content of multimedia data to uncover manipulation traces. Existing methods, such as [19], implement tamper trace extraction as a preprocessing step. However, it is inferred that a better approach could be extracting tampering traces locally and globally.

Conceptually, two significant enhancements have been proposed on the baseline of [28] that constitutes the proposed DenseTrace-Net architecture. First, the existing transition blocks, initially meant for spatial reduction, are replaced by the novel 'Local Attentional Tamper Trace Extractor' (LATTE) blocks to extract tampering traces locally. Next, a novel 'Global Attentional Tamper Trace Extractor' (GATTE) module aggregates multi-scale tampering traces globally across dense blocks of baseline architecture. These components are explained below in greater detail.

3.2.1 Local attentional tamper trace extractor (LATTE)

The Local Attentional Tamper Trace Extractor (LATTE) module has been designed to extract discriminative tampering traces locally at the block level. Additionally, a LATTE module also performs dimensionality reduction of extracted features. Figure 2 shows the detailed structure of the proposed LATTE module.

A dense block $\mathcal{D}(.)$ takes input features $g_{\text{dense, in}}$ and produces output features $g_{\text{dense, out}}$, as shown in Eq. (4).

$$g_{\text{dense, out}} = \mathcal{D}\left(g_{\text{dense, in}}\right) \tag{4}$$

The LATTE module $\mathcal{L}(.)$ takes features from a dense block and extracts tampering traces at the block level. $\mathcal{L}(.)$ is a composite function defined by Eq. (5).

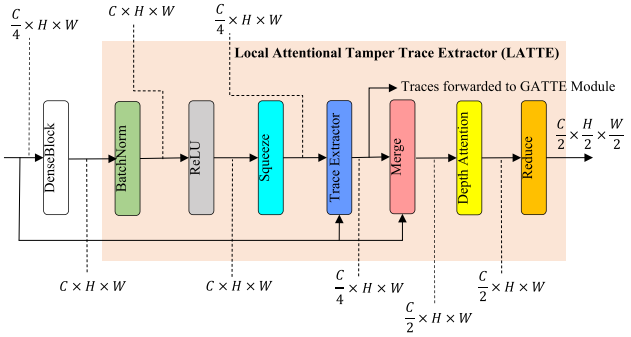


Fig. 2 Structure of local attentional tamper trace extractor (LATTE)

$$\mathcal{L}\left(g_{\text{dense, in}}, g_{\text{dense, out}}\right) = \mathcal{R}e\left(\mathcal{DA}\left(\mathcal{M}\left(\mathcal{TE}\left(\mathcal{S}q\left(\Psi\left(g_{\text{dense, out}}\right)\right), g_{\text{dense, in}}\right), g_{\text{dense, in}}\right)\right)\right) \quad (5)$$

Here, $\Psi(.)$ refers to a sequential operation of batch normalization and ReLU activation.$\mathcal{SE}(.)$ function squeezes features to a certain depth via 1×1 convolution operation. $\mathcal{TE}$ (.) computes the tampering residual via Eq. (6), which is also forwarded to the GATTE module for multi-scale tampering trace aggregation.

$$g_{\text{tampering trace}} = \mathcal{TE}\left(f_{\text{squeeze}}, g_{\text{dense, in}}\right) = g_{\text{dense, in}} - f_{\text{squeeze}} \quad (6)$$

Here, $f_{squeeze}$ refers to output features from $\mathcal{S}q(.)$.

$\mathcal{M}(.)$ performs channel-wise feature concatenation. $\mathcal{DA}$ (.) employs depth attention mechanism along feature channels, as explained in a later subsection. $\mathcal{R}e(.)$ reduces the spatial dimensions of features. Several alternate designs for LATTE were tested, and their results are discussed in the experimentation section.

3.2.2 Global attentional tamper trace extractor (GATTE)

While the LATTE module extracts local block-level tampering traces, the Global Attentional Tampering Trace Extractor (GATTE) aggregates multi-scale tampering traces. This helps capture the long-term contextual relationship of tampering residuals from multiple feature scales. The GATTE module $\mathcal{G}(.)$ can be expressed as shown in Eq. (7):

$$\mathcal{G}\left(_{i=0}{}^{2}\left[\text{trace}_i\right]\right) = \mathcal{AP}_{\text{adaptive}}\left(\mathcal{DA}\left(\mathcal{M}\left(_{i=0}{}^{1}\mathcal{R}e\left(\text{trace}_i\right), \text{trace}_2\right)\right)\right) \quad (7)$$

$\mathcal{G}(.)$ takes tampering residuals from local LATTE blocks represented by (trace_i) and aggregates multi-scale traces. $\mathcal{R}e(.)$, $\mathcal{M}(.)$ and $\mathcal{DA}(.)$ represent the same operations as described for LATTE. $\mathcal{AP}_{\text{adaptive}}(.)$ represents adaptive average pooling operation. The green block in Fig. 1 demonstrates the structure of the GATTE module.

3.2.3 Depth attention

With the recent rise of attention mechanisms, several methods have targeted important highlighting features along the depth [29–31]. These methods either focus only on highlighting important feature channels [29] or they try to exploit positional information such as in [32, 33].

This work uses the recently proposed depth attention from [31] to extract crucial depth information along discriminative feature channels. The main benefit of this depth attention lies in its incorporation of spatial positional embeddings into the channel attention mechanism. Specifically, instead of using the commonly used $2\mathcal{D}$ average pooling, it extracts important channel information from two $1\mathcal{D}$ feature vectors along the height and width to generate direction-aware and position-sensitive attention maps.

3.3 Lightweight architecture

The 121-layered variant of the baseline architecture contains four dense blocks and has 6.953 million trainable parameters. It is experimentally verified that the novel local LATTE modules and the global GATTE module provide complementary feature learning capabilities that can be leveraged to reduce the model size. The last two dense blocks that unnecessarily increase the number of trainable parameters without any significant increase in performance are removed. Hence, DenseTrace-Net contains only 1.378 million trainable parameters, making it extremely lightweight. Table 7 demonstrates this ablation study on the proposed model by removing zero, one, two and three dense blocks. The best results were achieved with two dense blocks removed, making the proposed model extremely lightweight without compromising the detection performance.

4 Experimental setup

This section explains the experimental setup followed for evaluating the proposed architecture.

4.1 Datasets

This section presents the publicly available face tampering datasets used to test the proposed model.

Table 1 Classification metrics for evaluating face tampering detection models

Metric	Formula	Range
Accuracy (ACC)	$\frac{TP+TN}{TP+TN+FP+FN}$	[0,1]
Precision (P)	$\frac{TP}{TP+FP}$	[0,1]
Recall (R)	$\frac{TP}{TP+FN}$	[0,1]
F1 score (F1)	$2 * \frac{Precision*Recall}{Precision+Recall}$	[0,1]
Area under curve (AUC)	–	[0,1]
Mathews correlation coefficient (MCC)	$\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	[− 1,1]

4.1.1 CelebDF

The CelebDF [34] dataset contains 5639 DeepFake videos and 590 actual videos extracted from YouTube, matching the celebrities of different genders, ages and races, totaling more than 2.3 million frames. These videos demonstrate various elements, including different face sizes, lighting setups and backgrounds.

4.1.2 DeepFake detection challenge

The DFDC or DeepFake Detection Challenge dataset [35] comprises 5214 videos featuring two unspecified facial modification algorithms. These methods are used to generate fake videos, which produce different quality swaps. Numerous actors with various genders, skin tones and ages were used to shoot the videos for the DFDC dataset. These videos feature a variety of lighting, head positions and visual backgrounds. Moreover, other features like gender, skin tone, age, etc. led actors to shoot videos against arbitrary backgrounds, adding visual variability.

4.1.3 FaceForensics + +

FaceForensics + + (FF + +) [36] is a forensics dataset having 1000 pristine video clips collected from the internet. Five automated face manipulation methods generate tampered videos, namely Face2Face, DeepFakes, FaceShifter, FaceSwap and NeuralTextures. Face2Face and FaceSwap manipulations are based on computer graphics algorithms, while NeuralTextures and DeepFakes are learning-based approaches. FaceShifter employs a two-stay face-swapping mechanism.

4.2 Classification metrics

Table 1 presents the list of classification metrics used in this experiment.

4.3 Hardware

All tests are run in parallel on two NVIDIA RTX A5000 GPUs with 128GB of system memory.

4.4 Preprocessing—resizing, normalization and data augmentation

This section presents the preprocessing operations followed in this experiment. All facial images are resized to 3 × 128 × 128. Random horizontal and vertical flipping is used to enhance the face images. The flipping probability is set to 0.5. The pixel values have been normalized to a range of [0, 1].

4.5 Hyperparameter specifications

This section presents the hyperparameter settings used to train the proposed model. Specifically, all experiments are run for 30 epochs. The batch size is set to 24. The learning rate (LR) is initialized at 0.0001, and a linear decay is implemented that reduces the learning rate after every epoch.

5 Results and discussion

This section presents the results obtained by DenseTrace-Net on publicly available DeepFake datasets and how they compare against results achieved by other state-of-the-art face tampering methods.

5.1 DenseTrace-Net performance on benchmark datasets

This section demonstrates the performance of DenseTrace-Net on FF + + , CelebDF and DFDC datasets. The accuracy (ACC), precision (P), recall (R), f1 (F1), AUC and MCC scores are shown in Table 2.

Figure 3 visually represents DenseTrace-Net performance on the FF + + , CelebDF and DFDC datasets. It achieves more than 0.95 accuracy and close to 0.9 0 MCC scores for all cases

Table 2 Results of DenseTrace-Net on CelebDF, FF + + and DFDC datasets

Datasets	ACC	P	R	F1	AUC	MCC
DeepFakes (FF + +)	0.9685	0.9688	0.9647	0.9667	0.9917	0.9369
Face2Face (FF + +)	0.9683	0.9629	0.9680	0.9655	0.9917	0.9362
FaceShifter (FF + +)	0.9615	0.9429	0.9845	0.9632	0.9912	0.9240
FaceSwap (FF + +)	0.9545	0.9672	0.9418	0.9544	0.9884	0.9094
NeuralTextures (FF + +)	0.8477	0.8299	0.8810	0.8547	0.9291	0.6964
Celeb DF	0.9745	0.9740	0.9972	0.9855	0.9843	0.8858
DFDC	0.9622	0.9786	0.9737	0.9761	0.9916	0.8858

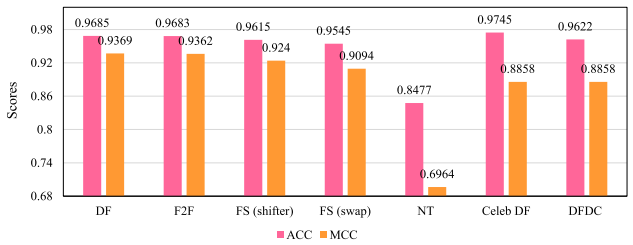


Fig. 3 MCC and accuracy scores of DenseTrace-Net on FF + + , CelebDF and DFDC datasets

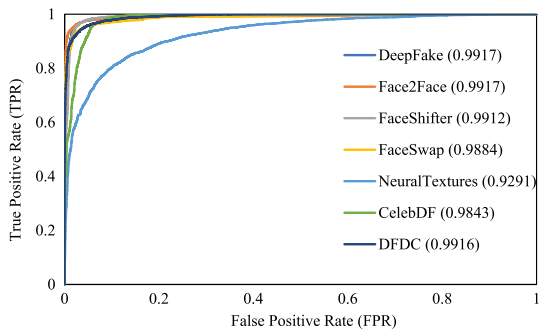


Fig. 4 AUC-ROC curve of DenseTrace-Net on FF + + , CelebDF and DFDC datasets

except NeuralTextures (NT) tampering, which is challenging. The high accuracy scores indicate that DenseTrace-Net can identify facially tampered images easily.

The MCC scores achieved by DenseTrace-Net prove that the model learns to identify original and tampered facial photographs. It is noteworthy that MCC lies in the range of $[-1, 1]$ and DenseTrace-Net reaches close to 0.9 0 MCC score for all cases except NT tampering.

AUC score measures a model's confidence in making a prediction. Figure 4 shows that DenseTrace-Net delivers potent AUC scores for all dataset variants. Most scores lie close to 0.9 9, signifying high confidence in the predictions.

5.2 Performance comparison against the state-of-the-art methods

This section compares the performance of DenseTrace-Net to existing state-of-the-art approaches for detecting face tampering.

Tables 3, 4 and 5 compare the DenseTrace-Net scores against existing methods on the FF + + , DFDC and CelebDF datasets, respectively, showing the accuracy (ACC) and area under curve (AUC) values. For the FF + + dataset, some existing methods specify manipulation category-specific scores, while others have mentioned overall aggregate scores across all categories. Hence, the unavailable scores from existing works are mentioned with a '–.' A '*' represents the existing models implemented and trained from scratch. The scores mentioned in bold indicate the best values. Similarly, for the DFDC and CelebDF datasets, '–' means unavailable scores, '*' represents implemented and trained models and the best scores are marked in bold.

Table 2 proves that DenseTrace-Net beats all state-of-the-art face tampering methods, achieving the highest accuracy and AUC values on the FF + + dataset across all manipulation types. DenseTrace-Net achieves excellent accuracy scores close to 0.9 5 for all manipulations except NT manipulation. Similarly, DenseTrace-Net scores close to 0.9 9 AUC scores except for NT, beating all the state of the art.

Table 4 shows the DenseTrace-Net performance comparison against the state-of-the-art methods on the DFDC dataset. Once again, the proposed model convincingly beats all state-of-the-art accuracy and AUC scores, proving the superiority of DenseTrace-Net.

Table 5 shows the performance comparison on the CelebDF dataset. Consistent with other dataset comparison results, DenseTrace-Net easily outperforms all current methods, making it the clear favorite.

5.3 Generalization study of DenseTrace-Net

A generalization study helps prove that a given model can identify unknown manipulations not witnessed during training. This is crucial since the types of facial manipulations

Table 3 Comparison of DenseTrace-Net against state-of-the-art methods on the FF + + dataset

Methods	Year	DeepFake		Face2Face		FaceShifter		FaceSwap		NeuralTexture		Average	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Yang et al. [37]	2023	–	–	–	–	–	–	–	–	–	–	–	0.8709
Guo et al. [38]	2023	–	–	–	–	–	–	–	–	–	–	–	0.9755
Xu et al. [39]	2023	–	–	–	–	–	–	–	–	–	–	–	0.9034
Lin et al. [40]	2023	–	–	–	–	–	–	–	–	–	–	0.9074	0.9486
Yang et al. [41]	2022	–	–	–	–	–	–	–	–	–	–	–	0.7888
Nirkin et al. [42]	2022	0.9450	–	0.8030	–	–	–	0.8450	–	0.7400	–	–	–
*AMTENnet [19]	2021	0.9285	0.9801	0.9204	0.9756	0.9273	0.9837	0.8995	0.9665	0.7728	0.8455	0.8897	0.9502
*Coordinate attention [43]	2021	0.8675	0.9430	0.5932	0.6295	0.7536	0.8276	0.7562	0.8441	0.6162	0.6514	0.7173	0.7791
Shang et al. [7]	2021	0.9563	–	0.9015	–	–	–	0.9493	–	0.8001	–	–	–
Liu et al. [44]	2021	0.9348	–	0.8602	–	–	–	0.9226	–	0.7678	–	0.8713	–
Hu et al. [45]	2021	0.9464	0.9800	0.8648	0.9400	–	–	0.8527	0.9400	0.8005	0.9000	–	–
Chen et al. [46]	2021	–	0.9595	–	–	–	–	–	0.9787	–	–	–	–
*CDCN + + [47]	2020	0.9074	0.9727	0.8715	0.9458	0.9081	0.9716	0.9086	0.9601	0.7639	0.8422	0.8719	0.9384
Baek et al. [48]	2020	0.7180	–	0.6860	–	–	–	0.6310	–	0.7070	–	–	–
Zi et al. [49]s	2020	0.9210	–	0.8390	–	–	–	0.9250	–	0.7820	–	–	–
Amerini et al. [20]	2019	–	–	0.8161	–	–	–	–	–	–	–	–	–
Rossler et al. [28]	2019	0.7450	–	0.7590	–	–	–	0.7090	–	0.7330	–	–	–
Afchar et al. [50]	2018	0.8727	–	0.5620	–	–	–	0.6117	–	0.4067	–	0.6132	–
DenseTrace-Net (proposed)	–	**0.9685**	**0.9917**	**0.9683**	**0.9917**	**0.9615**	**0.9912**	**0.9545**	**0.9884**	**0.8477**	**0.9291**	**0.9401**	**0.9784**

Table 4 Comparison of DenseTrace-Net against state-of-the-art methods on the DFDC dataset

Methods	Year	ACC	AUC
Mohiuddin et al. [51]	2023	0.7534	0.8567
Deng et al. [52]	2023	0.9216	0.9784
Guo et al. [38]	2023	–	0.9597
Asha et al. [53]	2023	0.8700	0.9100
Ke et al. [54]	2023	0.9108	–
Lin et al. [40]	2023	–	0.8847
Guo et al. [55]	2023	–	0.9827
Zhao et al. [56]	2023	0.9210	–
Yu et al. [57]	2023	0.9593	0.9896
Yang et al. [58]	2023	–	0.9911
Heo et al. [10]	2023	–	0.9780
Xu et al. [39]	2023	–	0.8037
Ganguly et al. [59]	2022	0.7321	0.8632
Ganguly et al. [60]	2022	0.7520	0.8359
Nadimpalli et al. [61]	2022	0.9100	0.8600
* Coordinate attention [43]	2021	0.8917	0.8830
* AMTENnet [19]	2021	0.9139	0.9402
Xu et al. [8]	2021	0.8453	0.9347
Shang et al. [7]	2021	–	0.9778
Li et al. [62]	2021	–	0.7600
Li et al. [63]	2020	–	0.8090
Luo et al. [64]	2021	0.8841	0.9496
Qi et al. [65]	2020	0.6410	–
Li et al. [66]	2020	0.8511	–
* CDCN + + [47]	2020	0.9174	0.9492
Mittal et al. [67]	2020	–	0.8440
Montserrat et al. [68]	2020	0.9188	–
Chugh et al. [69]	2020	–	0.9160
DenseTrace-Net (proposed)	–	**0.9622**	**0.9916**

available at training time are limited. Hence, a model that can effectively identify unseen facial manipulations has strong application in real-world scenarios. This section gives a study of the DenseTrace-Net model's generalization ability.

To evaluate DenseTrace-Net's ability to identify unknown face manipulations, the FF + + dataset has been used. There are five categories of face manipulations in the FF + + dataset: DeepFake Face2Face, FaceShifter, FaceSwap and NeuralTextures. In each round of training and testing, four categories of FF + + have been used to train the model; the fifth category is treated as the unseen test set. Training and testing in this cross-dataset manner ensure that the testing samples encountered by DenseTrace-Net were unavailable during the training phase. For example, in the first case (first row of Table 6), DenseTrace-Net is trained on a combined set of Face2Face, FaceShifter, FaceSwap and NeuralTextures categories. Then, the trained model was tested on the DeepFake category, which was not included during training.

The performance results in this cross-dataset evaluation are presented in Table 6. As expected, the unseen nature of the test set degrades the performance of DenseTrace-Net slightly compared to the same-dataset evaluation scores.

Figures 5 and 6 present a visual comparison of the DenseTrace-Net's accuracy and AUC performance on the same- and cross-dataset evaluation. It is clear from the presented visuals that the performance of DenseTrace-Net is reduced due to the unseen manipulations existing in the test set. However, DenseTrace-Net still achieves competent performance despite the unknown nature of manipulation in the test set.

In terms of the accuracy scores (Fig. 5), the drop in performance is less than 10% for the DeepFake (DF), Face2Face (F2F) and FaceShifter (FSh) manipulation categories. However, the FaceSwap (FSw) and NeuralTextures (NT) manipulation categories suffer a more severe drop in performance of 12% and 14%, respectively.

A similar story can be observed in the AUC score comparison of the same- and cross-dataset evaluation (Fig. 6). DF, F2F and FSh suffer minimal drops in performance. At the same time, FSw and NT have a relatively higher degradation of scores.

One possible explanation could be that the FSw and NT manipulation categories are more realistic and challenging than others, making it harder for the model to differentiate them from a non-manipulated facial image. A similar pattern arises in the same-dataset evaluation presented in Table 2, where the model scores accuracies higher than 0.96 for all categories except the FSw and NT. This implies that the FSw and NT examples are relatively more realistic and more challenging for the model to detect when compared to other manipulation categories. This performance degrades further in the generalization study due to the unseen nature of the test set.

5.4 Model size comparison against popular architectures

Figure 7 compares the size of DenseTrace-Net (blue) against famous deep learning architectures (green).

Specifically, DenseTrace-Net is extremely lightweight with only 1.378 million parameters. In comparison, other popular architectures like ResNet (25.438 million parameters), EfficientNet (66.347 million parameters), DenseNet (28.681 million parameters) and Vision Transformer-ViT_B_16 (86.566 million parameters) have a large number of trainable parameters. The lightweight nature of DenseTrace-Net makes it extremely easy to train, yet it outperforms all state-of-the-art methods.

Table 5 Comparison of DenseTrace-Net against state-of-the-art methods on the CelebDF dataset

Methods	Year	ACC	AUC
Guo et al. [38]	2023	–	0.6743
Ke et al. [54]	2023	0.8892	
Guo et al. [55]	2023	–	0.6950
Xu et al. [39]	2023	–	0.9332
Yang et al. [37]	2023	–	0.9108
Asha et al. [53]	2023	0.8800	0.8400
Li et al. [70]	2023	0.7169	0.7659
Yang et al. [41]	2022	–	0.8561
Nadimpalli et al. [61]	2022	0.6200	0.6700
Nirkin et al. [42]	2022	–	0.6600
Li et al. [62]	2021	–	0.7600
Hu et al. [63]	2021	–	0.7884
Trinh et al. [65]	2021	–	0.7176
Chen et al. [9]	2021	0.9575	–
Luo et al. [64]	2021	0.7435	0.9234
Hu et al. [45]	2021	0.8074	0.8700
Chen et al. [46]	2021	–	0.8765
*CoordinateAttention [43]	2021	0.8904	0.8026
*AMTENnet [19]	2021	0.9254	0.8804
*CDCN + + [47]	2020	0.9180	0.8982
Qian et al. [71]	2020	0.8706	0.8148
Dang et al. [25]	2020	–	0.7120
Li et al. [63]	2020	–	0.7476
Choi et al. [24]	2020	0.9200	0.9400
Afchar et al. [50]	2018	0.6750	0.6681
DenseTrace-Net (proposed)	–	**0.9745**	**0.9843**

Table 6 Generalization study of DenseTrace-Net on the FF + + dataset

Train datasets	Test datasets	ACC	P	R	F1	AUC	MCC
F2F + FSh + FSw + NT	DF	0.9112	0.9613	0.8529	0.9038	0.9725	0.8270
DF + FSh + FSw + NT	F2F	0.9014	0.9378	0.8575	0.8959	0.9677	0.8055
DF + F2F + FSw + NT	FSh	0.8882	0.9738	0.7878	0.8710	0.9706	0.7879
DF + F2F + FSh + NT	FSw	0.8309	0.8792	0.7753	0.8240	0.9170	0.6674
DF + F2F + FSh + FSw	NT	0.7006	0.6658	0.9515	0.7834	0.8555	0.4076

There are five categories of manipulations—DeepFake (DF), Face2Face (F2F), FaceShifter (FSh), FaceSwap (FSw) and NeuralTextures (NT). In each case, the model is trained on four manipulation categories and tested on the unseen fifth category

6 Ablation study

This section includes an in-depth ablation investigation to validate the contributions of DenseTrace-Net's different components. All ablation experiments are carried out using the DF (FF + +) dataset.

6.1 ResNet versus baseline model comparison

This section presents the relative performance of the two architectures on the DF (FF + +) dataset.

Figure 8 presents the performance comparison of the two architectures. It can be seen that the baseline model is superior to the more popular ResNet architecture. Hence, this model has been taken as the backbone of the proposed DenseTrace-Net.

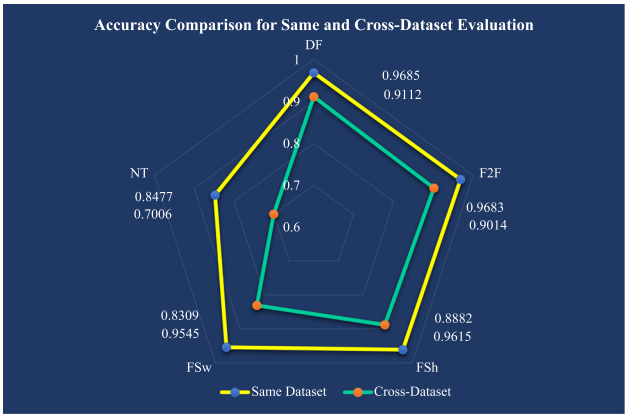


Fig. 5 Accuracy score comparison of DenseTrace-Net for same- and cross-dataset evaluation

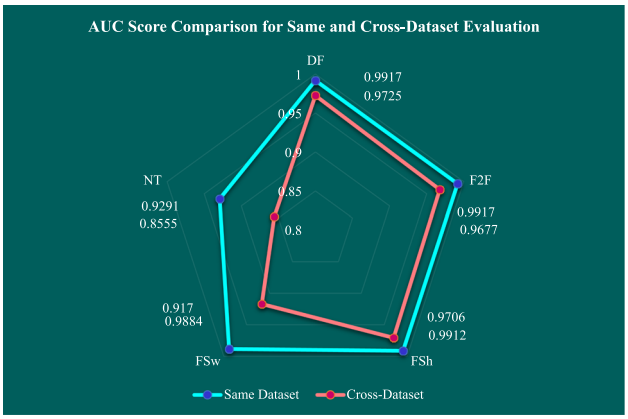


Fig. 6 AUC score comparison of DenseTrace-Net for same- and cross-dataset evaluation

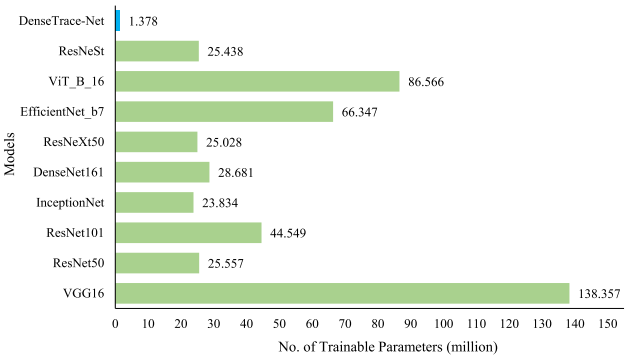


Fig. 7 Size comparison of DenseTrace-Net (blue) against popular deep architectures (green)

6.2 Weight initialization strategy

The first experiment of the ablation study explores the impact of the weight initialization strategy. Specifically, the 121-layer variant of [28] is trained and tested with two different weight initialization strategies. In the first case, the weights

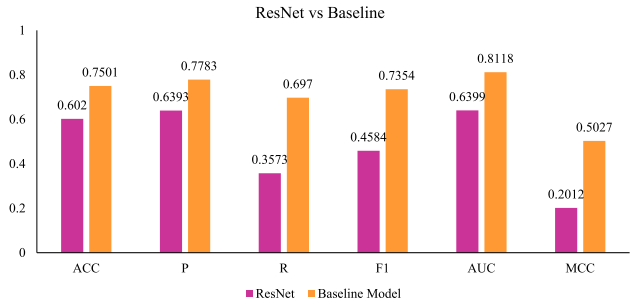


Fig. 8 Comparison of baseline model against the ResNet architecture

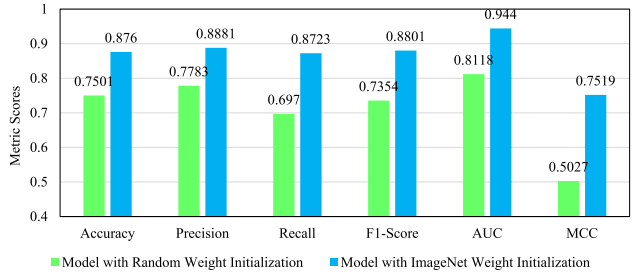


Fig. 9 Impact of random weight initialization versus ImageNet weight initialization

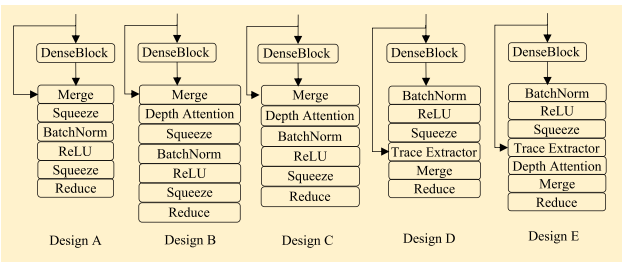


Fig. 10 Alternate designs tested for the local attentional tamper trace extractor (LATTE)

are randomly initialized. The model weights are initialized with pre-trained ImageNet weights for the second case. No layers are frozen for the second case, and the complete network is trainable. Figure 9 shows that initializing the model with pre-trained ImageNet weights achieves significantly better scores when compared to random weight initialization.

6.3 Design choice for local attentional tamper trace extractor (LATTE) module

This section explores several alternate design choices (Design A to E) tested for the Local Attentional Tamper Trace Extractor (LATTE) module, as presented in Fig. 10.

Figure 11 shows the performance of these designs. Design E scores the highest accuracy, MCC and F1 score and is finalized as the LATTE module design for DenseTrace-Net.

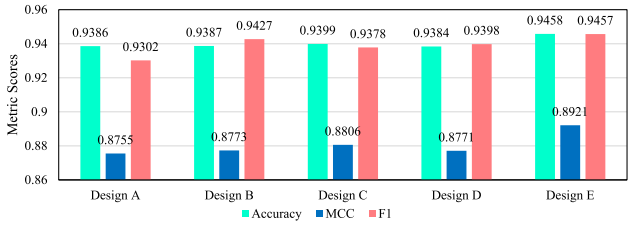


Fig. 11 Performance of alternate designs for the LATTE module. Design E scores the highest accuracy, MCC and F1 score and is used in the proposed model

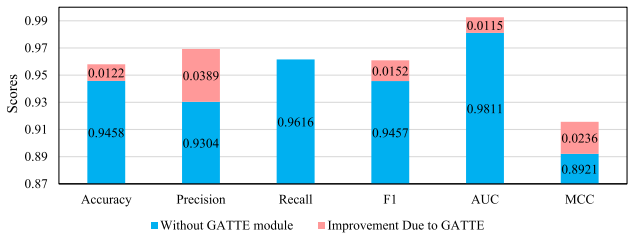


Fig. 12 Impact of introducing the GATTE module to the proposed model. All metrics are improved except the recall score, which confirms the validity of the GATTE module in capturing long-term contextual information from multi-scale tamper trace residuals

6.4 The benefit of adding the global attentional tamper trace extractor (GATTE) module

This section explores the impact of adding the GATTE module to the proposed DenseTrace-Net. Specifically, it demonstrates the improved performance of the model due to the proposed 'Global Attentional Tamper Trace Extractor' module aggregating multi-scale tampering traces globally.

Figure 12 shows that the performance is improved further by the GATTE module, which confirms the benefit of aggregating multi-scale tampering residual features to capture long-term contextual information. When trained and evaluated without the GATTE module (blue), the proposed model scores less accuracy of 0.9458. However, adding the GATTE module (pink) improves the model accuracy by 0.0122, making the final accuracy 0.9685. Similar improvements can be observed across other classification metrics as well. This clearly proves the benefit of adding the GATTE module to the proposed architecture.

6.5 Impact of reducing model complexity

As discussed in Sect. 3.3, introducing LATTE and GATTE modules provides complementary feature learning strength such that removing some dense blocks can reduce the model complexity. So far, in this ablation study, the proposed model contained four dense blocks, four local LATTE modules and one global GATTE module. In this section, dense blocks and their corresponding LATTE modules are progressively removed, and the results are presented in Table 7.

Table 7 indicates the reduced computational complexity of the proposed model. The four dense block architecture contains 7.545 million parameters. Similarly, the three, two and one dense block variants have 4.576, 1.378 and 0.370 million parameters, respectively. This clearly shows that removing two dense block units can substantially reduce the model size.

Table 7 shows that the best results are achieved by removing two dense blocks, which brings down the model complexity from 7.545 million to 1.378 million, while the model also scores the best accuracy, recall, F1 and MCC scores for this configuration.

Hence, the proposed DenseTrace-Net contains two dense blocks, two local LATTE modules and one global GATTE module, as shown in Fig. 1.

7 Conclusion, limitations and future trends

7.1 Conclusion

In this paper, a novel face tampering detection model, *DenseTrace-Net,* is proposed. DenseTrace-Net is designed for efficient feature reuse and improved information flow. It contains a novel Local Attentional Tamper Trace Extractor (LATTE) module to capture local tampering traces at the block level. It also includes a novel Global Attentional Tamper Trace Extractor (GATTE) module that aggregates multi-scale tampering traces to capture long-range contextual relationships of residuals. DenseTrace-Net achieves excellent scores on the CelebDF, FF + + and DFDC datasets (Table 2). The proposed model attains more than 0.95 accuracy and close to 0.9 0 MCC scores for all cases except for NT tampering. DenseTrace-Net outperforms all state-of-the-art facial tampering detection methods, as demonstrated in Tables 3, 4 and 5, establishing its clear superiority.

7.2 Limitations

The proposed DenseTrace-Net model is designed to be extremely lightweight and contains only 1.378 million trainable parameters. This lightweight design, coupled with the local and global tamper trace extractor modules, has proven highly potent in identifying face forgery, as demonstrated by the experimental results in Sect. 5. However, reducing the trainable parameters of a deep learning model also restricts its capability to capture complex features. While DenseTrace-Net has proven to be extremely effective in identifying forged samples from the FF + + , CelebDF and DFDC datasets, it may struggle to classify more realistically tampered facial

Table 7 Impact of reducing dense blocks to reduce model complexity

No. of DenseBlock + LATTE blocks	No. of parameters	ACC	P	R	F1	AUC	MCC
4	7,545,820	0.9580	**0.9693**	0.9526	0.9609	**0.9926**	0.9157
3 (the last block removed)	4,576,124	0.9533	0.9623	0.9413	0.9517	0.9864	0.9068
2 (last two blocks removed)	1,378,588	**0.9685**	0.9688	**0.9647**	**0.9667**	0.9917	**0.9369**
1 (last three blocks removed)	370,290	0.9122	0.8946	0.9403	0.9169	0.9653	0.8252

The best scores are indicated in bold

samples such as hyper-realistic facial images generated from generative adversarial network (GAN) models.

7.3 Future trends

Several interesting future research paths can be undertaken based on the DenseTrace-Net model proposed in this work. *Firstly*, a similar dual-order tamper trace extractor design can be used for localization applications in face tampering detection since these tamper trace extractors are highly potent in highlighting tampering clues. *Secondly,* as more sophisticated visual attention models are proposed, they can be utilized to boost the model performance even further. *Thirdly*, with the recent rise of vision transformer models, tamper trace extraction can be integrated with the patch-based self-attention mechanism to capture long-term dependencies often missed by CNN baseline models .

Author contributions Ankit Yadav: Software, Validation, Investigation, Data Curation, Writing – Original Draft, Visualization. Dhruv Gupta: Software, Validation, Investigation, Data Curation, Writing – Original Draft, Visualization. Dinesh Kumar Vishwakarma: Conceptualization, Methodology, Formal Analysis, Resources, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability The datasets generated during and/or analyzed during the current study are available through an online web repository via the following web links: CelebDF—https://cse.buffalo.edu/~siweilyu/celeb-deepfakeforensics.html. DFDC—https://ai.facebook.com/datasets/dfdc/. FaceForensics + +—https://github.com/ondyari/FaceForensics.

Declarations

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

References

1. "DeepFakes," GitHub, 14 August 2020. [Online]. https://github.com/deepfakes/faceswap. Accessed 08 July 2022
2. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: real-time face capture and reenactment of RGB videos. Commun. ACM **62**(1), 96–104 (2019)
3. "DeepFaceLab," GitHub, 18 March 2020. [Online]. https://github.com/iperov/DeepFaceLab. Accessed 08 July 2022
4. Wang, G., Jiang, Q., Jin, X., Li, W., Cui, X.: MC-LCR: Multimodal contrastive classification by locally correlated representations for effective face forgery detection. Knowl. Based Syst. **250**, 109114 (2022)
5. Hu, Z., Duan, Q., Zhang, P., Tao, H.: An attention-erasing stripe pyramid network for face forgery detection. Signal Image Video Process. (2023). https://doi.org/10.1007/s11760-023-02644-6
6. Masud, U., Sadiq, M., Masood, S., Ahmad, M., El-Latif, A.A.A.: LW-DeepFakeNet: a light-weight time distributed CNN-LSTM network for real-time DeepFake video detection. Signal Image Video Process. (2023). https://doi.org/10.1007/s11760-023-02633-9
7. Shang, Z., Xie, H., Zha, Z., Yu, L., Li, Y., Zhang, Y.: PRRNet: pixel-region relation network for face forgery detection. Pattern Recognit (2021). https://doi.org/10.1016/j.patcog.2021.107950
8. Xu, Z., Liu, J., Lu, W., Xu, B., Zhao, X., Li, B., Huang, J.: Detecting facial manipulated videos based on set convolutional neural networks. J. Vis. Commun. Image Represent. (2021). https://doi.org/10.1016/j.jvcir.2021.103119
9. Chen, Z., Yang, H.: Attentive semantic exploring for manipulated face detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto (2021)
10. Heo, Y.J., Yeo, W.-H., Kim, B.-G.: DeepFake detection algorithm based on improved vision transformer. Appl. Intell. **53**, 7512–7527 (2023)
11. Bonomi, M., Pasquini, C., Boato, G.: Dynamic texture analysis for detecting fake faces in video sequences. J. Vis. Commun. Image Represent. (2021). https://doi.org/10.1016/j.jvcir.2021.103239
12. Yang, J., Xiao, S., Li, A., Lan, G., Wang, H.: Detecting fake images by identifying potential texture difference. Futur. Gener. Comput. Syst. **125**, 127–135 (2021)
13. Yang, J., Li, A., Xiao, S., Lu, W., Gao, X.: MTD-Net: Learning to Detect Deepfakes Images by Multi-Scale Texture Difference. IEEE Trans. Inf. Forensics Secur. **16**, 4234–4245 (2021)
14. Caldelli, R., Galteri, L., Amerini, I., Bimbo, A.D.: Optical Flow based CNN for detection of unlearnt deepfake. Pattern Recogn. Lett. **146**, 31–37 (2021)
15. Kohli, A., Gupta, A.: Detecting DeepFake, FaceSwap and Face2Face facial forgeries using frequency CNN. Multimedia Tools and Applications **80**, 18461–18478 (2022)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR), Austria (2021)
17. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: CvT: Introducing convolutions to vision transformers. In: International Conference on Computer Vision (ICCV), Montreal (2021)

18. Heo, Y.J., Yeo, W.-H., Kim, B.-G.: DeepFake detection algorithm based on improved vision transformer. Appl. Intell. (2022)
19. Guo, Z., Yang, G., Chen, J., Sun, X.: Fake face detection via adaptive manipulation traces extraction network. Comput. Vis. Image Underst. (2021). https://doi.org/10.1016/j.cviu.2021.103170
20. Amerini, I., Galteri, L., Caldelli, R., Bimbo, A.D.: Deepfake video detection through optical flow based CNN. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul (2019)
21. Caldelli, R., Galteri, L., Amerini, I., Bimbo, A.D.: Optical flow based CNN for detection of unlearnt deepfake manipulations. Pattern Recogn. Lett. **146**, 31–37 (2021)
22. Hu, J., Liao, X., Wang, W., Qin, Z.: Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Trans. Circuits Syst. Video Technol. (2021). https://doi.org/10.1109/TCSVT.2021.3074259
23. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle (2020)
24. Choi, D.H., Lee, H.J., Lee, S., Kim, J.U., Ro, Y.M.: Fake video detection with certainty-based attention network. In: IEEE International Conference on Image Processing (ICIP), Abu Dhabi (2020)
25. Lu, C., Liu, B., Zhou, W., Chu, Q., Yu, N.: Deepfake video detection using 3D-attentional inception convolutional neural network. In: IEEE International Conference on Image Processing (ICIP), Anchorage (2021)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas (2016)
27. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: European Conference on Computer Vision, Amsterdam (2016)
28. Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu (2017)
29. J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018
30. Park, J., Woo, S., Lee, J.-Y., Kweon, I.S.: BAM: bottleneck attention module (2018). https://arxiv.org/pdf/1807.06514.pdf
31. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: European Conference on Computer Vision, Munich, Germany (2018)
32. Zhou, Y., Luo, A., Kang, X., Lyu, S.: Face forgery detection based on segmentation network. In: IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA (2021)
33. Zhang, J., Ni, J., Xie, H.: DeepFake videos detection using self-supervised decoupling network. In: IEEE International Conference on Multimedia and Expo (ICME), Shenzhen (2021)
34. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: a large-scale challenging dataset for DeepFake forensics. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA (2020)
35. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (DFDC) preview dataset (2019). https://arxiv.org/abs/1910.08854
36. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M.: FaceForensics++: learning to detect manipulated facial images. In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South) (2019)
37. Yang, G., Wei, A., Fang, X., Zhang, J.: FDS_2D: rethinking magnitude-phase features for DeepFake detection. Multimed. Syst. (2023). https://doi.org/10.1007/s00530-023-01118-6
38. Guo, Z., Yang, G., Zhang, D., Xia, M.: Rethinking gradient operator for exposing AI-enabled face forgeries. Expert Syst. Appl. (2023). https://doi.org/10.1016/j.eswa.2022.119361
39. Xu, K., Yang, G., Fang, X., Zhang, J.: Facial depth forgery detection. Multimed. Tools Appl. (2023). https://doi.org/10.1007/s11042-023-14626-4
40. Lin, H., Huang, W., Luo, W., Lu, W.: DeepFake detection with multi-scale convolution and vision transformer. Digit. Signal Process. (2023). https://doi.org/10.1016/j.dsp.2022.103895
41. Yang, G., Xu, K., Fang, X., Zhang, J.: Video face forgery detection via facial motion-assisted capturing dense optical flow truncation. Vis. Comput. (2022). https://doi.org/10.1007/s00371-022-02683-z
42. Nirkin, Y., Wolf, L., Keller, Y., Hassner, T.: DeepFake detection based on discrepancies between faces and their context. IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 6111–6121 (2022)
43. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA (2021)
44. Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., Yu, N.: Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA (2021)
45. Hu, J., Liao, X., Wang, W., Qin, Z.: Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Trans. Circuits Syst. Video Technol. **32**(3), 1089–1102 (2021)
46. Chen, H.-S., Rouhsedaghat, M., Ghani, H., Hu, S., You, S., Kuo, C.-C.J.: DefakeHop: a light-weight high-performance deepfake detector. In: IEEE International Conference on Multimedia and Expo (ICME), Shenzhen (2021)
47. Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA (2020)
48. Baek, J.-Y., Yoo, Y.-S., Bae, S.-H.: Generative adversarial ensemble learning for face forensics. IEEE Access **8**, 45421–45431 (2020)
49. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.-G.: WildDeepfake: a challenging real-world dataset for deepfake detection. In: 28th ACM International Conference on Multimedia (2020)
50. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: MesoNet: a compact facial video forgery detection network. In: IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China (2018)
51. Mohiuddin, S., Sheikh, K.H., Malakar, S., Velásquez, J.D., Sarkar, R.: A hierarchical feature selection strategy for deepfake video detection. Neural Comput. Appl. (2023). https://doi.org/10.1007/s00521-023-08201-z
52. Deng, L., Wang, J., Liu, Z.: Cascaded network based on EfficientNet and transformer for deepfake video detection. Neural. Process. Lett. (2023). https://doi.org/10.1007/s11063-023-11249-6
53. Asha, S., Vinod, P., Menon, V.G.: A defensive framework for deepfake detection under adversarial settings using temporal and spatial features. Int. J. Inf. Secur. (2023). https://doi.org/10.1007/s10207-023-00695-x
54. Ke, J., Wang, L.: DF-UDetector: An effective method towards robust deepfake detection via feature restoration. Neural Netw. **160**, 216–226 (2023)
55. Guo, Z., Yang, G., Wang, D., Zhang, D.: A data augmentation framework by mining structured features for fake face image detection. Comput. Vis. Image Underst. (2023). https://doi.org/10.1016/j.cviu.2022.103587
56. Zhao, C., Wang, C., Hu, G., Chen, H., Liu, C., Tang, J.: ISTVT: interpretable spatial-temporal video transformer for deepfake detection. IEEE Trans. Inf. Forensics Secur. **18**, 1335–1348 (2023)
57. Yu, Y., Zhao, X., Ni, R., Yang, S., Zhao, Y., Kot, A.C.: Augmented multi-scale spatiotemporal inconsistency magnifier for generalized DeepFake detection. IEEE Trans. Multimed. **99**, 1–13 (2023)

58. Yang, Z., Liang, J., Xu, Y., Zhang, X.-Y., He, R.: Masked relation learning for DeepFake detection. IEEE Trans. Inf. Forensics Secur. **18**, 1696–1708 (2023)
59. Ganguly, S., Ganguly, A., Mohiuddin, S., Malakar, S., Sarkar, R.: ViXNet: vision transformer with xception network for deepfakes based video and image forgery detection. Expert Syst. Appl. (2022). https://doi.org/10.1016/j.eswa.2022.118423
60. Ganguly, S., Mohiuddin, S., Malakar, S., Cuevas, E., Sarkar, R.: Visual attention-based deepfake video forgery detection. Pattern Anal. Appl. **25**, 981–992 (2022)
61. Nadimpalli, A.V., Rattani, A.: On improving cross-dataset generalization of deepfake detectors. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA (2022)
62. Li, G., Cao, Y., Zhao, X.: Exploiting facial symmetry to expose deepfakes. In: IEEE International Conference on Image Processing (ICIP), Anchorage (2021)
63. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face X-ray for more general face forgery detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle (2020)
64. Luo, Z., Kamata, S.-I., Sun, Z.: Transformer and node-compressed Dnn based dual-path system for manipulated face detection. In: IEEE International Conference on Image Processing (ICIP), Anchorage (2021)
65. Qi, H., Guo, Q., Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., Zhao, J.: DeepRhythm: exposing DeepFakes with attentional visual heartbeat rhythms. In: 28th ACM International Conference on Multimedia, Lisboa (2020)
66. Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., Xue, H., Lu, Q.: Sharp multiple instance learning for DeepFake video detection. In: Proceedings of the 28th ACM International Conference on Multimedia, Seattle WA USA (2020)
67. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emotions don't lie: an audio-visual deepfake detection method using affective cues. In: 28th ACM International Conference on Multimedia, Lisboa (2020)
68. Montserrat, D.M., Hao, H., Yarlagadda, S.K., Baireddy, S., Shao, R., Horváth, J., Bartusiak, E., Yang, J., Güera, D., Zhu, F., Delp, E.J.: Deepfakes detection with automatic face weighting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle (2020)
69. Chugh, K., Gupta, P., Dhall, A., Subramanian, R.: Not made for each other-audio-visual dissonance-based deepfake detection and localization. In: 28th ACM International Conference on Multimedia, Lisboa (2020)
70. Li, G., Zhao, X., Cao, Y.: Forensic symmetry for DeepFakes. IEEE Trans. Inf. Forensics Secur. **18**, 1095–1110 (2023)
71. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision (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.