# SCHOLARLY PUBLICATIONS

*A CURRENT AWARENESS BULLETIN*
*OF RESEARCH OUTPUT*

## *@DTU*

(149th Edition)

***MAY 2025***

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

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

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

This issue contains the information published during **May,2025**. The arrangement of the contents is alphabetical. The full text of the article which is either subscribed by the university or available in the web is provided in this bulletin.

**Central Library**

# CONTENTS


1. A Minimalist MOS-based Meminductor Emulator Circuit, ***6.Rahul Kumar Gupta, 3.Mahipal Singh Choudhry***, Varun Saxena, ECE, DTU

2. A review on breast cancer diagnostic techniques, Parikshana Mathur, a Saakshi Dhanekar and ***3.1 B. D. Malhotra***, Biotechnology, DTU

3. An energy and cost-effective realization of high-performance singlelayer vedic multiplier in quantum-dot cellular automata, ***6.Hemanshi Chugh and 3.Sonal Singh***, ECE, DTU

4. Assessment of arsenic contamination in drinking water sources and its efects on human health in rural regions: a study in the Malistan district, Ghazni province of Afghanistan, ***6. Ali Reza Noori, 3. S. K. Singh***, · Assadullah Reza, ENVT, DTU

5. An Efficient Attention-Based LSTM Framework for Blood Glucose Level Prediction***, 3.Sunny Arora , 3.Shailender Kumar***, Pardeep Kumar, Computer Science, DTU

6. A Novel Deep Learning-Assisted Framework for the Assessment of Real-Time Dust Accumulation Data on Solar PV Modules, ***6.Rahma Aman, 6.Astitva Kumar*** and ***3.M. Rizwan***, , Electrical Engg., DTU

7. Brain-Inspired Learning Controller for Grid-Connected Electric Vehicle Charging Junction with Improved Power Quality, ***6.1 Amarendra Pandey, 3.Alka Singh*** and ***6.1 Sudhanshu Mittal***, Electrical, DTU

8. Consequence of Ultrasonic Vibration-Assisted Cold Metal Transfer Welding Parameters on Metallurgical and Mechanical Properties of Ferritic Stainless-Steel-Welded Joints***, 6.Alka Sharma, 3.Reeta Wattal***, & ***3.M.S. Niranjan***, Mechanical, DTU

9. Cellulose Microfibers-Embedded Carboxymethyl Tamarind Kernel Gum Hydrogels as Soil Conditioners and Plant Nutrients Carriers, ***6.Tarun Kumar Gayen,*** Mohammad Amdad Ali, and ***3.Sudhir G. Warkar***, Applied Chemistry, DTU

10. Deciphering the Sustainable Stormwater Management Strategies for Urban Areas: a Review, ***6.Shivani Yadav, 6.Saurav Ambastha, 6.Harsh Pipil1 R. Reddy, 3.Anil Kumar Haritash*** and Krishna, Environmental, DTU

11. Design of DTMOS Based Preamplifier for Low Power and High Speed Comparator, ***6.Akshay Mann, 3.Neeta Pandey,*** *and* ***3.Maneesha Gupta***, ECE, DTU

12. Dispersion-engineered segmented core rib waveguide for coherent and broadband supercontinuum generation in both all-normal and anomalous regimes, ***6.Deepak Garg and 3.Ajeet Kumar***, Applied Physics, DTU

13. EAGLE: Enhanced Activity Recognition in Occluded Environments With Improved GAIN and Bi-LSTM, ***6.Diksha Kurchaniya*** and ***6..Sanjay Kumar***, CSE, DTU

14. Efficient Hardware Implementation of Multi-Armed Bandit Algorithms for RIS-Aided Wireless Networks, ***6.Ishaan Sharma ,3. Rohit Kumar*** and Sumit J Darak, ECE, DTU

15. Empathizing with parents and teachers for educating children with special learning needs, ***6.Shraddha Chadha, 3.Ravindra Singh*** and ***3. Lalit Kumar Das***, Design, DTU

16. Empirical Investigation into the Interlinkages Between the Price Movements of Energy Commodities and Stock Market in India, Dr. Nishi Sharma and ***7.1.Tarandeep Kaur***, Economics, DTU

17. Enhanced software change proneness prediction in android applications using balanced data techniques and advanced language models, ***3.Ruchika Malhotra*** and ***6.Jyoti Patidar,*** Software Engg., DTU

18. Evidence of Surface Biasing and Multiferroic Properties in Ferroelectric (BNT)-Ferrimagnetic (LFO) Nanocomposites, ***6.Bineet Kumar,*** Lagen Kumar Pradhan, Amar Dev, ***3.Amrish K. Panwar*** and Manoranjan Kar, Applied Physics, DTU

19. Examining the impact of domestic monetarypolicy on foreign portfolio investment flowsto India, ***3.Virender Kumar***,USME, DTU

**20.** EECRPOA: energy efcient clustered routing using pelican optimization algorithm in WSNs, ***6.Yogendra Mohan, 3.Rajesh Kumar Yadav*** and · Manisha Manjul, Computer Science, DTU

**21.** Exploring Blockchain-Based Patient-Centric Healthcare: A Comprehensive Review, ***6.Deepak Kumar Mishra*** and ***3.Pawan Singh Mehra***, CSE, DTU

**22.** Faces in the Fog: A Deep Dive into Face De-identification Techniques and Their Comparative Analysis, ***8.Shaonti Goswami***, ***8.Sudipty Konica Paul, 8.Sanyam Jawlia*** and ***3.Anurag Goel***, CSE, DTU

**23.** Flood-FireNet: a novel flood image classification architecture using the adaptive firefly algorithm***, 6.Vinay Dubey, 3.Rahul Katarya***, Computer Science, DTU

**24.** Finite Element Analysis of Concrete Bridge Under Moving Load Across and Along the Span, ***7.Bimal Sharma, 3.Shilpa Pal*** and Naveet Kaur, CIVIL, DTU

**25.** FeO NPs-Loaded Guar Gum/Xanthan Gum-Based Hydrogel: ASustainable Approach for Cationic Dye Removal, ***6.Priyanka Yadav, 6.Kanishka Kashyap, 3.Sudhir G. Warkar***, and ***3.Anil Kumar***, Applied Chemistry, DTU

**26.** GDI and Dynamic Logic: A synergy behind the modern electronics, ***6.Y. Parmar, 3.N. Pandey***, ECE, DTU

**27.** HEMT based Different Structures for High Threshold and Breakdown Voltages: A Review, ***8.Aastha Singh, 3.Sachin Dhariwal***, ECE, DTU

**28.** HEP-DHMI: The High Efficiency and Payload Data Hiding Scheme for AMBTC Compressed Medical Images, Chia-Chen Lin , En-Ting Chu, Iuon-Chang Lin and ***3.Rajeev Kumar***, CSE, DTU

**29.** High-Gain Fault-Tolerant Multiport Bidirectional DC–DC Converter With Time Multiplexing Control, ***6.Abhishek Kumar Gupta,3.Mayank Kumar and 3.Madhusudan Singh,*** Electrical, DTU

**30.** Intelligent Nanorobots for Precision Cancer Diagnosis and Treatment using Deep Reinforcement Learning, ***3.Kapil Sharma*** and ***6.Ravi Ranjan***, IT, DTU

**31.** Investigation of Pollution by Toxic Metals in the Bed Sediments of Ganga River Along Kanpur to Prayagraj City of Uttar Pradesh, India, ***8.1 Mohit Aggarwa***l, ***3.Srinivasan Anbukumar*** and ***3.Thangamani Vijaya Kuma***r, Civil, DTU

**32.** Keystroke Dynamics Based User Authentication Focusing on Fixed Text Approaches***, 6.Anurag Kamra, 8.Sanya Khurana and 3.Anurag Goel,*** CSE, DTU

**33.** Machine Learning for Advancing Electric Vehicles Security Leveraging CICEVSE2024 Anmol, ***8.Geetesh Kumar, 8.Amandeep*** and ***3.Sonika Dahiya***, Software Engg., DTU

**34.** Mitigating Thermal Effects in MEMS Gyroscopes: A Novel Substrate Integration Approach***, 6. Shaveta***, R. K. Bhan , S.S. Lamba , Vinita Gond and ***3.Rishu Chaujar,*** Applied Physics, DTU

**35.** Optimizing thermoelectric performance of halide perovskite Ca3AsBr3 through strain engineering, ***7.Sumedha Yadav, 6.Sangeeta, 6.Kulwinder Kumar, 3.Mukhtiyar Sing***, Applied Physics, DTU

**36.** Online Proctoring Systems - A Hardware Implementation using Enhanced Eye Detection and Gaze Tracking, ***6.Ashish Chotani*** and ***3.Rohit Kumar***, ECE, DTU

**37.** Photoluminescence Behavior of Sm3+-Doped $YPO_4$ Phosphors for Solid-State Lighting, Vijay Singh, ***6.Sheetal Kumari***, M. M. Haidari, Aadil Ahmad Bhat, ***3.A. S. Rao*** and Ji Bong Joo, Applied Physics, DTU

**38.** Privacy-preserving data sharing in blockchain-enabled IoT healthcare management system***, 6. Himanshu Nandanwar and 3.Rahul Katarya***, Computer Science, DTU

**39.** Polarization Controlled, Dual-Band, Active Plasmonic Switch Based on T-Shaped Nanopillars on a VO2-Coated Plasmonic Substrate, ***6.Asbah Masih, · 3.Yashna Sharma,*** ECE, DTU

**40.** Performance Analysis of a Dual FSO/RF Hybrid Link in the Presence of Malicious Intruder, 6.***Himani Mehta, 6.Vishal Kumar*** and ***3.Rohit Kumar***, ECE, DTU

**41.** Quantum Edge Cloud Computing: Revolutionizing IoT, ***8.Aanya Tiwari, 8.Ahsaan Ahmad Ahanger, 8.Anjali Meena*** and ***3.Pawan Singh*** Mehra, CSE, DTU

**42.** QNet: exploring deep learning for quantum code smell detection***, 3.Ruchika Malhotra 6.1 Bhawna Jain*** and Marouane Kessentini, Software Engg., DTU

**43.** Risk Analysis of Indian Banking Stocks Using Machine Learning Techniques, ***6.Vidit Tayal, 6.Snehal Kumar, 3.Shweta Meena***, Software, DTU

**44.** Routing Optimization in Hybrid Software-Defined Networks: A Heuristic Approach, ***6.Samiullah Mehraban & 3.Rajesh K Yadav***, Computer Science,DTU

**45.** Recent Advances on Catalysts Based on MoS 2 : A Review onDiverse Application, Priya, Suman Srivastava and ***6.1 Ritika Khatri***, Physics, DTU

**46.** Structural, Judd–Ofelt and Photoluminescence Properties of Dy3+ Ions Doped SrO-MgO-Na2O-Al2O3-B2O3 Glasses for Visible Laser and W-LED Applications, G. Muralidhar, Mahamuda Shaik, · P. Sailaja· K. Swapna, M. Venkateswarlu, G. Dedeepya and ***3.A. S. Rao,*** Applied Physics, DTU

***47.*** Sum Rate Maximization and Performance Analysis of NOMA-Based IRS-Aided Uplink Visible Light Communication System, ***6.Vishal Kumar, 3.Rohit Kumar,*** ECE, DTU

***48.*** Swarm Intelligence in Robotics: Optimizing PSO Parameters for Target Search in Single and Multiple Scenarios, ***6.Vikas Sharma, 3.Rahul Gupta*** and ***3.Kapil Sharma***, IT, DTU

***49.*** Sustained release of epigallocatechin gallate from calcium carbonate nanoparticles: promising therapy for colorectal cancer, ***6.Ritu, 3.Prakash Chandra*** and ***3.Asmita Das***, Biotechnology, DTU

***50.*** Technological Optimization, Health Risk Reduction, and Economic Analysis of Electrokinetic Remediation of Cadmium-contaminated Soil, ***6. Sonam Taneja,*** Öznur Karaca and ***3.Anil Kumar Haritash***, Environmental, DTU

***51.*** Towards secure healthcare: SA-GBO-ODBN model utilizing Blockchain and deep learning for data handling and diagnosis***, 6. Nikhil Sharma*** and ***3.Prashant Giridhar Shambharkar,*** CSE, DTU

***52.*** Traffic Engineering Optimization in Hybrid Software-Defined Networks: A Mixed Integer Non-Linear Programming Model and Heuristic Algorithm, ***6. Samiullah Mehraban*** and ***3.Rajesh K. Yadav***, CSE, DTU

***53.*** Twitter Sentiment Analysis of the Electric Vehicle Industry in India using NLP***, 8. Rishabh Sharma***, Meha Joshi, EE, DTU

***54.*** Unpaved Road Damage Detection via Semantic Segmentation: Enabling Smarter ADAS, ***6. Ruchi Tripathi, 3.S.Indu*** and ***3.Rohit*** **Kumar**, ECE, DTU

***55.*** Up-conversion luminescence study of thermally stable Sr2ZnSi2O7: Er3+ phosphors, ***6. Seema , 3.A.S. Rao and 3.Sarita Baghel***, Applied Physics, DTU

***56.*** Ultrasensitive terahertz metasurface for biomolecule detection based on toroidal dipole resonance, ***6.Pooja Agarwal, 3.Kamal Kishor*** and ***3.Ravindra Kumar Sinha***, Applied Physics, 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*

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-025-10222-w

RESEARCH ARTICLE-ELECTRICAL ENGINEERING

# A Novel Deep Learning-Assisted Framework for the Assessment of Real-Time Dust Accumulation Data on Solar PV Modules

**Rahma Aman[1] · Astitva Kumar[2] · M. Rizwan[1,3]**

Received: 3 November 2024 / Accepted: 14 April 2025
© King Fahd University of Petroleum & Minerals 2025

**Abstract**
Soiling, usually caused by dust deposition, is a significant difficulty for photovoltaic (PV) systems because it obstructs light absorption, resulting in reduced power output and the creation of hot spots. Monitoring the soiling ratio can give critical insights into PV module performance under dirty circumstances, allowing for power output prediction. This study gives a complete performance analysis of PV modules impacted by soiling based on real-time data obtained over six months in Badli, New Delhi, India. A soiling monitoring system captured key variables such as soiling ratio, transmission loss, and temperature, which served as the foundation for the development of a mathematical model to estimate PV power production under soiled circumstances. The findings show that dust deposition causes a considerable 17% loss in power production, emphasizing the necessity for effective soiling mitigation techniques. Deep learning models, especially stacked long short-term memory (LSTM) and bidirectional LSTM, were used to forecast power output under soiling circumstances. Stacked LSTM outperformed Bi-LSTM, with a $R^2$ score of 0.9913 and a mean squared error (MSE) of 0.0078. Training time was 17.35 s. By precisely estimating dirty power output, this study makes it easier to schedule cleaning cycles, improves PV module performance, and contributes to sustainable solar energy generation in dust-prone areas.

**Keywords** Dust accumulation · Power prediction · Solar modules · Soiling ratio · Deep learning model

## 1 Introduction

As global fossil fuel reserves deplete, there is an increasing shift toward renewable energy sources. Solar photovoltaic (SPV) systems have significant promise, particularly for supplying power to rural places. These technologies also help to reduce greenhouse gas emissions, which generates increased global interest and investment. However, shifting to a completely renewable energy system on a global scale poses major problems [1]. Due to meteorological circumstances, the efficiency of a solar PV module is restricted, thus it is vital to consider elements such as ambient temperature, solar irradiance, airborne dust, humidity, and wind speed. Accumulation of dust and other particles on solar collectors leads to a decrease in transmission efficiency. This applies to the losses in energy transmission in both PV and concentrated photovoltaic (CPV) systems, as well as the losses due to reflection in concentrated solar power (CSP) systems. Figure 1 illustrates the decline in performance of solar plants due to factors such climatic conditions, due to properties of SPV Systems. The ideal positioning of PV power plants relies on the geographical coordinates, which are irradiance, latitude, longitude, tilt angle, and orientation, among other variables. However, the solar PV modules are affected by several climatic factors that significantly influence the power production, conversion efficiency, and energy cost. Orientation of SPV system helps further sun exposure, coatings, electrostatic charge, tilt angle, and wind interaction all have

✉ Rahma Aman
rahmaaman_2k21phdee08@dtu.ac.in

Astitva Kumar
astitvakumar.27@gmail.com

M. Rizwan
rizwan@tsue.uz

[1] Department of Electrical Engineering, Delhi Technological University, New Delhi, Delhi 110042, India

[2] Department of Electrical Engineering, Netaji Subash University of Technology, New Delhi, Delhi, India

[3] Department of Information Systems and Technologies, Tashkent State University of Economics, Tashkent, Uzbekistan

Published online: 17 May 2025

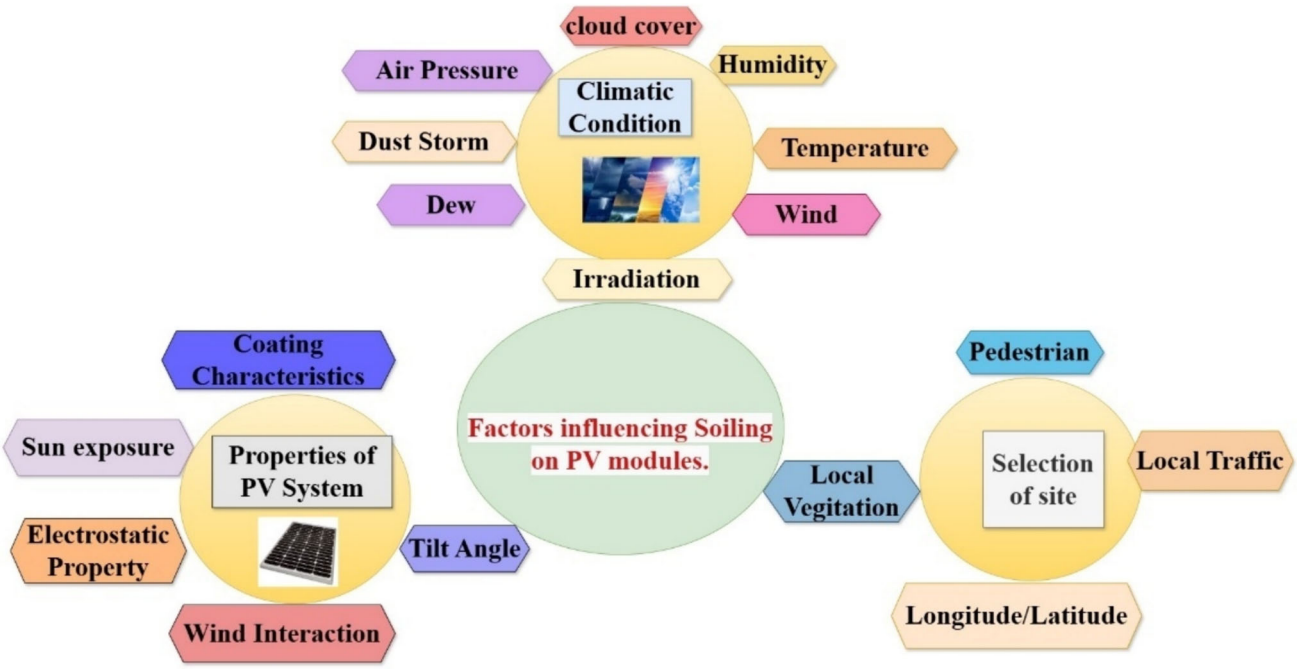



**Fig. 1** Factors effecting dust deposition on SPV systems

an impact on dust collection, which affects soiling rates and module performance overall.

### 1.1 Literature Review

Dust significantly impacts the performance and productivity of SPV modules. The characteristics of dust (such as kind, size, shape, and meteorology) vary according on the geographical location [2].

Environmental variables directly affect PV cell performance and efficiency, making them important in mathematical modeling of PV systems. By including these parameters, models may better anticipate PV system behavior under different situations. To achieve the optimal design of PV systems with great dependability, mathematical models are used, and the impacts of weather conditions and solar irradiation are included in these models. Researchers wanted to provide models and methods to determine PV system variables. Literature studies have modeled transmittance ratio [3], reduction in solar energy gain [4], degradation rate [5] solar cell power, efficiency, and I-V characteristics [6].

The sophistication of models is enhanced by the inclusion of additional factors that influence the system's performance. However, there is always a trade-off between the accuracy, complexity, and number of factors in the derived model.

More than 100 parameters have been examined in [7] to predict the soiling losses in PV systems at 20 stations in the USA. Particulate matter (PM), particularly PM10 and PM2.5, as well as certain precipitation parameters and the average number of days between consecutive rainfalls, were identified as the most effective soiling predictors. In [8], a neural network (NN) model was proposed that predicts SPV power output by considering the physical, chemical, and spectral characteristics of soil. The model utilizes a hybrid data clustering algorithm to provide efficient data preprocessing and a new data division technique to ensure that the training data set contains an adequate amount of data. In comparison to other networks, the neural network hybrid (NNH) models demonstrated superior performance for both known and unknown soil samples. A regression model (RM) and neural network (NN) were devised in another study [9] to forecast the loss of PV power as a result of artificial soiling. The NN model was found to be more effective at certain irradiance levels and soil types, while the RM model was more effective at other levels and types. Both models were experimentally validated at a diversity of irradiance levels and five types of soils. In order to determine the power loss resulting from soiling in 1 MWp plants in Southern Italy, four Bayesian neural network (BNN) models were created to be used both before and after cleansing [10]. In [11], author discusses that after global semiconductor materials science improved solar panel efficiency, solar energy has become a clean, renewable source of electricity in high demand. Solar panel performance is not limited by design or materials; however, flaws can reduce or disrupt it. Some ways are suggested to address these dysfunctions and identify them early. The literature classifies them as statistical, analytical, and AI. After reviewing recent

diagnosis methods, the authors in this work proposed a simple, reliable, and efficient method based on the fill factor FF and the maximum short-circuit current Isc. In [12] author seeks to create a dust-solar cell model. The single-diode model models PV cell performance based on current–voltage relationships. Soiled solar modules affect installation production. The module's angle of inclination and climatic conditions—aerosols, relative humidity, ambient temperature, pressure, etc.—cause this effect. In [13] authors setup an experimental PV plant to analyze soiling losses and anticipate PV plant production in one to three hours. The author in [14] develops a computational model to research the impacts of residue and encompassing temperature on the presentation of a PV framework. The models assess change productivity using improved counterfeit brain organization ANN models' multi-input multi-output. The model-building approach is displayed and confirmed for accuracy using several measures. Daily cleaning boosts solar PV system performance, and an artificial neural network boosts accuracy to 99.8%. The author in [15] uses artificial neural networks to assess and anticipate the power production of a grid-connected 20-kWp solar power plant in Tiruchirappalli, India, a reputable manufacturing industry. A multilayer perceptron-based ANN model is suggested for power generation day-ahead prediction. A monocrystalline silicon solar module's power was predicted by an artificial neural network. PV module soiling refers to the deposition of dust, soil, and microfibers from the environment, as well as the growth of minute particles such as moss and fungi. It is a lesser-known factor that substantially reduces the power output by operating as a barrier for photons of effective light used by a module. The anticipated loss in irradiance and power can be calculated using a soiling ratio ($SR$) parameter [16]. Author in [17] presents methodologies to quantify $SR$ and soiling rate ($S_{rate}$) for two representative commercial technologies, polycrystalline or multi crystalline silicon (mc-Si) and thin-film cadmium telluride (CdTe) modules, through soiling monitoring stations deployed in the selected climate regions. These methodologies can be applied to polycrystalline or multi crystalline silicon (mc-Si) and thin-film cadmium telluride (CdTe) modules. Data used to compute $SR$ and $S_{rate}$ in accordance with the international standard IEC 61724–1/2017 were used. Since solar module cleaning is essential, constant monitoring and evaluation are required to optimize these processes [18]. The integration of advanced machine learning algorithms has proven efficacy in enhancing model accuracy and predictive capacities in photovoltaic (PV) systems. Recent developments in finite-element-informed neural networks have been effectively utilized in parametric simulations within structural mechanics, improving generalization and efficiency in simulating intricate mechanical processes. [19] introduced a particle filter methodology for structural health monitoring that incorporates finite-element-informed learning to estimate time-invariant parameters, therefore enhancing reliability and robustness in predictive modeling.

Moreover, physics-informed neural networks (PINNs) have exhibited exceptional efficacy in addressing differential equations related to solid mechanics issues. [20] presents physics-informed neural networks (PINNs) as a framework that integrates physical constraints into neural networks, hence guaranteeing stable and precise solutions for nonlinear partial differential equations (PDEs). This methodology has been extensively employed in mechanical and materials sciences to improve model interpretability and generalization. Moreover, Fourier Neural Operators (FNOs) have arisen as an effective instrument for addressing parametric PDEs, enhancing the capacity to simulate fluctuations in system behavior. [21] introduced FNOs as a sophisticated deep learning method that learns mappings across function spaces, facilitating fast and scalable solutions to intricate physics-based challenges. This approach has been very effective in simulating dynamic environmental variables influencing photovoltaic performance.

Integrating data-driven machine learning frameworks with dust accumulation modeling may substantially improve solar energy forecasting, optimize power loss prediction, and boost operational efficiency in solar photovoltaic systems. These developments allow the development of more precise, real-time adaptive solutions that mitigate the damaging effects of dust on photovoltaic module performance, hence enhancing solar power production and maintenance techniques.

### 1.2 Problem Statement

Dust deposition lowers the solar cells' exposure to sunlight, resulting in a reduction in the total energy production. Despite its significance, there is lack of comprehensive models that precisely forecast the influence of dust on solar PV efficiency. The objective of this work is to create a strong and accurate model that can measure and reduce the impact of dust on solar PV systems. In recent years, most PV power plant owners and businesses have been extremely concerned about this issue. It is necessary to realize unattended periodic cleaning, intelligent dust removal, and snow removal of PV modules through low cost and high reliability, and to improve the efficiency of cleaning dust on the surface of solar modules by using intelligent cleaning robots increases PV power plant lifespan.

### 1.3 Motivation and Gaps

Despite the extensive research conducted on the influence of environmental factors on PV systems, there are still gaps in the understanding and prediction of the effects of soiling on solar modules power output. After reviewing several papers,

the aspect of dust accumulation can be considered and its impact in SPV prediction need to be quantified. Further the various research suggests that using machine learning techniques may enable precise prediction of several data such as temperature, relative humidity, solar radiation, rainfall, and wind speed, hence facilitating reliable solar prediction and wind speed estimation [22, 23].

The primary focus of current research is to access the impact of dust on solar irradiance, temperature and solar power production. However, there are shortage of comprehensive models for hourly power loss estimation from soiling utilizing advanced techniques like deep learning (DL) indicates to a major research opportunity. A development of such models would significantly enhance the capacity to precisely forecast contaminated power loss and maximize solar system performance, providing much more attention to work in this area of research. In addition, it is imperative to continue monitoring and evaluating the performance of PV modules in order to optimize cleaning processes and improve efficiency. To address the above issues, this paper develops a mathematical model to investigate the effects of dust accumulation on photovoltaic modules. Furthermore, the research analyzes power loss due to dust accumulation and proposes soiled SPV power forecast models. The suggested frameworks employ deep learning approaches such as Stacked LSTM and Bi LSTM. These deep learning-based soiling models are rarely available in the literature. Furthermore, power prediction models should increase hierarchical feature extraction, long-term dependency learning, redundancy reduction, and better scalability for multi-step power forecasting with less computing cost.

### 1.4 Contribution of the Work

The objective of this paper is to rectify these deficiencies by employing soiling ratio data obtained from a Dust IQ soil monitoring system, a specialized instrument designed to evaluate dust accumulation which is installed on a 5-kW solar PV system located at the rooftop of Utilization lab in Electrical department, Delhi technological university, New Delhi. The study suggests the utilization of two deep learning algorithms, Stacked Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM), to forecast soiled solar PV power at one-hour intervals. The research aims to quantify power loss due to dust accumulation by forecasting the power output of contaminated solar PV modules and comparing it to clean modules through the training of these algorithms on the supplied dataset.

The main contribution of the proposed work:

- Real-Time Performance Evaluation: Performed a comprehensive examination of soiled PV modules utilizing live data for a duration of six months, with a specific emphasis on the ratio of soil accumulation, loss of light transmission, and temperature.
- Mathematical Modeling: Develop mathematical model to forecast the power generation of soiled SPV modules, improving the capacity to control and reduce performance degradation caused by dust accumulation.
- Effective Deep Learning Models: Stacked LSTM and Bidirectional LSTM models were used to forecast mathematically calculated soiled PV power for better accuracy.

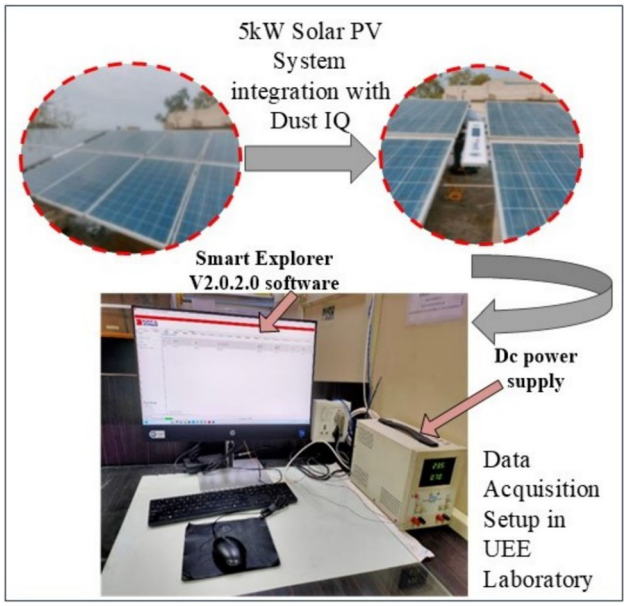



**Fig. 2** Dust IQ setup on a 5 kW Solar panel

The paper is organized into five sections. The first section provides a thorough literature review of relevant studies and existing methodologies. The second subsection describes the Dust Accumulation methodology utilized in this investigation. The third section presents the prediction model fourth section is the mathematical modeling of effect of dust on solar module, fourth section is results of the analysis and offers a thorough analysis of the findings. This permits a comprehensive examination of the effect of dust accumulation on the performance of solar modules. The last section concludes the paper by summarizing its most important findings and contributions.

## 2 Methodology of Work

### 2.1 Collection of Data

For the study analyses of soiling loss from the Dust IQ soil monitoring system installed on the rooftop of Utilization lab, New Delhi, the solar module is first clean Fig. 2 represents

the 5 kW solar PV system and the data are been collected from the Dust Explorer software in which the soiling loss and transmission losses are recorded for 6 months from March, 2023 to August 2023 i.e., 180*24 = 4320 data points.

Data are collected at log interval of 60 min. This monitoring system is installed on 5 kW PV system a total of 20 panels are arranged in a layout of 10 × 2, with 2 arrays of 10 panels in each array having solar modules of 250 $W_p$ rating. The location of the monitoring system is in center of the 5 kW PV system as shown in Fig. 2b. It should be aligned with the solar panels in terms of both azimuth and tilt angle. This guarantees that the soiling conditions assessed accurately reflect the circumstances that impact the panels. The weather inputs considered are considered from solcast site [24] which is at every 60 min for 6 months. The dataset includes hourly temperature and wind speed data gathered over six months to evaluate their influence on dust buildup and SPV performance. Inconsistencies such as temperature spikes and wind gusts were rectified by statistical thresholding and interpolation, hence assuring data reliability. This enhanced dataset improves the accuracy of soiled SPV power predictions and establishes a robust basis for performance evaluation.

## 2.2 Dust IQ Soil Monitoring System

To determine the quantity of dust and dirt that is present on PV modules, this soil monitoring system uses an optical sensor technology. This technique involves evaluating the light that is transmitted and reflected from the surface of the module. It provides useful data that can be utilized to improve maintenance schedules and boost the overall performance of solar power systems. This is accomplished by continually measuring the amounts of soiling that are present.

The unique, high-quality soil monitoring system device measures module soiling-induced light loss. The soiling of nearby PV modules causes light loss. This device has industrial data collecting and control system-optimized digital signal processing and interfaces. An intuitive interface with RS-485 Modbus® data transfer from Kipp & Zonen connects PLCs, inverters, digital control devices, and digital data recorders. Measurement of reflected light and local dust calibration are needed to determine light loss with the requisite accuracy. Table 1 lists soil monitoring system specifications for data recording on a 5-kW solar module.

Following are the steps of proposed work.

Step 1: Soiling ratio data collection:
Use the soil monitoring device to collect reliable measurements of the degree of soiling on PV modules. This system uses advance sensing technology to analyze the build-up of dust, dirt, microfibers, and other particulate matter on module surfaces. Maintain frequent monitoring and data collecting to establish a comprehensive understanding of the soiling ratio over time.

**Table 1** Specification of dust IQ soil monitoring system

| Range of transmission loss (TL) | 0–50% |
|---|---|
| Range of soiling ratio (SR) | 100–50% (SR = 100–TL) |
| Transmission loss measurement accuracy | ± 1% and ± 1/10 of reading |
| Ambient working temperature | − 20 to + 60 °C |
| Storage temperature | − 20 to + 80 °C |
| Power (DC) | 12 to 30 VDC, 200 to 70 mA |

Step 2: Soiled solar power calculation:
Calculate the power output produced by the PV system considering climatic characteristics such as irradiance, temperature, windspeed, which is obtained from solcast these variables affect SPV power. And further SR is obtained from soiled monitoring system.
Step 3: Comparing soiled and cleaned power output:
Analyze the calculated power output without taking into account cleaning activities (Soiled power). Compare the uncleaned power output to the power output produced after the solar modules consisting no dust accumulation. Calculate the difference between two power values further used to assess the impact of cleaning on power generating efficiency.
Step 4. Forecasting of soiled SPV power: The deep learning model i.e., stacked LSTM and bi-LSTM is used to estimate soiled SPV. The dataset was preprocessed, standardized, and divided into 80% training and 20% testing. Performance-optimized hyperparameters were used to train the models on soiled SPV power. Stacked LSTM collects hierarchical temporal information using two successive LSTM layers, whereas Bi-LSTM processes data forward and backward for accuracy. Further a k-fold cross-validation is done where the model is verified over numerous subsets of the dataset, confirming that the model's performance is consistent across a variety of data distributions. In addition, data augmentation is performed in order to bring unpredictability into the training data, which resulted in the models being more adaptive to inputs that were not previously encountered.
Step 5. Performance calculation: MSE and $R^2$ scores were used to evaluate model performance, comparing computational efficiency and prediction accuracy. The proposed architecture is described in Fig. 3.

## 2.3 Stacked LSTM

A stacked LSTM architecture is described as an LSTM model with numerous LSTM layers Fig. 3. [26] developed a stacked

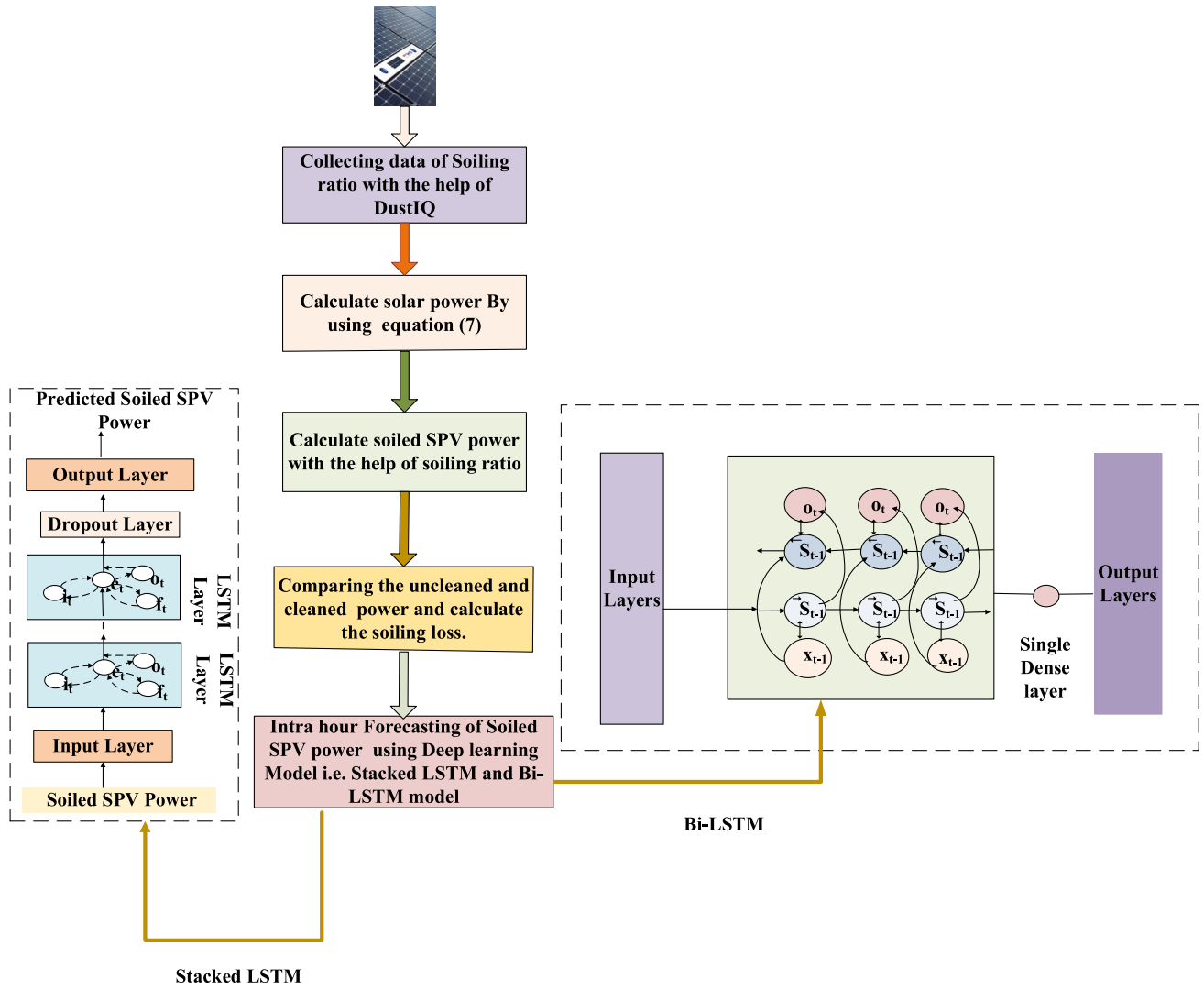



**Fig. 3** Architecture of proposed work

LSTM, sometimes known as deep LSTM. The stacked LSTM model, like the DRNN model, employs numerous LSTM layers that are stacked before being sent to a dropout layer and output layer at the final output. In a stacked LSTM, the first LSTM layer generates sequence vectors that are utilized as input to the succeeding LSTM layer. Furthermore, the LSTM layer gets input from its preceding timestep, enabling for the capture of data patterns. To minimize overfitting, the dropout layer additionally excludes 10% of the neurons. The fundamental structure of LSTM, as shown in the LSTM layer in Fig. 3 consists of an input gate $i_t$, output gate $o_t$, forget gate $f_t$, and memory cell $c_t$. A single LSTM layer features a second-order RNN architecture that excels in storing consecutive short-term memories and retrieving them at a later time. Equation (1–4) are used in stacked LSTM

$$f_t = \sigma\left(W_f.x_t + U_f.h_{t-1} + b_f\right) \tag{1}$$

$$i_t = \sigma\left(W_f.x_t + U_i.h_{t-1} + b_i\right) \tag{2}$$

$$o_t = \sigma(W_o.x_t + U_o.h_{t-1} + b_o) \tag{3}$$

$$c_t = f_t.c_{t-1} + i_t.\sigma(W_c.x_t + U_c.h_{t-1} + b_c) \tag{4}$$

## 2.4 Bi-LSTM

The bidirectional LSTM model detects temporal relationships inherent in data in both directions. This means that, given a vector of inputs, the model can understand the link between past and future samples moving in both the past-to-future and future-to-past directions. Such models have been used in research [25], and the architecture of the bidirectional LSTM layer is shown in Fig. 3 of the proposed work. This

**Table 2** Hyperparameter configuration for model training

| Hyperparameter | Stacked LSTM | Bi-LSTM |
|---|---|---|
| Number of LSTM Layers | 2 | 2 |
| Number of Units per Layer | 64, 32 | 64, 32 |
| Activation Function | ReLU | ReLU |
| Optimizer | Adam | Adam |
| Learning Rate | 0.001 | 0.001 |
| Batch Size | 32 | 32 |
| Dropout Rate | 0.1 | 0.1 |
| Number of Epochs | 50 | 50 |

design uses the same input, output, and one dense layer as the normal model. As previously stated, the bi-directional LSTM layer looks for temporal relationships in the data in two directions, while the traditional LSTM layer only scans in one direction. In the above two discussed Deep learning models are trained with soiled SPV power further performance—optimized hyperparameters are used for training the models as discussed in Table 2.

## 3 Mathematical Modeling for Proposed Work

This section develops a mathematical relation between dust and solar power and helps in calculating soiled power with the help of soiling ratio.

### 3.1 Soiling Ratio

The SR is the ratio of the irradiance used by a module that has been soiled to the irradiance used by a module that has been cleaned in order to create the corresponding short-circuit current or power [26]. The SR is obtained from soil monitoring system representing in Eq. (5).

$$SR = \frac{Actual\ Power\ Output(Ps)}{Expected\ clean\ power(Pc)} \tag{5}$$

$$t_{loss} = 1 - SR \tag{6}$$

### 3.2 Transmission Loss

The transmission loss $t_{loss}$ is represented in Eq. (6), refers to the transmission loss that occurs as a result of soiling. There are many different mathematical models that may be used to predict the performance of PV modules as a function of the amount of irradiance and the temperature.

### 3.3 Modeling of Dust

In this work, an empirical correlation for dust deposition density $\rho_D$ based on an optimal tilt angle $\beta$ is developed. The study team included this association, power equation to compute dust in non-tracking tilting PV modules in Eq. (7). They also used math to examine the dust shield's effect on transmittance during tilt angle changes as in Eq. (11). Thus, they provided a second planned correlation Eqs. (8–14).

$$\rho_D = 680.96 * e^{-0.0469 * SR} \tag{7}$$

$$T_m = T_{amb} + \frac{S_m}{26.9 + 6.2 * w_{in}} \tag{8}$$

$$T_a = T_m - 25^o C \tag{9}$$

$$W_{in} = \frac{{D_m}^{0.2}}{D_a} * w_s \tag{10}$$

$$T = T_{amb}\Big\{ < spanclass =' convertEndash' > 1 - 34 < /span > .37erf\Big[0.17\Big( - 8.5 \times 10^{-3}\rho_{D_0}\beta + 0.82\rho_D\Big)^{0.8473}\Big]\Big\} \tag{11}$$

$$S_{mt} = S_{\text{tilted}}.T \tag{12}$$

$$S_{\text{tilted}} = S_B R_b + S_D(1 - A_i)\left(\frac{1 + \cos\beta}{2}\right)\left[1 + f\sin^3\left(\frac{\beta}{2}\right)\right] + S_D\left(\frac{1 - cos\beta}{2}\right) \tag{13}$$

$$S_m = \frac{S_{mt}}{S_{stc}} \tag{14}$$

$S_B$ is direct radiation, $S_D$ diffuse radiation, the ratio of direct radiation on an inclination to horizontal radiation is $R_b$, $S_{stc}$ is the standard irradiance which is 1000 w/m$^2$, Ai is the anisotropic index, and f is the correction coefficient, using $S_{mt}$ from Eq. (12) calculate accurate solar irradiation considering tilt angle and effect of transmittance T. Density of dust on horizontal surface $\rho_{D_0}$ ($\beta$= 0).

### 3.4 Effect of Soiling on Power

To model PV power, the irradiance is incident on the solar PV module, the exact wind speed on the PV module, and the solar PV module temperature must be estimated. The solar power equation is as follows.

$$P_1 = P_r S_m(1 + K_1\log(S_m) + K_2\log(S_m)^2 + K_3 T_a + K_4 T_a\log(S_m) + K_5 T_a\log(S_m)^2 + K_6 {T_a}^2) \tag{15}$$

$$P_2 = P_1/SR \tag{16}$$

$P_r$ is the rated power, $W_{in}$ is wind speed on solar PV modules, $T_a$ is wind-induced temperature change, $T_m$ is module Temperature, $T_{amb}$ is the ambient temperature, and $K_1 - K_6$ are PV module constants. $D_a$ is the distance between the PV module and the ground surface, and $D_m$ is the distance between the anemometer and the ground surface [27]. $S_{mt}$ here is calculated by Eq. (12) and then the $S_m$ is calculated from Eq. (14) and Ps is calculated from Eq. (15).

Equation (16) gives the cleaned power, and Eq. (15) gives soiled power which is used for further soiled power forecasting. (18) shows the effect of soiling on power the old power equations discussed in [28] do not have soiling parameters Use the new $S_{mt}$ calculated from (15).

The mathematical relation between soiling ratio and module temperature ($T_m$) curve fitting analysis is done shown in Eq. (17). It illustrates an inverse correlation, signifying that when the soiling ratio reduces (with more dust collection), the module temperature is increased. The effect is primarily due to decreased light absorption efficiency, resulting in buildup of heat on the module surface when the SR decreases to 83% owing to dust collection, the module temperature may rise to 36.7 °C, whereas a clean module (SR = 100%) maintains a lower temperature of around 25.23 °C. This correlation helps in understanding the relation between Tm and SR it is inverse relation which shows that lesser the soling ratio which means more dust, temperature is high.

$$SR = 0.0146 * {T_m}^3 - 1.4196 * {T_m}^2 + 44.2150 * T_m - 346.9506 \quad (17)$$

$$P_s = -0.002 * SR^3 + 0.94 * SR^2 - 79 * SR - 0.002 \quad (18)$$

The correlation between soiled power and soiling ratio is shown in Eq. (18) which is polynomial type of equation. It measures the direct impact of dust collection on PV module power loss. Power production declines nonlinearly when SR drops because less irradiance is available for power generation. Further correlation between dust deposition and module temperature is shown in Fig. 4. This relation shows how the soiling ratio affects the module temperature and further effects the soiled power this soiled power is further forecasted which helps in scheduling cleaning.

The mathematical modeling methodology utilized in this work offers a systematic way for quantifying power losses due by soiling, allowing for a more exact assessment of its influence on PV performance. By combining real-time soiling data, transmission loss measurements, and climatic factors, the model reflects the dynamic link between dust deposition and power loss. This enables a data-driven knowledge of how varied amounts of soiling affect energy production in various environmental circumstances. The insights acquired from this model help in identifying important thresholds were performance declines dramatically, directing the deployment of optimized cleaning schedules to minimize efficiency loss. Furthermore, the use of deep learning algorithms, namely Stacked LSTM and Bi-LSTM, improves prediction capacities by understanding complicated temporal relationships in soiling patterns and power variations. These models use past power generating data and weather parameters to properly predict dirty power output, allowing a proactive approach to PV system repair.

## 4 Results and Discussion

In this work, examination of the effect of soiling ratio, transmission loss on solar power system performance is analyzed. Evaluation of the performance of clean and soiled PV system is done. To evaluate the soiling ratio, a smart explorer software is used. Further the soiled power is mathematically calculated and Prediction is done in order to forecast the soiled power for further evaluation and cleaning schedule. In Sect. 4, 1st part shows the effect of modeled soiled power by accumulated dust on solar module using soil monitoring system-based real-time data and second part shows how sensitivity analysis of rainfall intensity on dust deposition and SR and in last part the comparison of predicted soiled PV power using DL models i.e., Bi-LSTM and stacked LSTM is discussed.

### 4.1 Performance Analysis of Dust on PV System Using Soil Monitoring System

Performance analysis of dust is required to find out the relation between dust and power. Figure 2 also shows the setup for capturing the data which collects data for a month for study aspects. The soiling ratio for 6 months i.e., March, April, May, June, July and August, 2023 is shown in Fig. 5a. The soiling ratio indicates the amount of dust that has accumulated on the solar modules. The soiling ratio progressively reduced during the first 12 days in the month of march, suggesting a decrease in PV system performance owing to dust collection. On the twelfth day, the soiling ratio increased rapidly and dramatically, reaching 100%, indicating no soiling loss. This fast rise implies a natural cleaning of the solar modules, maybe caused by rain. As illustrated in Fig. 5a the soiling ratio steadily increase after the natural cleaning event. During the span of 6 months 6 times natural cleaning occurred, and in the month of July it can be seen that there is maximum accumulation of Dust on PV system when no cleaning action takes place. Figure 5b depicts the transmission loss to further examine the influence of the soiling ratio on solar power output transmittance loss is less when the modules are cleaned as at day 12 natural cleaning of modules

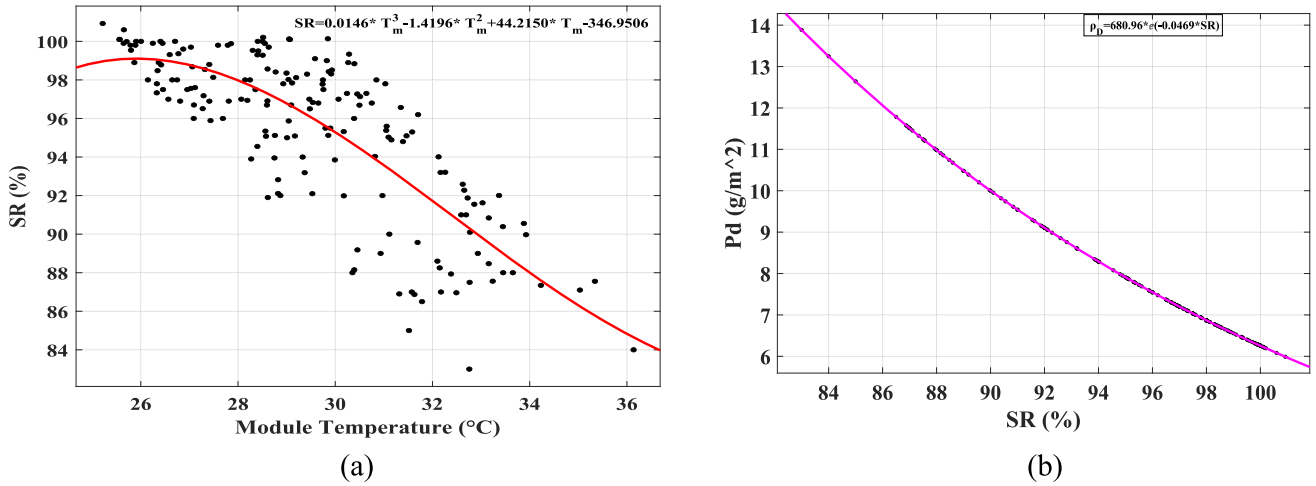



**Fig. 4** Correlation plot of Soiling Ratio with **a** Module temperature **b** dust deposition density

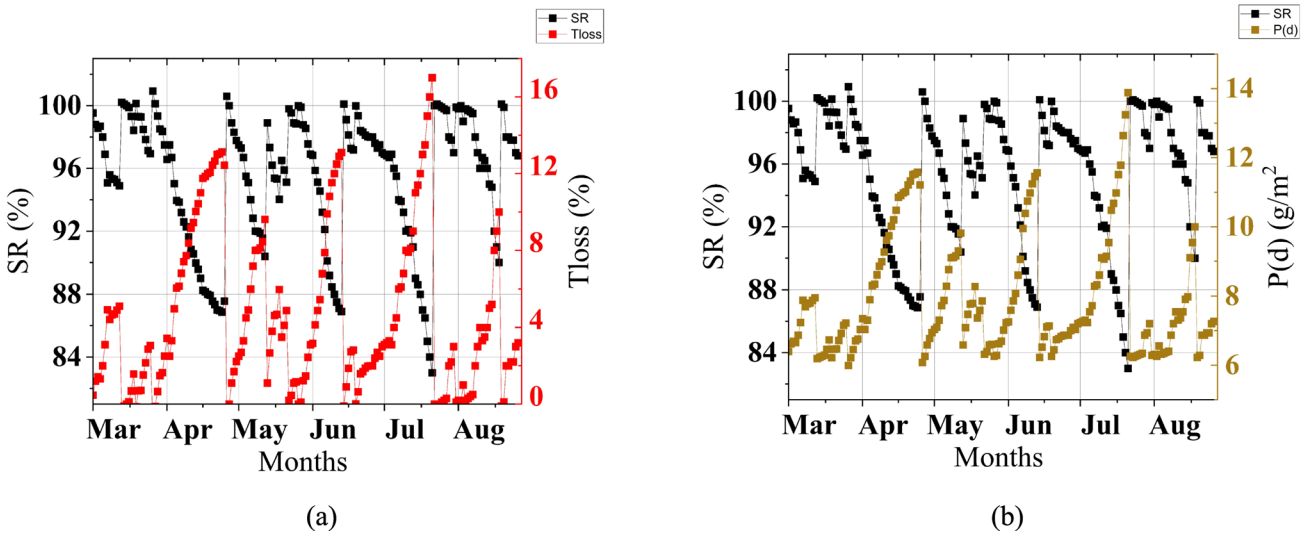



**Fig. 5** Comparison plot of **a** Soiling Ratio and solar PV loss **b** Soiling Ratio and Dust deposition density

occurs so the loss is zero. Transmission loss is the reduction in the quantity of sunlight traveling through the soiled modules, resulting in lower power production. Relation between the soiling ratio and the resulting drop in power production can be learned by analyzing this plot.

In conclusion, the examination of the soiling ratio and its influence on solar power output demonstrates that the modules gradually accumulate dust. Periodic natural cleaning events, on the other hand, aid in the general cleanliness of the modules. The decline in the soiling ratio and the resulting transmission loss highlights the need of routine cleaning and maintenance in order to optimize the performance of solar power systems. (3) is used to determine the cleaned output power, which takes temperature, irradiance, and windspeed into consideration. This estimate takes into account the data received by solcast, a platform that delivers solar irradiance and meteorological data.

It is been observed that when there is slight rainfall cleaning of solar modules are not efficient still there are dust accumulated as it is shown in Fig. 5 on 16 May there is slight rainfall so SR is 98.06% because of which proper cleaning is not done solar power output is less on 30 March and May there is a heavy rainfall hence losses are less with heavy rainfall 0% PV loss and with slight rainfall 1.94% PV loss occurs.

The information was gathered over the course of 6 months from March 1, 2023 to August 31, 2023. After accounting for -environmental variables such as temperature, irradiance, and windspeed, it estimates the power provided by the solar modules. It is crucial to remember that dust collection on solar modules might result in a decrease in output power.

The soiling ratio obtained, measures the level of dust deposition on the modules. The comparison between Soiling ratio and $\rho_D$ is represented in Fig. 5b. Soiled power calculated

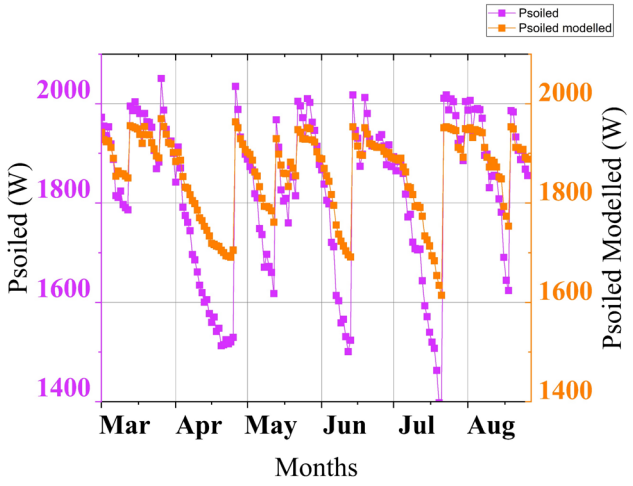



**Fig. 6** Comparison of soiled power plotted from (15) and (18)

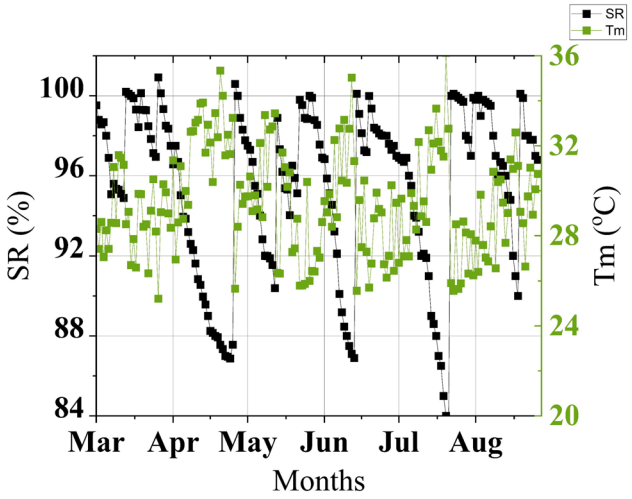



**Fig. 8** Comparison plot of Temperature and soiling ratio

from (12) and mathematically modeled soiled power (18) is shown in Fig. 6.

Soiled power is calculated with the help of cleaned power and soiling ratio. Figure 7a shows a comparison of the power output of soiled modules and cleaned modules. According to Fig. 7a, the highest power loss owing to the presence of dust on the modules is 285.35 W in the month of July. This decrease in power output demonstrates the influence that dust collection has on solar module performance. The soiling ratio is critical in calculating power loss due to dust collection. In order to ensure accurate and effective analysis, it is necessary to evaluate the relationship between $SR$ and $P_{soiled}$ in (18) and then compared with modeled soiled power (12).

Comparison of PV power with Soiling ratio is shown in Fig. 7b as the dust is accumulated the PV power is reduced when there is natural cleaning the PV power is enhanced.

Figure 8 shows the relation of temperature with soiling ratio it can be analyzed that when soiling is more accumulated temperature is increased the maximum temperature is 36.70 ° C where SR is 83% and minimum temperature is 25.23 °C where SR is 100%. Figure 9 shows the comparison of maximum, minimum and average of soiling ratio, mathematical modeled soiled power and soiled power calculated normally using Eq. (12), respectively, which helps in analyzing the impact of soiling and further shows how natural cleaning effects the PV system performance. SR reaches minimum of 83% in the month of July among the six months showing transmission loss of 13% from Fig. 9a.

Maximum power is obtained in the month of March which is 2051.06 W and the least power is obtained in the month of July which is 1393.2 W.

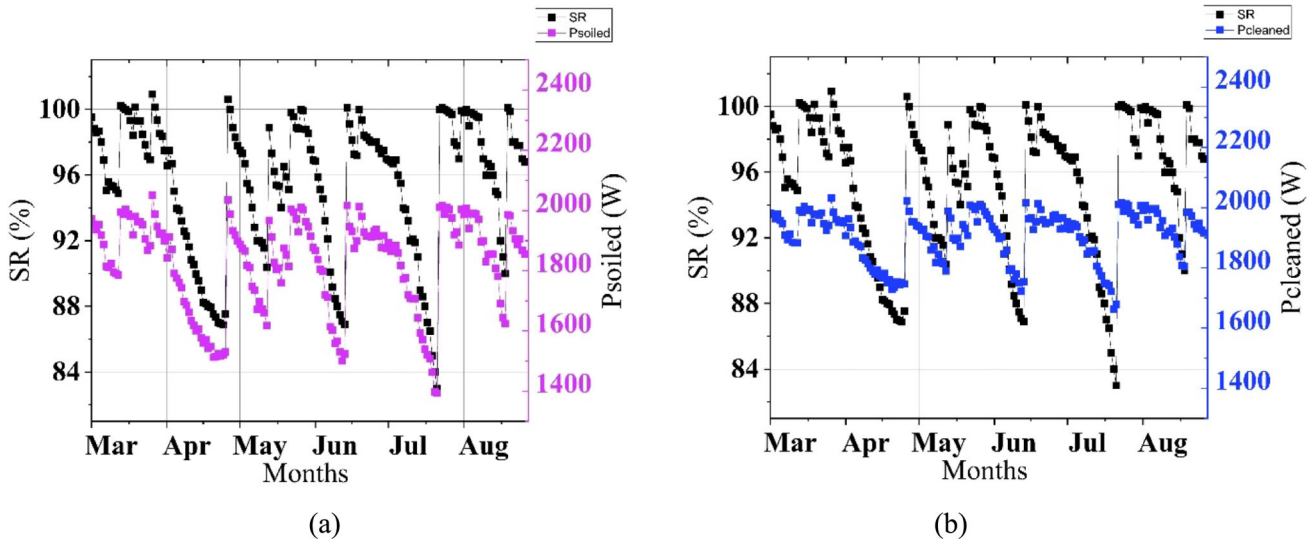



**Fig. 7** Comparison plot of **a** clean and dusty solar modules **b** soiling ratio and PV power

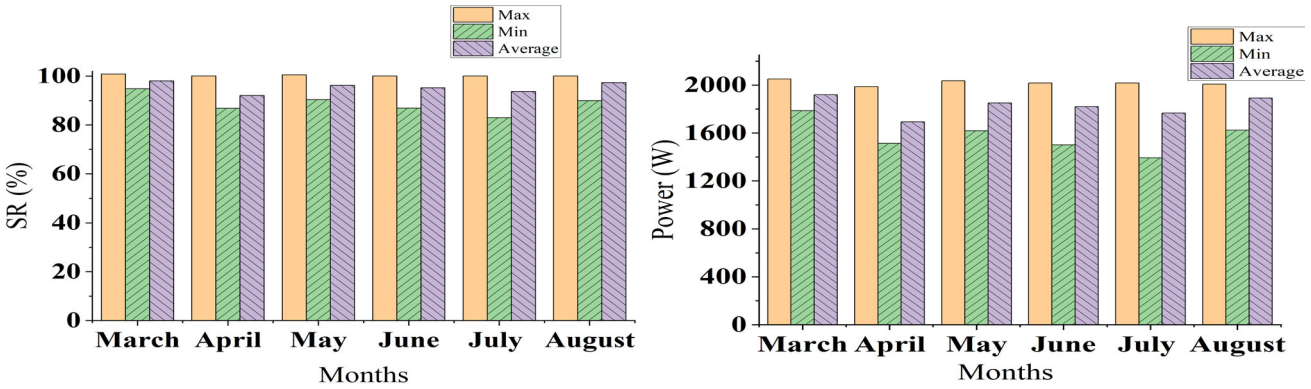



**Fig. 9** Performance analysis of **a** soiling ratio **b** soiled power. Influenced by dust accumulation on PV modules

**Table 3** Sensitivity analysis of rainfall intensity on dust deposition and SR recovery

| Rainfall category | Rainfall amount (mm) | SR recovery (%) | Dust deposition density (P(d), g/m$^2$) before cleaning | Dust deposition density (P(d), g/m$^2$) after cleaning | Remarks |
|---|---|---|---|---|---|
| Light rain | < 1 mm | 96–98 | 9 | 8 | May form mud layers, leading to re-adhesion of dust |
| Moderate rain | 1–5 mm | 90–96 | 10 | 7 | Efficiency depends on duration and wind conditions |
| Heavy rain | > 10 mm | 85–90 | 14 | 6 | Transmission loss reduced to zero, ensuring maximum cleaning efficiency. And SR becomes 100% |

## 4.2 Sensitivity Analysis

Light rain (< 1 mm) may not effectively remove dust off SPV modules owing to weak water impact and poor SPV module surface wetting, resulting in partial cleaning or even mud formation. Moderate rain (1–5 mm) enhances cleaning efficiency but may still leave residual dust, particularly in areas where fine particulate matter prevails. Heavy rain (> 10 mm) has been shown to be the most effective in restoring module cleanliness because it causes more surface runoff, effectively removing dust and debris. Correlating these rainfall categories with soiling ratio improvements allows us to calculate the minimal threshold necessary for successful natural cleaning. Wind may affect dust re-deposition even after a rainstorm event, lowering the total cleaning impact. Humidity may affect dust adhesion; greater humidity levels may enhance cohesiveness between dust particles, making cleanup more difficult; in dry circumstances, dust may be carried by wind rather than washed away [29]. The results highlight the need for alternate cleaning solutions in areas where light rainfall is common yet inefficient. The SR and dust deposition in Fig. 5b are divided in three different cases as shown in table below where it is observed that after light rain, dust is still accumulated there is very less change in dust deposition while after a heavy rainfall the dust is reduced drastically showing maximum cleaning efficiency. There is a moderate change in dust deposition after a moderate rain fall. Table 3 is made with the help of Fig. 5b.

## 4.3 Comparison of Predicted Soiled PV Power

Soiling ratio helps in getting soiled Solar PV power, soiled power data are modeled for every one hour, this soiled power is the input which is given to the two DL model i.e., Stacked LSTM and Bi LSTM for soiled PV power prediction. This prediction of soiled dust will further help in developing cleaning schedules for dusty modules.

This prediction of soiled power helps in getting knowledge about the power reduction of solar PV module because of dust accumulated on solar modules the two DL comparison

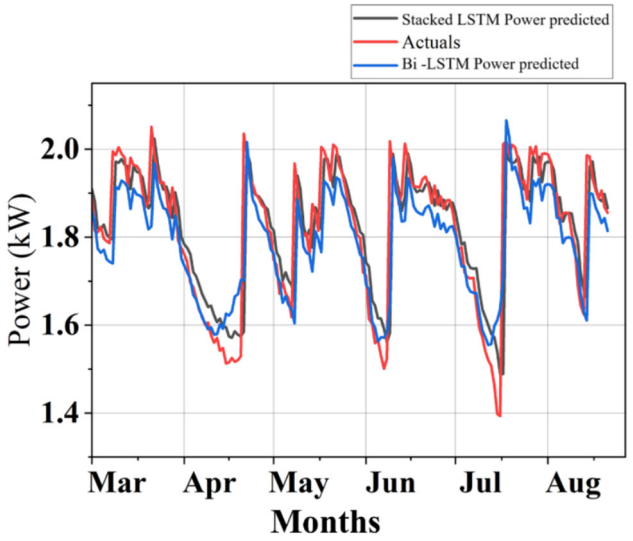



**Fig. 10** Soiled PV power comparison plot of Proposed forecasting models i.e., Stacked LSTM and Bi LSTM

of predicted soil power is shown in Fig. 10. This figure is difficult to comprehend; for greater clarity, it is depicted for two different days: first, on July 26, when maximum dust accumulates and results in a decrease in power output as shown in Fig. 11a; and second, after cleaning occurs with minimal power losses and a significant increase in power output shown in Fig. 11b. The practicality of the two DL models, Stacked LSTM and Bi-LSTM, for accurate predictions in the context of predicting the performance of the models must be evaluated.

Performance estimate is an important aspect in establishing a model's forecast accuracy. Variables used in comparison analysis include root mean square error (RMSE), moving average error (MAE), mean square error (MSE), and $R^2$. MAE helps to analyze the prediction model's constant prediction mistakes, RMSE evaluates the model's overall accuracy, and R2 reflects the strength of the connection between the projected and actual values of the models [30]. Table 4 shows the estimation errors for the DL models.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|A_i - P_i| \tag{19}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(A_i - P_i)^2 \tag{20}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - Pi)^2} \tag{21}$$

$$R^2 = 1 - \frac{\sum_k (A_i - P_i)^2}{\sum_k \left(A_i - \overline{A}\right)^2} \tag{22}$$

where"N"number of windows,"$A_i$"actual value, and Pi is the Predicted value. The arithmetic mean of the actual value is $\overline{A}$

In deep leaning prediction there is an issue of over fitting but this model is not over fitted. To avoid excessive model complexity, L1/L2 regularization is used in this work to implement weight penalties this helps in ensuring that the models did not overfit the training data. Early stopping was used to stop training whenever the testing loss slowed down, eliminating additional training cycles that might damage generalization. Further a dropout layers were utilized to

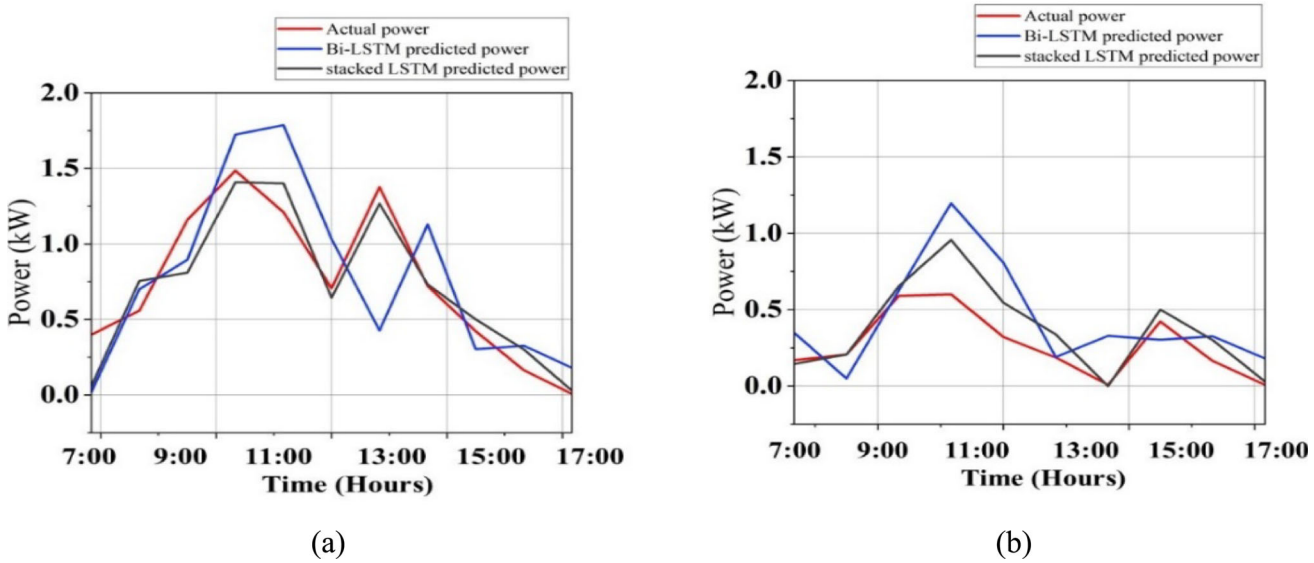



**Fig. 11** Soiled PV power comparison plot of forecasting models at hourly interval i.e., Stacked LSTM and Bi LSTM for **a** Maximum soiled **b** After cleaning

**Table 4** Estimation errors for prediction model

| Models (hourly) | MAE | MSE | RMSE | $R^2$ | Training time (sec) |
|---|---|---|---|---|---|
| Stacked LSTM | 0.051 | 0.0078 | 0.00883 | 0.9913 | 17.35 |
| Bi—LSTM | 0.0588 | 0.0082 | 0.00907 | 0.9850 | 61.70 |

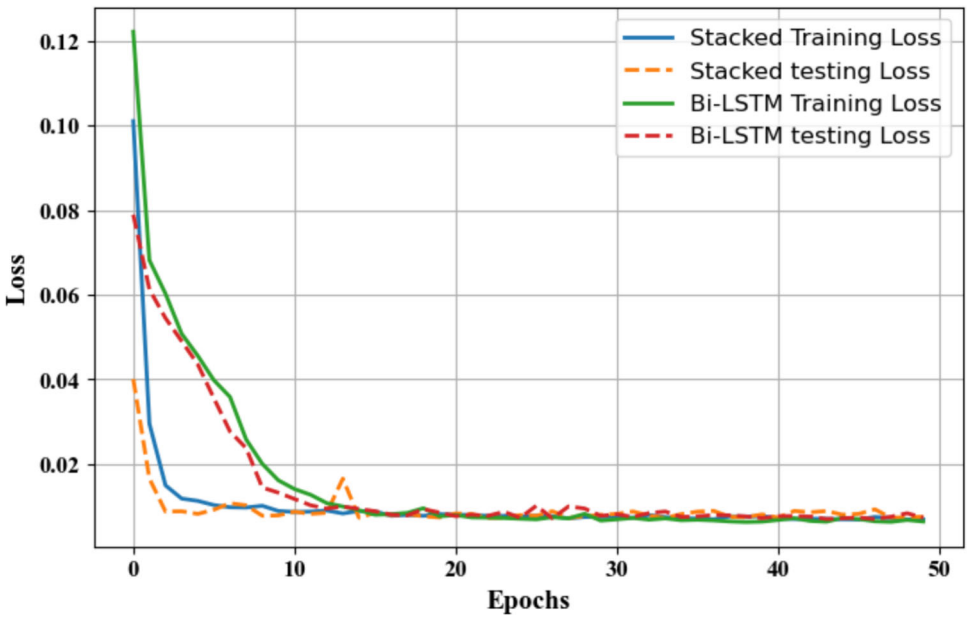



**Fig. 12** Comparison of Training and testing loss function graph for both Stacked LSTM and Bi-LSTM models

randomly deactivate neurons, forcing the model to learn additional features rather than recalling patterns from the training set.

Figure 12 shows the comparison of the training and testing loss curves for the Stacked LSTM and Bi-LSTM models the graph shows their learning efficiency across numerous epochs. Both models' loss values steadily decrease throughout training, demonstrating optimization and convergence. No substantial difference between training and testing loss reduction curves shows the models are not overfitting. Additionally, the Stacked LSTM model consistently has a lower total loss than the Bi-LSTM model, indicating better learning efficiency and generalization.

Further performance metrics calculated from Eqs. (19–22) shown in Table 4, stacked LSTM is the most efficient model according to the obtained result, providing an ideal combination of computational cost and prediction accuracy. While Bi-LSTM improves temporal learning by processing input in both forward and backward directions, it greatly increases computational complexity while providing no significant accuracy benefit. The findings reveal that Stacked LSTM not only trains approximately 3.5 times quicker than Bi-LSTM, but it also has somewhat higher accuracy, making it a more practical option. In real-time applications requiring quick forecasts and frequent model retraining, Stacked LSTM is the recommended model for efficient and dependable forecasting because to its cheap computing cost and good predictive performance. In addition when compared to different studies as: In [31] discusses neural network-based modeling and sensor data to estimate PV system power output when soiled. The results show that soiled PV modules may be predicted with accuracy of 97%. Particle size composition, incident horizontal, and irradiance are used to train the neural network + multi linear regression model to predict soiled power with 98% accuracy [9].

[32] presents a model to correlate soiled solar panel power output, irradiance, and soil particle size composition. Multi linear regression predicts power with 97% accuracy. [8] hybrid clustering for neural network training data preparation. It also uses data division to anticipate soiled solar panel power production with 98% accuracy. The proposed deep learning approach, forecasts soiled power more accurately and effectively.

## 5 Conclusion

Power generated by solar modules is affected by dust and other particulates on their surfaces. These dusts adhere to solar PV panels, causing scratches and corrosion and shortening their lifespan. Solar plant design must consider several

factors, including local conditions and dust characteristics. In order to develop dust-cleaning methods that are both safe and effective for solar modules, it is helpful to first investigate the challenges posed by the accumulation of dust on modules, in addition to the negative impacts. In this paper, the real-time soiling ratio measured from the soil monitoring system helps in modeling the accurate soiled power which helps in further analysis of loss of power in presence of dust and study the impact of temperature on dust accumulation. The maximum loss achieved is 285.35 W. And it will further degrade if continuous dust accumulation occurs, so cleaning of solar modules should be at a prior importance to avoid solar power loss, this paper further helps in prediction of soiled PV power using stacked LSTM and Bi-LSTM models and the trained results shows that stacked LSTM performs much better than Bi- LSTM with 99.13%. Findings show that for SPV performance forecasting, Stacked LSTM surpasses Bi-LSTM in both computational efficiency and accuracy. The results also show that more advanced data-driven maintenance strategies are needed to increase solar energy generation in dust-prone areas. Further this work uses data from one site, which may restrict its application to diverse environmental situations. Additionally, the model does not account for real-time cleaning optimization, which might improve SPV performance, efficiency and generalizability of the model. Future research should use various climate datasets, optimize cleaning schedules, and investigate alternative deep learning models.

**Acknowledgements** The authors would like to thank Anusandhan National Research Foundation (ANRF) to carry out this research under CRG scheme (CRG/2022/003902).

## References

1. Osmani, K.; Haddad, A.; Jaber, H.; Lemenand, T.; Castanier, B.; Ramadan, M.: Mitigating the effects of partial shading on PV system's performance through PV array reconfiguration: a review. Therm. Sci. Eng. Prog. **31**, 101280 (2022). https://doi.org/10.1016/j.tsep.2022.101280
2. Kumar, A.; Alaraj, M.; Rizwan, M.; Alsaidan, I.; Jamil, M.: Development of novel model for the assessment of dust accumulation on solar PV modules. IEEE J. Photovoltaics. **13**, 150–157 (2023). https://doi.org/10.1109/JPHOTOV.2022.3220923
3. Elminir, H.K.; Ghitas, A.E.; Hamid, R.H.; El-Hussainy, F.; Beheary, M.M.; Abdel-Moneim, K.M.: Effect of dust on the transparent cover of solar collectors. Energy Convers. Manag. **47**, 3192–3203 (2006). https://doi.org/10.1016/j.enconman.2006.02.014
4. Niknia, I.; Yaghoubi, M.; Hessami, R.: A novel experimental method to find dust deposition effect on the performance of parabolic trough solar collectors. Int. J. Environ. Stud. **69**, 233–252 (2012). https://doi.org/10.1080/00207233.2012.664810
5. Biryukov, S.: An experimental study of the dry deposition mechanism for airborne dust. J. Aerosol Sci. **29**, 129–139 (1998). https://doi.org/10.1016/S0021-8502(97)00037-2
6. Kaldellis, J.K.; Kokala, A.: Quantifying the decrease of the photovoltaic panels' energy yield due to phenomena of natural air pollution disposal. Energy **35**, 4862–4869 (2010). https://doi.org/10.1016/j.energy.2010.09.002
7. Micheli, L.; Muller, M.: An investigation of the key parameters for predicting PV soiling losses. Prog. Photovoltaics Res. Appl. **25**, 291–307 (2017). https://doi.org/10.1002/pip.2860
8. Pulipaka, S.; Kumar, R.: Power prediction of soiled PV module with neural networks using hybrid data clustering and division techniques. Sol. Energy **133**, 485–500 (2016). https://doi.org/10.1016/j.solener.2016.04.004
9. Pulipaka, S.; Mani, F.; Kumar, R.: Modeling of soiled PV module with neural networks and regression using particle size composition. Sol. Energy **123**, 116–126 (2016). https://doi.org/10.1016/j.solener.2015.11.012
10. Massi Pavan, A.; Mellit, A.; De Pieri, D.; Kalogirou, S. A.: A comparison between BNN and regression polynomial methods for the evaluation of the effect of soiling in large scale photovoltaic plants, Appl. Energy. 108:392–401 (2013). https://doi.org/10.1016/j.apenergy.2013.03.023.
11. Bouzaher, A.; Terki, A.; Bouzaher, M.T.: Photovoltaic panel faults diagnosis: based on the fill factor analysis and use of artificial intelligence techniques. Arab. J. Sci. Eng. **48**, 6471–6487 (2023). https://doi.org/10.1007/s13369-022-07409-w
12. Kazem, H. A.; Al-Waeli, A. H. A.; Chaichan, M. T.; Sopian, K.: Modeling and experimental validation of dust impact on solar cell performance, (2022)https://doi.org/10.1080/15567036.2021.2024922.
13. Alonso-Montesinos, J.; Barbero, J.; Batlles, F. J.; Rodríguez-Martínez, F.; López, G.; Polo, J.; Martín-Chivelet, N.; Alonso, M.; Vela, N.; Marzo, A.; Ferrada, P.; Cortés, M.: Relevance analysis of atmospheric variables in the production of an experimental pv power plant considering dust deposition in the Mediterranean coast, (2019). https://doi.org/10.18086/swc.2019.43.01.
14. Pal, U.; Chandra, S.: ANN based performance forecasting for dusty solar panel. Mater. Today Proc. (2023). https://doi.org/10.1016/j.matpr.2023.05.558
15. Muhammad Ehsan, R.; Simon, S. P.; Venkateswaran, P. R.: Day-ahead forecasting of solar photovoltaic output power using multilayer perceptron, Neural Comput. Appl. 28 3981–3992. (2017) https://doi.org/10.1007/s00521-016-2310-z.
16. P. Nepal, Effect of Soiling on the PV Panel kWh Output, (2018). https://repository.tudelft.nl/islandora/object/uuid%3A59b6a100-4db9-4262-bc5e-d0f2e5c6f57a (accessed May 4, 2023).
17. Costa, S.C.S.; Kazmerski, L.L.; Diniz, A.S.A.C.: Impact of soiling on Si and CdTe PV modules: case study in different Brazil climate zones. Energy Convers. Manag. X. **10**, 100084 (2021). https://doi.org/10.1016/j.ecmx.2021.100084
18. Sayyah, A.; Horenstein, M.N.; Mazumder, M.K.: Energy yield loss caused by dust deposition on photovoltaic panels. Sol. Energy **107**, 576–604 (2014). https://doi.org/10.1016/J.SOLENER.2014.05.030
19. Chatzi, E.N.; Smyth, A.W.: Particle filter scheme with mutation for the estimation of time-invariant parameters in structural health monitoring applications. Struct. Control Heal. Monit. **20**, 1081–1095 (2013). https://doi.org/10.1002/stc.1520
20. Huang, B.; Wang, J.: Applications of physics-informed neural networks in power systems—a review. IEEE Trans. Power Syst. **38**, 572–588 (2023). https://doi.org/10.1109/TPWRS.2022.3162473
21. Atif, M.; Dubey, P.; Aghor, P. P.; López-Marrero, V.; Zhang, T.; Sharfuddin, A.; Yu, K.; Yang, F.; Ladeinde, F.;Liu, Y.; Lin, M.; Li, L.: Fourier neural operators for spatiotemporal dynamics in two-dimensional turbulence, in: SC24-W Work. Int. Conf. High Perform. Comput. Networking, Storage Anal., pp. 41–48, IEEE (2024). https://doi.org/10.1109/SCW63240.2024.00013.
22. Malik, P.; Gehlot, A.; Singh, R.; Gupta, L.R.; Thakur, A.K.: A review on ANN based model for solar radiation and wind speed

prediction with real-time data. Arch. Comput. Methods Eng. **29**, 3183–3201 (2022). https://doi.org/10.1007/s11831-021-09687-3
23. Kayri, I.; Gencoglu, M.T.: Predicting power production from a photovoltaic panel through artificial neural networks using atmospheric indicators. Neural Comput. Appl. **31**, 3573–3586 (2019). https://doi.org/10.1007/s00521-017-3271-6
24. No Title, "solcast".[Online] https://Toolkit.Solcast.Com.Au/Historical/Timeseries/Request. (n.d.). https://toolkit.solcast.com.au/historical/timeseries/request.
25. Singla, P.; Duhan, M.; Saroha, S.: A hybrid solar irradiance forecasting using full wavelet packet decomposition and bi-directional long short-term memory (BiLSTM). Arab. J. Sci. Eng. **47**, 14185–14211 (2022). https://doi.org/10.1007/s13369-022-06655-2
26. Nepal, P.; Korevaar, M.; Ziar, H.; Isabella, O.; Zeman, M.: accurate soiling ratio determination with incident angle modifier for PV modules. IEEE J. Photovoltaics. **9**, 295–301 (2019). https://doi.org/10.1109/JPHOTOV.2018.2882468
27. Alaraj, M.; Kumar, A.; Alsaidan, I.; Rizwan, M.; Jamil, M.: Energy production forecasting from solar photovoltaic plants based on meteorological parameters for Qassim Region, Saudi Arabia. IEEE Access. **9**, 83241–83251 (2021). https://doi.org/10.1109/ACCESS.2021.3087345
28. Kumar, A.; Rizwan, M.; Nangia, U.: A hybrid intelligent approach for solar photovoltaic power forecasting: impact of aerosol data. Arab. J. Sci. Eng. **45**, 1715–1732 (2020). https://doi.org/10.1007/s13369-019-04183-0
29. Bamisile, O.; Acen, C.; Cai, D.; Huang, Q.; Staffell, I.: The environmental factors affecting solar photovoltaic output. Renew. Sustain. Energy Rev. **208**, 115073 (2025). https://doi.org/10.1016/j.rser.2024.115073
30. Rajasundrapandiyanleebanon, T.; Kumaresan, K.; Murugan, S.; Subathra, M.S.P.; Sivakumar, M.: Solar energy forecasting using machine learning and deep learning techniques. Arch. Comput. Methods Eng. **30**, 3059–3079 (2023). https://doi.org/10.1007/s11831-023-09893-1
31. Shapsough, S.; Dhaouadi, R.; Zualkernan, I.: Using linear regression and back propagation neural networks to predict performance of soiled PV modules. Procedia Comput. Sci. **155**, 463–470 (2019). https://doi.org/10.1016/j.procs.2019.08.065
32. Mani, F.; Pulipaka, S.; Kumar, R.: Characterization of power losses of a soiled PV panel in Shekhawati region of India. Sol. Energy **131**, 96–106 (2016). https://doi.org/10.1016/j.solener.2016.02.033

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.

Sensors & Diagnostics

CRITICAL REVIEW

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d5sd00016e

# A review on breast cancer diagnostic techniques

Parikshana Mathur, [a] Saakshi Dhanekar *[abc] and B. D. Malhotra [de]

Received 7th February 2025,
Accepted 1st May 2025

DOI: 10.1039/d5sd00016e

rsc.li/sensors

Breast cancer occurs when cells grow abnormally and form tumors. It is currently one of the most prevalent cancers in women, and it is known to cause serious detrimental effects if not detected on time. Thus, early detection and screening may tremendously contribute to a patient's medical treatment. The boom in cancer diagnostics resulted from the demand to overcome the limitations of bulky and time-consuming conventional detection methods. The new and advanced methods are simpler, faster and easily deployable. This review elucidates various techniques used for breast cancer detection, which include optical, electrochemical, mechanical, electrical, thermal and color- and breath-based methods. An overview of different techniques is presented with additional information related to the available commercial options. This review also presents the integration of artificial intelligence and Internet of Things into futuristic diagnostic techniques. The unmet needs and challenges are also discussed. Overall, this review is a comprehensive package for researchers who want to dive into the advances of breast cancer diagnostics.

## 1. Introduction

Breast cancer (BrC) is a type of tumor that generally occurs in the inner lining of epithelial tissues, milk ducts, and/or lobules. The reason for its occurrence is linked to several factors, such as inheritance of mutations in the genes, past exposure to radiation, and hormonal imbalance. It has now become the foremost cause for cancer-linked illnesses and mortality in females.[1] According to the WHO reports of 2024, it was the most common cancer in women with 670 000 deaths reported worldwide in recent times.[2] Early screening and diagnosis may improve the treatment efficiency. However, majority of the tests carried out in hospital laboratories are time-consuming and expensive.[3] The tests recommended for the detection of BrC at an early stage require professionally trained personnel to operate on bulky equipment, such as those used for ultrasounds, MRI, and mammograms, which are presently not easily affordable by the mass population. Further, the presence of specific BrC biomarkers (*e.g.*, HER2, ER, PR, CA15-3, and CD44) formidably demands precise detection.[4] Owing to these limitations and an urgent need for improvement in the current diagnostics, the research focus has shifted towards the development of detection techniques that are relatively less painful, more accurate, affordable, rapid, and easy to use with a low limit of detection. In addition to this, with the advent of nanotechnology, the use of nanoparticles has become prominent in various BrC detecting methods owing to their small size and unique physiochemical properties.[5]

[a] *Sensekriti Technology Solutions Pvt. Ltd., Technology Innovation and Start-up Centre (TISC), IIT Jodhpur, NH 62, Nagaur Road, Karwar, Jodhpur 342030, Rajasthan, India. E-mail: saakshi@iitj.ac.in*
[b] *Inter Disciplinary Research Division – Smart Healthcare, IIT Jodhpur-AIIMS Jodhpur, NH 62, Nagaur Road, Karwar, Jodhpur 342030, Rajasthan, India*
[c] *Department of Electrical Engineering, IIT Jodhpur, NH 62, Nagaur Road, Karwar, Jodhpur 342030, Rajasthan, India*
[d] *Environmental Sciences & Biomedical Metrology, CSIR-National Physical Laboratory, Dr K.S. Krishnan Road, New Delhi 110012, India*
[e] *Department of Biotechnology, Delhi Technological University, Main Bawana Road, Delhi 110042, India*



**Parikshana Mathur**

*Parikshana Mathur obtained her BSc from JNVU, Jodhpur, and MSc and PhD from IIS deemed to be University, Jaipur. She secured a gold medal for her Master's and has trained at DRDO, Jodhpur. She has served as an Assistant Professor at Central University of Rajasthan in the Department of Biotechnology. Later, she worked as a Research Scientist at Sensekriti Technology Solutions Pvt. Ltd., Technology Innovation and Start-up Centre (TISC), IIT Jodhpur. Currently she is working as a Postdoctoral Fellow at Durban University of Technology, South Africa. She has published several research papers in reputed journals and 9 book chapters. Her past work involves plant-microbe interaction, endophytes, metabolomics and nanoparticle synthesis. Her current and future research interests include target-specific drug delivery and biosensors for societal applications.*

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

This review will take readers through a wide variety of recent techniques that are currently used for BrC diagnostics and their advantages and disadvantages (Table 1). Different aspects of sensing, commercial kits and future perspectives with an essence of AI and IoT methods used for BrC detection are covered.

## 2. Techniques in breast cancer (BrC) screening and diagnosis

### 2.1 Color-based techniques

**2.1.1 Lateral flow immunoassay (LFIA).** LFIA is a paper-based, qualitative, labelled, cost-effective, point-of-care diagnostic technique.[21] It is an immunochromatographic assay that facilitates the separation, capture, and detection of the target analyte. The intensity of colored lines on the LFA strip is directly related to the specific amount of analytes received and reacted upon.[22] The design of LFIA structure uses a strip (nitrocellulose membrane), primary and secondary antibodies and labels for the detection of target analytes.

Conventionally, gold nanoparticles have been used as labels. However, considering their cost, other materials such as graphenes have also been explored.[23] The most significant advantage of this method is that it provides real-time information, can be read by the naked eye and is extremely rapid in delivering results.

The traditional LFIAs can be used for the naked observation of qualitative or semi-quantitative results for samples having concentration above that of the assay. For quantitative analysis of the targets, the advancements in LFIA techniques have emerged, which are employed with portable optical readers. Ye *et al.* have developed a sandwich LFIA-based HER2 immunosensor, wherein HER2-ECD protein in serum was detected with a limit of detection (LOD) of 1.7 ng $mL^{-1}$, and detection in the range of 1.7 to 400 ng $mL^{-1}$ using a portable sensing method.[24] Another such advancement is fluorescent LFIA presented by Deng *et al.* they have synthesized second near-infrared (NIR-II) $Ag_2Se$ polystyrene beads that can be used as a fluorescent probe in LFIA for fast (within 15 min) and accurate detection of BrC markers (CEA and CA153) with a very low limit of detection (CEA: 0.768 ng $mL^{-1}$, CA153: 1.192 U $mL^{-1}$) and high recoveries (93.7% ± 6.2% to 108.8% ± 4%). This method is suitable for clinical and quantitative detection.[25]

Wang and co-workers created an optofluidic metasurface composed of silicon nanoposts (SNP) for lateral flow-through identification of BrC biomarker *ErbB2*. It consists of a top thin silicon layer, with a coating of graphene oxide (GO) nanosheets functionalized with anti-*ErbB2* (Fig. 1a). The GO coating enhances the loading capacity of anti-*ErbB2* molecules due to the presence of oxygenated groups like –COOH and –CHO, which assist in covalent bond formation with anti-*ERbB2* through EDC-NHS coupling. The analyte-ligand binding is monitored from the shift in the wavelength as observed in the reflectance spectra. The shift was apparently displaced in case of bare SNP to SNP with anti-*ErbB2* (Fig. 1b). The resonance dip shifts to longer wavelengths with the increase in *ErbB2* concentration (Fig. 1c). The dose–response curve and reflectance spectra, as



**Saakshi Dhanekar**

*Saakshi Dhanekar received her PhD degree in Electronics from the Faculty of Engineering, Jamia Millia Islamia, New Delhi, India, in 2012. She has been with the Centre for Applied Research in Electronics (CARE) and the Centre for Biomedical Engineering (CBME), Indian Institute of Technology, New Delhi, from 2013 to 2019. She has been an Associate Professor at the Department of Electrical Engineering, IIT Jodhpur, India, since 2019. She has more than 40 publications, four granted patents, and a few patents at different stages of evaluation in the field of sensors and diagnostics. Her research interests include sensor development using nanostructured silicon, metal oxides, IoT, and MEMS for industry and healthcare applications. She is the co-Founder and the Director of Sensekriti Technologies Solutions Private Ltd. She is actively working on diagnostic systems for the detection of diseases through non-invasive breath tests.*



**B. D. Malhotra**

*Dr B. D. Malhotra received his PhD from the University of Delhi, Delhi in 1980. Dr Malhotra has published more than 250 refereed papers in international journals, has filed 10 patents, and has edited/co-edited books on biosensors, organic nano-composites and polymer electronics. His research papers are cited widely and carry an h-index of 98. Prof. Malhotra has established the internationally recognized Department of Science & Technology Centre on Biomolecular Electronics at the CSIR-National Physical Laboratory, New Delhi, India. He is a recipient of the National Research Development Corporation Award 2005 for the invention of 'blood glucose biochemical analyzer', is a Fellow of the Indian National Science Academy, National Academy of Sciences, India and Academician of Asia Pacific Academy of Materials. His current research activities include nanobiomaterials, biosensors, ordered molecular assemblies, conducting polymers, Langmuir–Blodgett films, self-assembled monolayers, nano-biotechnology, biomedical engineering and biomolecular electronics.*


© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

**Table 1** Advantages and limitations of various techniques used for breast cancer screening

| Technique | Advantage | Limitation | Ref. |
|---|---|---|---|
| LFIA | Cost-effective, point of care | Semi-quantitative with possibility of cross-reactivity | 6 |
| Flow cytometry | High sensitivity with multiparameter analysis | Complex sample preparation | 7 |
| SERS | Highly sensitive and specific | Reproducibility and standardization of the SERS signal | 8 |
| ELISA | Can be observed with naked eye | Less sensitive | 9 |
| Fluorescence-based techniques | Chromosomal aberration at single cell level can be analysed | Provide semi-quantitative results | 9 |
| $^{1}$H-MRS | Detailed soft tissue imaging without ionizing radiation exposure | Time consuming, expensive, lack of anatomical information | 10 |
| EIT | Generates tomographic image using non-ionizing radiation | Poor spatial resolution | 11 |
| Electrochemical techniques | Real-time monitoring, and cost-effective | Less sensitive | 12 |
| Breath-based techniques | Easy sample collection | Time-consuming data processing, less specificity | 13 |
| MEMS | Highly sensitive real time measurement with minimal sample consumption | Lack of standardization, complex fabrication leading to increased cost | 14 |
| PMUTs | Reduced size, enhanced bandwidth and sensitivity, lower power consumption | Lower signal-to-noise ratio | 15 |
| SAW | Real time quantification of cells in 2D and 3D cultures | Analyte range is design limited | 16 |
| QCM | Detect very small (nanogram) mass variations | Low reproducibility | 17 |
| Infrared thermography | Non-invasive, fast imaging time | Highly prone to false-negative and false-positive results depending on size of tumour and body temperature | 18 |
| DOI/OM | Generates 2D and 3D images when combined with other imaging modalities | Less spatial resolution when used alone | 19 |
| Microwave-based technique | Non-ionizing radiation, non-invasive, inexpensive, comfortable | Not available in clinic or hospital | 20 |

LFIA: lateral flow immunoassay; SERS: surface enhanced Raman spectroscopy; ELISA: enzyme-linked immunosorbent assay; $^{1}$H-MRS: proton magnetic resonance spectroscopy; EIT: electrical impedance tomography; MEMS: microelectromechanical systems; PMUTs: piezoelectric micromachined ultrasonic transducers; SAW: surface acoustic waves; QCM: quartz crystal microbalance; DOI/OM: diffuse optical imaging/optical mammography.

shown in Fig. 1d and e, indicate that the *ErbB2* free samples form a very small shift (<1 nm) in their resonance wavelength, whereas *ErbB2*-conjugated samples show a wavelength shift of 6 nm. It can be concluded that the device selectively targets *ErbB2* antigen. The results of the study (Fig. 1f) are coinciding in simulation and experiment for *ErbB2* and anti-*ErbB2* binding. The LOD of the SNP-based biosensor is 0.7 ng mL$^{-1}$ with a sensitivity of 2 nm nM$^{-1}$ for *ErbB2*. It is also suggested that the lateral flow-through feature, along with optical bound states, supports the increase in the sensitivity of molecule-bond changes in refractive index.[26]

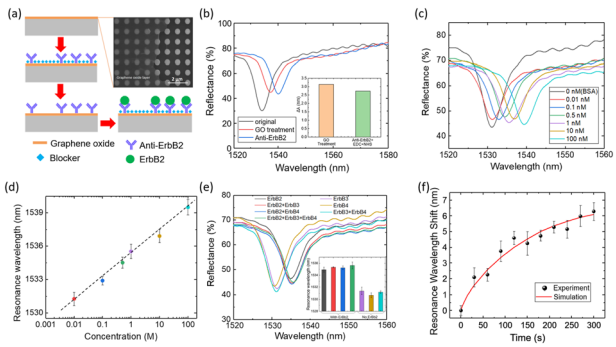



**Fig. 1** (a) Schematic of the label-free assay and SEM of the SNPs. (b) Reflection spectra of the bare sensor, GO layer, and anti-*ErbB2* antibody coating. (c) Reflection spectra in the presence of *ErbB2*. (d) Dose–response curve for the detection of *ErbB2*. (e) Reflection spectra for seven different combinations of *ErbB2*, *ErbB3*, and *ErbB4* antigens. Inset: resonance wavelength in presence of antigens and interfering molecules. (f) Simulation and experimental results for the binding of *ErbB2* and anti-*ErbB2* antibodies. Reproduced from ref. 26 with permission from Elsevier, copyright 2018.

**2.1.2 Flow cytometry.** This lab-based technique measures or counts the cells or particles (in suspension). It comprises a fluidic system and an optical system (lasers), along with a data acquisition system, adjusted with filters and detectors. This system captures fluorescence and scattered signals. The cell sorting and analysis are done on the basis of their fluorescence or light scattering characteristics.[27] This technique is now being commonly used in classifying leukemias and lymphomas. The disproportionate growth and distribution of cells in body, unusual expression of surface receptors, along with tumor antigens, can be characterized by using different flow cytometry-based techniques.[28] Another advancement in flow cytometry has been reported by Bhattacharya *et al.*, in which they employed photoacoustic (PA) flow cytometry to identify circulating BrC cells in human blood. A Q-switched laser, particularly targeting BrC cells having chromophores, is applied to the passing cells and the optically absorbing cells emit an ultrasound wave due to the PA effect. The red curve is characteristic of the presence of only leukocytes, and the blue curve depicts prominent PA

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

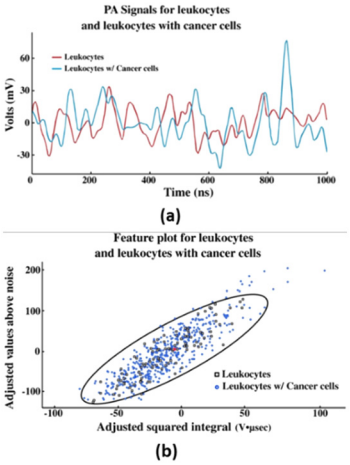



**Fig. 2** (a) PA waveforms obtained from the PAFC test from centrifuged blood. (b) Feature plot of each PA waveform from leukocytes and leukocytes with cancer cells. Reproduced from ref. 29. Published under the terms and conditions of the Creative Commons Attribution 4.0 Unported license.

emission between 850 and 900 ns (Fig. 2a). A plot for PA waveforms obtained during the detection of leukocytes and leukocytes with cancer cells is shown in Fig. 2b. Most of the waveforms are present inside the red boundary and cannot be differentiated from leukocytes. However, towards the top-right, the 17 outliers outside of the red boundary were classified as cancer-containing cells by the automated system. Further, these droplets were captured and imaged to identify separated cancer cells. The features of cells seen at the bottom-left were found to be small and their voltage amount did not present PA signals from cancer cells. Therefore, they are not classified as cancer cells. The study established that the PA flow cytometry can capture single circulating cells with ~43% retrieval rate (25–45 cancer cells per mL blood). The presented method can be employed to govern the disease state of the effected person and their response to the treatment. Further, it can be employed for *in vitro* drug trials and genetic testing.[29]

## 2.2 Optical techniques

**2.2.1 Surface-enhanced Raman spectroscopy (SERS).** SERS is a surface-sensitive, non-destructive vibrational spectroscopy technique that works by identifying the analytes with low concentrations and generating strong electromagnetic fields ("hot spots"), amplifying the scattered signals generated due to the excitation of localized surface plasmons.[30] It presents high sensitivity and specificity and has the ability to detect molecular information from biological samples.[31] Xu *et al.*[30] designed a multiplexed $Fe_3O_4$@CP@SERSnanotag for early and quantitative detection of BrC biomarkers (miRNA: miR-21, miR-155, and let 7b). $Fe_3O_4$@CP@SERSnanotags alone show an increase in Raman signals, whereas, in the presence of biomarkers, the target miRNA specifically binds with capture probe (CP) and forms DNA/miRNA heteroduplexes. The removal of SERS nanotags from $Fe_3O_4$ NPs showed a decrease in the SERS intensity, which is directly related to the amount of target miRNAs. The proposed method presented simultaneous detection of miR-21 (LOD 0.050 fM), miR-155 (LOD 0.063 fM), and let 7b (LOD 0.037 fM).

The authors in ref. 32 report an evaluation of 253 total serum samples from unaffected subjects and patients for differential diagnosis of breast, oral, lung, ovarian, and colorectal cancer. Forty-two patients are diagnosed with BrC using SERS-based detection. The SERS spectra of deproteinized serum were obtained using a 532 nm laser, and the SERS substrates were presented by hydroxylamine-reduced silver nanoparticles (AgNPs). Further estimation was performed using principal component analysis-linear discriminant analysis (PCA-LDA) model for identifying the specific cancer of each sample. The PCA-LDA model depicted 76% efficacy level for BrC. The results indicate that the differential expression of the metabolites in the SERS signal is more profound in breast, ovarian, and oral cancers than that of colorectal and lung cancers. A similar technique has been employed,[33] wherein BrC was diagnosed from the urine sample. For this, urine samples from 53 BrC patients and 22 as control were collected. The SERS spectra were recorded using Ag NPs obtained by reduction with hydroxylamine hydrochloride and further activated with $Ca^{2+}$. This further promoted the adsorption of the anionic purines including uric acid, xanthine and hypoxanthine. By employing PCA-LDA and SERS, the samples from BrC patients were classified with 81% sensitivity and 95% specificity with an overall accuracy of 88%. Both the studies [32,33] suggest the translation of SERS in the clinical setting and emphasize the potential of SERS as a novel and prominent diagnosis strategy for the detection of BrC.

**2.2.2 Enzyme-linked immunosorbent assay (ELISA).** It is the most frequently used immunoassay, which employs enzymatic activity and antibody–enzyme binding as a measure to detect the analytes. The presence of an antigen is detected by the addition of an appropriate substrate, which leads to color change that can be further quantified using a microtiter plate reader. The use of microtiter plate was believed to limit its portability. With the developments in image capturing and processing techniques, it is possible to use ELISA as a point of care device.[34] The paper-based ELISA is an advancement of the conventional ELISA and can be used for diagnostic applications owing to their short analysis time, low cost, and portability, that allows the use of a small quantity of sample. Carneiro *et al.* designed a similar quick, simple and cheaper paper-based colorimetric assay to detect CA15-3 cancer antigen. The linear range of the assay is between 2 and 1100 U $mL^{-1}$ in buffer and 2 and 200 U $mL^{-1}$ in human serum.[35] Choi and group developed an antibody-lectin sandwich test to detect glycosylation of CA15-3 BrC markers. Glycosylation is considered a sensitive indicator of cancer progression. CA15-3 biomarker is a heavily glycosylated molecule, therefore detecting its glycosylation


© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

can be considered a sensitive approach to measure CA15-3 as compared to the conventional ELISA. In this method, monoclonal antibody anti-CA15-3 serves as a capture antibody and binds with CA15-3 in the serum. The detection of CA15-3 glycosylation is done by concanavalin A (ConA) lectin. The CA15-3 glycosylation reportedly increased with the progression of BrC stages. The technique showed proficient discrimination of BrC stage I (69% specificity, 63% sensitivity), IIA (75% specificity, 77% sensitivity), IIB (86% specificity, 69% sensitivity) and III (65% specificity, 80% sensitivity) from BrC.[36]

**2.2.3 Fluorescence-based methods.** The fluorescence-based methods are an advancement of optical methods, which employ fluorescence dyes and analyze the amount of fluorescence emitted from a contrast agent attached with the specific biomarkers. Lee *et al.*[37] demonstrated the use of pH-based fluorescence lifetime for BrC diagnosis by establishing the relation between fluorescence lifetime and pH using fluorescein. The fluorescence lifetime is the mean time of a molecule, spent in the excited state after absorbing a short-pulsed laser energy. Based on the energy exchange between the fluorophore and its environment, it can be used to obtain biochemical characterization of the cells.[38] Since the normal tissues exhibit a neutral pH, whereas cancer tissues show an acidic pH (due to the release of lactic acid through aerobic metabolic processes), it causes differences in the fluorescence lifetime, which can be used to monitor the progression of cancer. The method further presents a non-invasive technique for real-time, *in vivo*, and *in situ* cancer detection using fluorescence lifetime of the target tissue.[37]

Another work[39] depicts the synthesis of a fluorogenic material functionalized with peptides using self-assembly between molybdenum disulfide ($MoS_2$) and a fluorescent peptide probe for fluorescence detection. The developed material presented for the detection of protein C receptor (PROCR), which is a transmembrane biomarker for triple-negative BrC (TNBC) cells and their imaging. The study used TAMRA (tetramethyl-rhodamine, a fluorescent label) to bind with AN33 (peptide), to form TAMRA-AN33 followed by the synthesis of TAMRA-AN33/$MoS_2$ composites. To evaluate the fluorescence imaging ability of the synthesized composite, a TNBC cell line MDA-MB-231 (high PROCR expression) and an MCF-7 BrC cell line (low PROCR expression) were used. It was observed that upon increase in TAMRA-AN33 concentration, the fluorescence in MDA-MB-231 cells increased proportionally, whereas, not much change was observed in the fluorescence of MCF-7 cells (Fig. 3a–c). Similarly, upon incubation of TAMRA-AN33/$MoS_2$, a fluorescence increment was seen in MDA-MB-231 when $MoS_2$ concentration was increased up to 20 μg $mL^{-1}$, but MCF-7 cells remained unaffected (Fig. 3d and e). This indicates that $MoS_2$ enhances the peptide probe binding with the membrane-bound PROCR by clustering effects.[40] Additionally, the cell proliferation assay suggested good compatibility of the peptide probe for pharmacological investigations. The targeting capability of the composite was also investigated and it presented with a strong imaging signature for the BrC cells.

**2.2.4 Proton magnetic resonance spectroscopy ($^1$H-MRS).** The malignant cancer cells show increased proliferative activity, leading to changes in the cell membrane metabolism and subsequently increasing the choline levels that serve as image-based cancer biomarkers.[41] $^1$H-MRS is used to measure the resonance of protons to generate a spectrum, which assists in the identification of BrC biomarkers. A systematic review and metanalysis reports mention the use of tCho as diagnostic biomarkers for BrC detection. The specificity and sensitivity of $^1$H-MRS are found to be 78–80% and 71–74%, respectively.[42] Another systematic review based on 159 studies and 26 papers confirmed that the metabolic concentrations obtained using $^1$H-MRS can be related to the diagnosis and treatment response in BrC patients.[43] MRS-

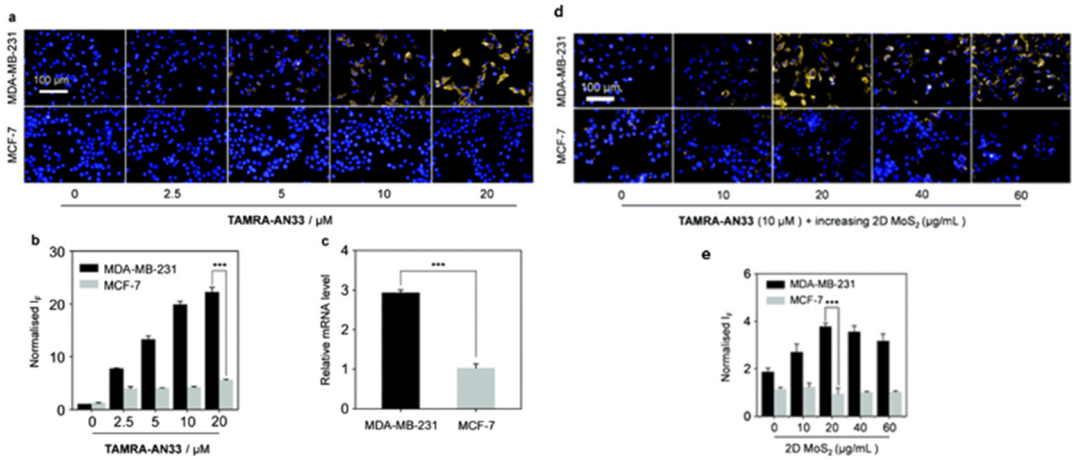



Fig. 3 (a) Fluorescence imaging, (b) quantification of MDA-MB-231 and MCF-7 cell lines treated with TAMRA-AN33, (c) relative mRNA level of PROCR in MDA-MB-231 and MCF-7 cells using RT-qPCR method, (d) fluorescence imaging, and (e) quantification of MDA-MB-231 and MCF-7 treated with TAMRA-AN33 and $MoS_2$. Reproduced from ref. 39. Published under the terms and conditions of the Creative Commons Attribution 3.0 Unported license.

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

based lipid profiling may provide a multivariate approach for BrC characterization. MRS spectra, through the stimulated echo acquisition mode sequence, were obtained from patients. Fat peaks were quantified using the LCModel software. The results presented that the mean concentrations of all lipid metabolites and PUFA were much lower in tumor cells ($p \leq 0.002$) than the normal tissue ($p \leq 0.04$). Similar results were observed in case of HER2-positive *vs.* HER2-negative tumors.[44] A study to establish the diagnostic accuracy of MRS in breast lesions was conducted, in which 159 diagnosed patients underwent $^{1}$H-MRS using a 1.5 Tesla MR system and fast scans in multiple planes were produced. It was concluded that $^{1}$H-MRS is a reliable diagnostic technique for malignant breast lesions and can be employed as an efficient early detection tool for risk-prone age groups.[45]

### 2.3 Electrical techniques

Electrical impedance tomography (EIT) also known as electrical impedance scanning is used to estimate various electrical properties of breast tissue. It is a method where there is no or zero exposure to any form of radiation, and the breast tissue can be safely examined without facing any hazardous conditions.[46] This is completely a non-invasive and user-friendly method with facilities of mobile screening. The mechanism followed by this method is the measurement of impedances using dense electrodes to be placed on the surface of the patient's body. The result is further generated as reconstructed tomographic images, 2D or 3D. Here, impedance mainly involves both conductivity and/or permittivity, and as it changes, the EIT imaging starts its activity. The important mechanisms it follows are data acquisition and image reconstruction.[47]

Lee *et al.*[48] designed and proposed an 8-channel 10 MHz wide bandwidth EIT-AFE-IC having a wide dynamic range, reconfigurable front-end architecture, and a phase compensation loop for high precision detection of BrC. This wide bandwidth EIT system could go up to 10 MHz and portrayed a small phase error of 4.32 degree and could detect an object of 0.5 cm. The results were confirmed using phantom experiments.

A mobile-based EIT demonstrated by Hong *et al.* has a brassiere shape and EIT IC is integrated using a fabric-based board, which consists of an EIT electrode array, electronic circuitry and imaging device.[49] This IC has a chip size of 0.18 μm size 1P6M CMOS process with a consumption of 53.4 mW. The system offers 4.9 mΩ sensitivity and detects a 5 mm cancer mass (0.1% breast volume) within an agar test phantom. The complete time taken for EIT measurement is less than 10 s. in a recent study by Wan *et al.*,[50] HER2 and CA15-3 biomarkers were included into commercially accessible disposable strips, which are functionalized to detect BrC. These strips resemble the widely used glucose detection strips. The findings showed that the limits of detection for these two biomarkers were as low as 1 fg mL$^{-1}$, which is significantly lower than the range of 1–4 ng mL$^{-1}$ for the traditional enzyme-linked immunosorbent test. The detected signal was obtained by applying 10 1.2 ms voltage pulses to the transistor's drain electrode and sensing strip electrode using a synchronized double-pulse method. The average of the 10 digital output readings that corresponded to those 10 voltage pulses was the detected signal. For HER2 and CA15-3, the sensor sensitivity values were attained at about 70/dec and 30/dec, respectively.

### 2.4 Electrochemical techniques

Electrochemical sensors are simple, sensitive, and cost-effective devices for cancer detection. The electrochemical techniques detect the change in electrochemical parameters such as resistivity, current, and redox reactions at the electrode and interpret them for diagnostic results.[51] Zare *et al.*[52] report the synthesis of an electrochemical biosensor showing prominent electrochemical activity, porosity, and dispersity, along with a large surface area for better functionalization. The biosensor is made of a haemoglobin-capped Ag nanocluster conjugated with a Cu(II)-porphyrin-metal organic framework in a graphene-incorporated nanohybrid probe. The Hb-AgNCs@MOF-G nanocomposite immobilizes anti-HER2 on a glassy carbon electrode to detect the presence of HER2-positive BrC cells using electrochemical impedance spectroscopy (EIS) "signal on" and square wave voltammetry (SWV) "signal off" techniques. It reportedly detects 2 cells per mL using EIS "signal on" and 16 cells per mL by SWV "signal off". The sensor demonstrates 15–16 fold selectivity with a recovery efficiency in the range of 94.8 to 106% by the EIS method and 95.4 to 111% *via* the SWV method.

Another multiplexed electrochemical biosensor has been fabricated using screen-printed carbon electrodes functionalized with gold nanoparticles in conjugation with graphene quantum dots and graphene oxide (AuNPs/GQDs/GO) to detect BrC-related microRNAs (miRNA-21, miRNA-155, and miRNA-210). It further involves the use of anthraquinone, methylene blue, and polydopamine as redox indicators, which hybridize with the complementary targets, miRNA-21, miRNA-155, and miRNA-210, respectively. The qualitative and quantitative estimation of each miRNA is identified by its respective peaks and their intensity. The developed biosensor demonstrated efficient sensing of miRNA. It renders a wide linear range of 0.001–1000 pM with low detection limits of 0.04 (miRNA-21), 0.33 (miRNA-155), and 0.28 fM (miRNA-210).[53]

A similar electrochemical sensor has been demonstrated using a GO-IL-AuNP (graphene oxide-ionic liquid-gold nanoparticles) hybrid nanocomposite on a glassy carbon electrode to detect the presence of a BrC biomarker, CD44 antigen. The functionalization of IL with GO helped in tremendously increasing the conductivity of GO. The biocompatibility of IL augments the loading capacity of the anti-CD44 molecule and prohibits the lumping of the


© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Table 2 Electrochemical diagnostics for BrC with parameters included

| Biomarker | Technique | Transducer | LOD | LR | Sensitivity | Ref. |
|---|---|---|---|---|---|---|
| HER2 | Differential pulse voltammetry (DPV) | 8 × SPE | 1.8–2.6 ng $mL^{-1}$ | 0–20 ng $mL^{-1}$ | Validated | 55 |
| | | AuE | 0.995 pg $mL^{-1}$ | 0.01–10 ng $mL^{-1}$, 10–100 ng $mL^{-1}$ | 5.921 μA mL $ng^{-1}$ | 56 |
| | | SPCE | 6.0 ng $mL^{-1}$ | 0–30 ng $mL^{-1}$ | 6 ng $mL^{-1}$ | 57 |
| | | Au@PdAg DBNR/GCE | 0.25 pg $mL^{-1}$ | — | | 58 |
| | Electrochemical impedance spectroscopy (EIS) | CILE | 7.4 ng $mL^{-1}$ | 10–110 ng $mL^{-1}$ | Validated | 59 |
| | | GSPE | 6.0 μg $mL^{-1}$ | 0–40 μg $mL^{-1}$ | — | 60 |
| | | Au/AuNP | 5.0 ng $mL^{-1}$ | $10^{-5}$–$10^{2}$ ng $mL^{-1}$ | — | 61 |
| | | IDμEs | 1 pM $L^{-1}$ | 1 pM–100 nM | 0.035 μF/log([HER2] pM) | 62 |
| | Linear sweep voltammetry (LSV) | SPCE | 4.4 ng $mL^{-1}$ | 15–100 ng $mL^{-1}$ | — | 63 |
| | Amperometry | SPCE | 1.0 μg $mL^{-1}$ | 1–200 μg $mL^{-1}$ | — | 64 |
| CA15-3 | DPV | GCE | 0.015 U $mL^{-1}$ | 2.0–25.0 U $mL^{-1}$, 0.05–2.0 U $mL^{-1}$ | — | 65 |
| | | | 0.012 U $mL^{-1}$ | 0.1–20 U $mL^{-1}$ | — | 66 |
| | | | $5 \times 10^{-6}$ U $mL^{-1}$ | $2 \times 10^{-5}$–40 U $mL^{-1}$ | — | 67 |
| | Cyclic voltammetry (CV) | GCE | 0.040 U $mL^{-1}$ | 0.1–160 U $mL^{-1}$ | — | 68 |
| | | AuE | 0.640 U $mL^{-1}$ | 2.0–240 U $mL^{-1}$ | — | 69 |
| | | | 0.6 ng $mL^{-1}$ | 1.0–240 ng $mL^{-1}$ | | 70 |

8 × SPE: eight screen-printed electrochemical cells, AuE: gold electrode, AuNPs: gold nanoparticles, Au@PdAg DBNR: Au@PdAg dog-bone-like nanorod, CILE: carbon ionic liquid electrode, CV: cyclic voltammetry, DPV: differential pulse voltammetry, EIS: electrochemical impedance spectroscopy, GCE: glassy carbon electrode, GSPE: graphite screen-printed electrode, HER2: human epidermal growth factor receptor, IDμEs: interdigitated gold micro-electrode arrays on silicon/silicon oxide wafers, LOD: limit of detection, LR: linear range, LSV: linear sweep voltammetry, SPCE: screen-printed carbon electrode.

nanocomposite materials improving their stability. The AuNPs in the nanocomposite increase the surface area, which favors the immobilization of anti-CD44 antibodies, and also improve the conductivity of the nanocomposite. All the components result in the synthesis of a highly specific and sensitive device for CD44 detection with a wide range of detection of 5.0 fg $mL^{-1}$ to 50.0 μg $mL^{-1}$ and an LOD of 2.80 fg $mL^{-1}$.[54] Table 2 summarizes a few electrochemical diagnostics for BrC with respect to the biomarkers, techniques and quantifiable parameters.

Bioimpedance is an electrochemical-based technique in which current is passed to the object through electrodes, and the electrical impedance generated gives an impedance score based on its electrical characteristics. The impedance value can be used to differentiate between cancerous (low impedance) and normal tissues (higher impedance).[71] In the study by Mansouri and group,[72] a low-frequency current of 1 kHz, 0.9 mA is introduced through a device to measure the differences in the extracellular resistance (from 0.6 g $l^{-1}$) between the breasts. It is suggested to visit the doctor if the difference is more than 50 Ω. The breast afflicted by the cancer can be determined by looking at the resistance with the lowest value.

### 2.5 Breath-based techniques

BrC can alter the metabolic constitution and synthesis of volatile metabolites in the breath of patients, which can serve as a non-invasive diagnostic marker. The ER, PR and HER2 biomarker receptors exhibit unique metabolomic expression in BrC patients.[73] There are different methods to identify breath VOCs. One is traditional, GC-MS, and the other is advanced, electronic nose. The latter uses an array of sensors and machine learning algorithm to map the breath print to breath VOCs.[74] Types of breath tests for BrC analysis have been presented by Yang *et al.*[75] based on volatile compounds found in the exhaled alveolar breath, with an electronic nose (E-nose) made of an array of carbon nanotube sensors. The volatile organic compounds (VOCs) interact with the array and generate a change in their electric resistance. Further, machine learning was employed to create prediction models for BrC detection and its molecular phenotyping. The breath samples were analyzed using HPPI-TOFMS (high-pressure photon ionization-time-of-flight mass spectrometry). The E-nose system can execute rapid breath biopsy, (in 30 min), which generally requires 7 days. The test reports 86% sensitivity, 97% specificity, and 91% prediction accuracy using the random forest model for 899 volunteers. Another study was conducted on 5047 women based on VOC markers, wherein 465 (9.21%) participants were identified with BrC. It employed similar breathomics and forest algorithm techniques, as reported in the previous study. The results indicated 96.97% detection rates and 87.70% specificity for ductal carcinoma.[76]

The profiling of BrC with 276 subjects using VOCs is portrayed[77] in Fig. 4, wherein the breath analysis was done by GC-MS and nanoarray technology. Twenty-three compounds including alkanes, methylated alkanes, chlorinated alkanes, ethers aldehydes, alcohols, ketones, acetic acid and benzene derivatives were identified as belonging to various molecular sub-types of BrC. These VOCs were found similar to the previously reported results of breast-derived cell lines.[78] Particularly, 2-ethyl-1-hexanol was identified as a potential biomarker for BrC progression.

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

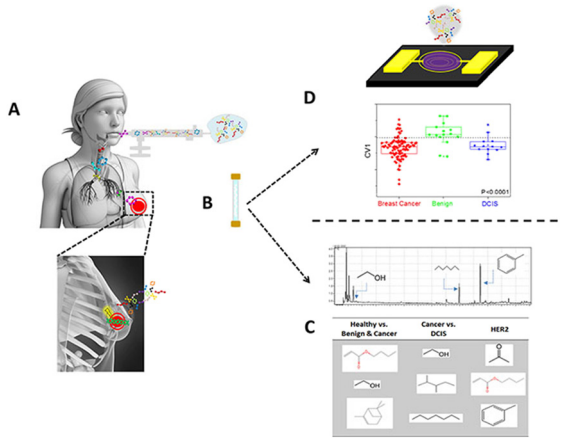



**Fig. 4** (A) Patient exhales into a collection bag. (B) Collection and concentration of breath on Tenax® (C) exposure of the Tenax® tube to GC-MS for detecting the presence of specific compounds. (D) Exposure to the AI nanoarray for volatolomic signature of BrC genetic mutations. Reproduced from ref. 77 published under the terms and conditions of the Creative Commons Attribution license.

Furthermore, certain VOCs were found to be the products of DNA damage caused by the action of reactive oxygen species (ROS) present in tumor cells. ROS also induces the upregulation of cytochrome p-450 enzyme, which further leads to the peroxidation of polyunsaturated fatty acids in the cell membranes. This results in the overexpression of volatile alkanes and alkane-derivatives in the breath.[79] The discriminant function analysis (DFA) of the nanoarray recognized specific differences in the volatolomics between affected and non-affected cells, and different molecular subtypes with 82–87% accuracy, 76–96% specificity and 81–88% sensitivity. Another work presented by Herman-Saffar *et al.* demonstrated that commercial electronic noses and GC-MS can be used to detect initial stages of BrC. They processed exhaled breath samples from 48 BrC patients and followed artificial neural networks for statistical analysis. The model showed up to 95.2% accuracy for the classification of BrC patients.[80] In another study, 443 exhaled breath samples, having 262 diagnosed with BrC, were collected. The results indicated that the breath prints of BrC patients differed from those of healthy women, with a valid classification rate of 98% and a range up to 98.8%. The ROC curve indicates that the sensitivity, specificity, negative predictive value, and positive predictive value were all 100%. This study shows that the electronic nose can distinguish between BrC patients and healthy individuals.[81]

### 2.6 Mechanical techniques

**2.6.1 Microelectromechanical systems (MEMS).** MEMS consist of an electrical and a mechanical component combinedly working as sensors, and can be used to analyze samples at the micro scale.[82] The changes in the biophysical parameters of breast tissues during the progression of tumors can assist as a label-free marker to distinguish tumors from unaffected tissues. On a microscale level, these transformations are imitated as changes in thermal, electrical, and mechanical responses which can render a more objective assessment. One such MEMS-based study[83] presents a conjugation of an electrothermal sensor on a microchip, equipped with a mechatronic actuation system, along with a graphical user interface. The developed sensor provides quantitative information of the thermal conductivity, and also bulk and surface resistivities of the sample. A high value of surface resistivity indicates high surface irregularity and poor conformability, which is present in the cancerous tissue.[84] The caner infected tissues are stiffer than normal tissues, which make them less conformal to the microchip surface. They also have a higher volume of temperature sensitive collagen fibers, which further cause an increase in resistivity.[85,86]

Khosla[87] has presented a novel nano electrode array-based electroimpedance tomography for non-invasive and radiation-free medical imaging. It has been reported as the first device employing a nano-electrode array integrated to a flexible textile sensor fabricated using MEMS. It can scan through a wide range of frequencies and allow for simultaneous scanning of both the breasts. Further, it is reported as the first sensor that parallelly scans lymph nodes present under the arm. These features can assist in tracking cancer progression and treatment.

A RapidET:MEMS-based platform has been evidenced that can quickly demarcate between normal breast and biopsy tissue. The technique is label free and the microchip is based on electrothermal sensing. This MEMS platform contains a platinum microheater, interdigitated electrodes (IDEs), and resistance temperature detectors (RTDs). Fig. 5a depicts the design of the microchip with other elements fabricated on a $SiO_2$ wafer. The optical photograph and SEM are shown in Fig. 5b and c, respectively. A thermal isolation trench of 350 μm depth was designed around the microheater for maintaining a uniform temperature around the microheater. This also reduces the temperature dissipation to other elements such as RTDs. This trench was first simulated in COMSOL to assess its effects.[88] Fig. 5d displays the sample holder that comes into contact with the sensor holder (Fig. 5e). The detailed picture of the sample holder is shown in Fig. 5f, and the enlarged image of microchip is shown in Fig. 5g and h.

The parameters bulk resistivity, surface resistivity and thermal conductivity were measured, and a pathologist confirmed whether the tissue is normal or tumorous. The biopsy tissues were deparaffinized formalin-fixed paraffin-embedded (FFPE) tissues and formalin-fixed fresh tissues. Using this RapidET system, the microheater provided different temperatures to the tissues and both bulk ($\rho_B$) and surface resistivity ($\rho_S$) were measured.

It was observed that both these values were higher for tumor tissues at higher temperatures. A similar trend was found for both formalin fixed and deparaffinized tissue samples. The mean $\rho_B$ was measured to be 4-fold for tumor


© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

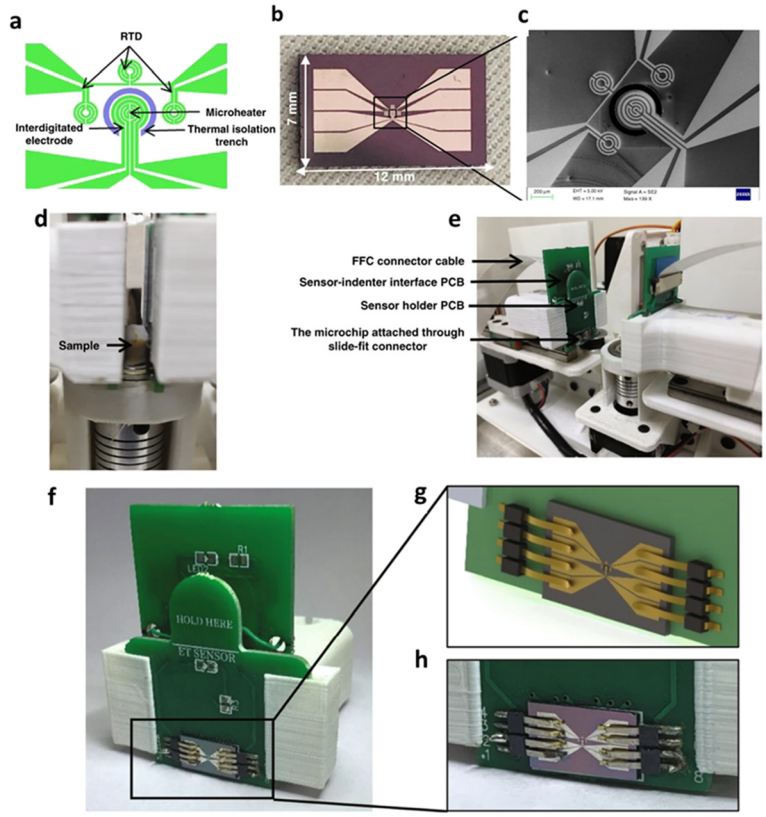



**Fig. 5** (a) Microchip design, (b) optical photograph of the fabricated microchip, (c) SEM image of the active area of the microchip, (d) sample loaded in the system, (e) subsystem with the PCBs, (f) microchip packaged into the sensor holder structure through slide-fit contacts, (g and h) rendered and actual image of the microchip connected to the slide-fit contacts. Reproduced from ref. 88 published under the terms and conditions of Creative Commons Attribution 4.0 International license.

tissues in comparison to the 3-fold change in case of normal tissues. A lower thermal conductivity ($K$) was exhibited by tumor tissues than normal tissues. The mean $K$ of the formalin-fixed tumor tissues was a smaller value of 0.309 ± 0.02 W $m^{-1}$ $K^{-1}$ compared to K of 0.563 ± 0.028 W $m^{-1}$ $K^{-1}$ of normal tissues. Further fisher's combined probability test and regression analysis asserted that these three parameters play a key role in identifying tumor tissues from BrC patients.

Palpreast is a wearable device that is specially designed to imitate the process of breast self-examination in women for the detection of cancer. The device is non-invasive and uses pressure-sensing-based textiles and further enhances confidence and self-awareness among women. The device is based on large- and medium-sized breast phantoms. It mimics the exact function of a human hand. The textile pressure sensor and the inflator system, as shown in Fig. 6, are embedded inside the smart top or fabrics used by women, and it protects the entire breast. To avoid any excessive expansion of the inflator, the smart wearable device consists of an internal pocket. The inflator system has a bulb pump with an air release valve, and a manometer to verify the pressure, along with properly designed T-valve structures to balance the air flux present in each compartment, as a result of which balloons inserted in each compartment layer inter-connected to the pumping system through rubber hoses are inflated.[89]

**2.6.2 Piezoelectric micromachined ultrasonic transducers (PMUTs).** These are extensions of MEMS devices, which assist in *ex vivo* characterization of breasts using single cells, tissue sections, and whole tissues. PMUTs show better sensitivity in differentiating normal cells from cancerous cells by assessing




**Fig. 6** Prototype design of the Palpreast device, including a smart wearable and inflator system. Reproduced from ref. 89 published under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

the vascular density, mechanical stiffness, acoustic impedance, and viscoelastic parameters.[90] Further, such devices are associated with microneedles equipped with tissue-specific targeting for the surgical removal of cancers.[91] A 3D ultrasound imaging ring system using a PMUT linear array has been designed with a circular scanning method for BrC detection. This setup can provide 3D scanning of the breasts along with the identification of the size, location, and outline of the cancerous tissue. The 1 × 128 PMUT linear array is placed at 90° in a cross-vertical manner and the transducer surrounds the mammary gland, facilitating a non-contact detection.[92]

**2.6.3 Surface acoustic waves (SAWs).** They detect targeted cells or tissues in a fluid or solid state by using mass loading on acoustic-wave devices. SA waves are produced at the piezoelectric material surface by applying voltage to the metal interdigital transducer (IDT). Such waves travel through the delay line region to reach the IDT output and cause a mechanical shift, which further creates a voltage difference. Any modification in the device surface by selective coatings can be done to deploy mass loading. The interaction of the functionalized device surface with the entity being sensed creates changes in the resonant frequency, which are further measured to analyze the sample. SAW sensors in combination with complementary metal-oxide semiconductor (CMOS) has been employed to detect hMAM, a BrC biomarker, using streptavidin/biotin-based immunoassay. It can detect BrC biomarkers at low cost, higher sensitivity and enhanced repeatability of fabrication. Chang and co-workers developed another such method for detecting BrC using a novel 2 × 3 model of a leaky surface acoustic wave (LSAW) aptasensor array. The aptamer was assembled on a gold electrode surface of 100 MHz $LiTaO_3$ piezoelectric crystal. The interactions between the aptamer and MUC1 protein expression on the cancer cell surface could successfully capture the MCF-7 cells. The aptamer-cell complex amplified the mass loading capacity of the LSAW sensor.[93] In another study, an ultra-miniaturized microfluidic chip using a narrow-path travelling surface acoustic wave (np-TSAW) has been made for the isolation and purification of BrC cells (MCF-7) from blood with 98% isolation efficiency and 93% purity.[94]

**2.6.4 Quartz crystal microbalance (QCM).** QCM is a piezoelectric sensor in which the mass loaded on the sensor causes a shift in the resonator oscillation frequency. The interactions between the surface immobilized biomolecules can be studied by measuring the frequency shift of the resonator.[95] This phenomenon has been employed to detect the changes in cell surfaces of BrC cells such as over expression of Notch-4 receptors. The techniques involve the development of QCM biosensors conjugated with PHEMA-NPs (poly(2-hydroxyethyl methacrylate) nanoparticles) further functionalized with a Notch-4 receptor antibody. It detected MDA-MB-231 cells with a detection limit of 12 cells per mL.[96] Another such QCM-based biosensor coated with hydroxyethylmethacrylate (HEMA) and ethylene glycol dimethacrylate (EDMA) polymeric nanoparticles and modified with an HER2/neu antibody has been synthesized with a detection limit of 10 cells per mL. SKBR3 cells were taken for the expression study.[97] QCM biosensors are highly sensitive, rapid, reusable, and cost-effective biosensors for cancer cell detection.

### 2.7 Thermal-based technique (infrared thermography)

It is a non-invasive technique that creates a series of infrared scans of the breasts by mapping the temperature difference. The temperature of the tumor tissue is different from that of the non-cancerous tissue, and this difference is generated as a particular thermographic pattern, which can be analyzed for disease characterization. The physiologic image obtained from infrared imaging is in sync with the anatomic image created through mammography, with a very high sensitivity and a negative predictive value. The scans of IR thermography can be additionally used for ensuring the presence of a breast abnormality seen in mammograph.[98] The images are further analyzed by computer algorithms to compare the infrared patterns. It can also assist in differentiating the benign or malignant nature of the lesion.[99] Ekici and Jawzal[100] developed an algorithm for early detection of BrC using image-processing techniques for analyzing thermal breast images. The proposed algorithm is based on convolutional neural networks optimized by Bayes algorithm with an accuracy rate of 98.95%. Likewise, Resmini *et al.*[101] used genetic algorithms and support vector machine for early diagnosis of BrC using infrared thermography with 97.18% of accuracy and 94.79% area under the receiver operating characteristic curve.[86] Another extensive work describes a system that produces thermographic images of breast and further characterizes them as cancerous and non-cancerous. The study suggests that a combination of gradient vector flow and convolutional neural network can detect BrC using the thermographic image characterization.[102]

### 2.8 Imaging technique

It is a lesion detection technique based on functional abnormalities using near-infra-red light, evidenced to detect up to 85% lesions. The breast lesions are identified on the basis of increased hemoglobin concentration and reduced oxygen saturation, which are related to the increased vascularization of tumor cells. DOI techniques include *Transillumination*, which creates 2D images, and diffuse optical tomography (DOT), which generates 3D images.[103] A digital breast tomosynthesis (DBT)/DOT fusion imaging for initial diagnosis of BrC on 28 women has been explored. The receiver operating characteristic curves showed significant improvement, and the technical success rate was 96.4%.[104] Diffuse optical tomography breast imaging system (DOTBIS) non-invasively measures hemoglobin concentration, which is a potential biomarker for short-term response to neoadjuvant chemotherapy. The study conducted by McGuinness[105] analyzed 7 postmenopausal


© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

women with stage I-III BrC enrolled in pre-surgical studies and performed DOTBIS pre- and post-therapy in the affected and unaffected breast. It was observed that DOTBIS-derived observations are modifiable with pre-surgical targeted therapies. The findings support DOTBIS as a potential imaging tool for assessing neoadjuvant targeted therapies in early-stage BrC. Further, to explore the clinical applicability of the diffuse optical inspection device, a compact multi-wavelength DOT system for breast imaging with a fiber-free parallel-plane structure is developed for acquiring the 3D optical properties of the breast.[106]

### 2.9 Microwave-based technique

Microwave-based diagnostics for BrC offers a pain-free method with less scanning time and cost effectiveness.[107] These can be classified into wearable, RF and imaging (radar and tomography). This technique is a contrast to the electrical method, wherein sensors sense the change in signal passing through waveguides, transmission lines lumped capacitors, *etc.*, against the mass/temperature of the tissues.[108]

A fully textile antenna-based detection system envisioned as "smart bra", which can be worn on the breasts with dimensions 24 × 45 × 0.17 $mm^3$, has been presented.[107] An impedance bandwidth from 1.6 GHz up to 10 GHz at $|S11| \leq -6$ dB (VSWR ≤ 3) and from 1.8 to 2.4 GHz and from 4 up to 10 GHz at $|S11| \leq -10$ dB (VSWR ≤ 2) has been realized. The authors also used machine learning algorithms to classify different scanning states using recorded *S* parameters. In the imaging process, the antenna sends signals to the breasts and receives the backscattered signals.[109] Islam *et al.*[110] developed a small-side slotted antenna for the microwave imaging of the breast tumor. A heterogeneous breast phantom has been synthesized to simulate the human breast, and it contained the same dielectric properties as those of humans. Since the tumor cells have a higher water content than that of the healthy ones, the dielectric value of the tumor cells is also very high. This caused a change in the backscattered signals received from both tumor and healthy cells. The system consisted of a 1 × 9 antenna array with an RF switching system to control the receivers, signal processing and image reconstruction unit. The collected back scattered signals were processed using the iteratively corrected delay and sum imaging algorithm, which reconstructed the image of the breast interior for multiple tumor sensing.

A clinical prototype as a wearable device for breast health monitoring comprising 16 flexible monopole antennas embedded in a bra has been developed.[111] The data received from this prototype were compared with the table-top version. It works on the principle of multistatic radar. A small pulse is generated, amplified and transmitted through a switching matrix to a transmitting antenna. A 16-element wideband antenna array surrounds the breast under test. The back scattered waves received from the breasts after the waves pass through the breasts are received at 15 receiving antennas. A picoscope was employed for collecting and storing data. The wearable prototype has reduced the uncertainties by improvement in the fixed positioning of breast and the collection of stronger signal levels, leading to a high signal-to-noise ratio. Since wearables can be directly worn, they reduce the chances of offset error, which is caused in the table-top version due to an additional medium required.

## 3. Commercially available solutions for cancer detection

### 3.1 Blue box device

The blue box[112] is an electronic nose containing chemical sensors that detect BrC biomarkers present in the urine samples. It is a test at-home BrC detection device. Designed with an integrated artificial intelligence (AI) algorithm along with sophisticated technologies, it is used to trace late-stage BrC. There are in total 6 chemical sensors placed inside the box. Within 30 minutes of sample collection, one can know results on the user's phone through an app. In terms of accuracy, the device produces an excellent result of 95%. It renders non-irradiating, painless, and low-cost diagnosis along with a simple and in-home BrC screening. It offers a sensitivity of 88%, an accuracy of 83% and a specificity of 75%.

### 3.2 Human BrC BRCA1 protein ELISA kit

MBS008497 from MyBioSource[113] is a hassle-free ready-to-use strip plate with an ELISA kit to identify the presence of the BRCA1 target analytes in the biological samples. The kit concentration gradient and positive controls offer a detection range in biological samples containing BRCA1. The kit follows the quantitative sandwich-immunoassay procedure. The MBS008497 kit exploits the BRCA1 antibody-BRCA1 antigen interaction (immunosorbent) along with an HRP-conjugated colorimetric detection system to identify BRCA1. It has a detection range as 0.625–20 ng $mL^{-1}$ with a sensitivity of 0.1 ng $mL^{-1}$.

### 3.3 Human cancer antigen CA15-3 ELISA kit

Human cancer antigen CA15-3 ELISA kit (ab108633_abcam)[114] is another product intended to quantitatively measure the concentration of CA15-3 protein in the human serum sample. In patients with breast tumors, there is an elevated amount of CA15-3 (*i.e.* 25% more) concentrations being estimated (the normal concentration of CA15-3 usually ranges up to 30 units per mL). This protein is mostly shredded by the tumor cells into the bloodstream and acts as a disease-specific tumor marker. Using this kit, the minimum detectable concentration of CA15-3 is estimated to be 5 U $mL^{-1}$.

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

### 3.4 HercepTest kits

HercepTest is an FDA-approved semi-quantitative immunohistochemical assay for the recognition of HER2 proteins that are overexpressed in BrC-associated tissues.[115] The kit comprises various reagents needed for the immunohistochemical staining. The HercepTest™ is interpreted as negative for Her2 protein at 0 and 1± score, equivocal at 2+ score, and positive for 3± score.

### 3.5 Xpert® BrC STRAT4

The Xpert® BrC STRAT4[116] test kit provides an accurate, standardized, and reliable biomarker for the qualitative estimation of ER, PR, HER2, and Ki-67 mRNA within 2 hours. This strip follows the semi-quantitative method for the detection of biomarkers in the case of invasive BrC. The method involves robust testing and workflow that does not need any PCR laboratory testing.

## 4. Implantable devices for BrC detection

In recent times, medical devices in the form of *in vivo* sensors and therapeutics have played an important role in the early diagnosis and treatment. The traditional hospital-centered diagnostic process and monitoring is now moving towards patient-centered monitoring. Such type of remote healthcare monitoring needs reliable and affordable solutions. Thus, automated biomedical devices are being developed to address the needs of patients. In this context, implantable devices, which are electronic devices used in medical field, have played a key role. The research is, however, also shifting towards biodegradable implantable devices that degrade harmlessly in the human body and work in the human system without being rejected by the patient's immune response. This, further, rules out the dependency of battery and surgical removal after use which imposes a financial load and severe pain on patients. Triboelectric nanogenerators, which utilize mechanical energy and further transform it into electricity, are also new-age biodegradable implantable devices, showcasing massive potential in cancer therapeutics.[117,118]

An implantable device capable of detecting a BrC biomarker (lactate) has been reported by Gil and his team.[119] The device has demonstrated high sensitivity, a pH less than 0.1 (4.2 mV potential) and 1 mM lactate (70 nA current) level. It can operate at distances of 50 mm (10 μW) as tested inside an anatomical model of the human breast. Such safe and specific devices are highly beneficial for BrC screening and provide real-time access to intrabody tissue monitoring post-surgery. This simultaneously assists in assessing the healing or development of any infection in the tissue.

A microporous poly(ε-caprolactone) (PCL) scaffold has been developed as an implant that can identify metastatic BrC cells at early stage. The number of immune cells at the PCL scaffold, which had stabilized prior to tumor inoculation, altered after tumor inoculation. This scaffold can be used in combination with label-free imaging of early stage metastatic cells, or implanted in patients for monitoring and scheduling follow-up visits using the optical imaging technique.[120]

A multifunctional implantable anticancer device has been fabricated for postsurgical BrC treatment. It is a nanocomposite made of iron oxide nanoparticles, graphene oxide and doxorubicin in a nanofiber matrix that enables target-specific and sustained drug release over 60 days *in vitro*. It has shown significant results in eradicating the local regional recurrence of BrC. Additionally, it reportedly helps in breast reconstruction post-surgery *via* the adipogenic process. The implantable scaffold exhibits enhanced specificity with real-time monitoring and long-term anticancer efficacy.[121]

## 5. Future diagnostics and key challenges for BrC detection

The Internet of Things (IoT) has emerged as a novel and progressive way in the last decade, providing a way for real-time as well as past data analysis. With the use of artificial intelligence (AI), wireless sensor network (WSN), and machine learning (ML), it has advanced rapidly into medical diagnostics. AI and ML tools render enhanced computer-aided diagnosis, medical image processing, interpretation, retrieval, and analysis with add ons of predicting diseases, thus aiding to early diagnosis.[122] Onasanya and Elshakankiri[123] proposed an IoT-enabled system for enhanced detection, and monitoring of malignant cancer using cancer care services and business analytics/cloud services. Using these services, the availability of patient data can be streamlined for further treatment. The use of smart devices and WSN can improve cancer care services by seamless and secure integration of the equipment used in chemotherapy procedures. Moreover, follow-up and monitoring can be shifted into the home through remote monitoring, process automation, and alert communication. In addition to this, WSNs can be employed in achieving better planning of treatment and finalizing appropriate prescription doses. All the systems including pharmacy, medical and radiation oncology servers may allow access to a detailed patient chart that can be enabled from any smart device *via* a remote VPN access from anywhere.

In another such study,[124] an ML-based classifier support vector machine (SVM) has been demonstrated for classifying the malignant and benign BrC by using a recursive feature selection algorithm to include more information from the dataset. The results reveal that the proposed algorithm identifies and selects the best subsets, and the SVM classifier reported optimal classification performance with 99% classification accuracy and specificity and 98% sensitivity. The proposed system is suitable for early detection of BrC with more effective recovery and treatment.


© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Salvi and Kadam[125] also presented an ML-based model, trained through CNN algorithm with 200 000 images. The study proposes that the raspberry pi-integrated thermal imaging sensor can be used by the patient to capture cancer cell images, which is further given as an input to the CNN trained model to identify the status of the image (benign or malignant). In addition, using the Wi-Fi and cloud technology, the captured images are wirelessly sent to the doctor for further examination.

Deep learning systems (DLSs) have been used to predict the presence of ER, PR, and HER2 using the 'patch level model' and hematoxylin-and-eosin-stained images as input.[126] The conventional and commercially existing methods employ costly preparation methods and produce inconstant scoring systems. Moreover, the contradiction in the histology and the expected biomarker results lead to repeated testing. In the study, computer software is trained to identify and classify the image features in order to find ER/PR/HER2 biomarkers in digitized images, instead of tissue stains. ML automatically predicts the status of the biomarkers in pathology images and further testing confirms the specific image features that assist in making the predictions. This method is cost-effective and less time consuming, and provides high-quality control in marker detection.

Despite the high advances made in various fields of cancer diagnostics, certain challenges need to be addressed. The most common of all issues is creating reproducible calibration and separation methods. Besides, identifying the appropriate technology while keeping the highest specificity and sensitivity is another challenge. Further, the device should offer the lowest detection time and detection limit with minimum consumption of energy and cost.[127] Research is also directed on identifying novel biomarkers using a single gene/protein for precise detection. It also includes the use of nanomaterial-based substrates along with a combination of different mechanisms on a common platform, resulting in multiple outputs from a single sensor.

Nanomaterials have a large surface-to-volume ratio that assists in better functionalization of the substrate for the binding of biomolecules, and thus, specific detection of cancer cells. Nanomaterial-based point-of-care devices have shown exceptionally low limit of detection towards the BrC biomarker HER-2.[128–130] Nanorobots are biocompatible, biodegradable, and biomimetic nanostructures that can detect and kill BrC cells. These can perform multiple tasks on a small scale, including targeted drug delivery and localized minimally invasive microsurgery. DNA nanorobots are specifically known to cause lysosomal degradation of BrC proteins, in different pH environments inside the cell and can also be used for imaging purposes through fluorescent tags.[131] Similarly, nanomotors, nanobodies,[132] nanoflowers,[133] nanosubmarines,[134] nanotrains,[135] nanomachines, nanoballons,[136] nanostars,[137] nanocages,[138] *etc.*, are all futuristic programmable DNA-based nanostructures that can be employed for safe imaging and therapeutic purposes. These nanostructures can autonomously navigate throughout the body and are able to find and kill cancerous cells.[139] However, scalable fabrication or synthesis of nanomaterials is still a challenge. Moreover, the nanosafety concern must be developed to handle the unique and specific behavior of nanomaterials and further understand their potential toxicity.[140]

Most methods can identify one analyte at a time due to the limitation of a single biomarker. Therefore, for early diagnosis, simultaneous detection of multi-biomarkers is also necessary.[141] In the case of IoT-based solutions and systems, the issues of confidentiality and privacy are a major concern. Thus, it is imperative to ensure the operational and security challenges, about patient-sensitive information to avoid the breach of patient data. For this, it is necessary to ensure that all the connected devices conform to industry standards and provide necessary information with minimal equipment failure and lesser human intervention. AI, though a boom for medical diagnostics, requires a large amount and span of data sets to train the models. Thus, it will take some time for AI-based techniques to come up to full use till the current devices fetch, store and keep the data for the AI models.

## 6. Conclusion

Despite the conventional techniques still being used today for BrC diagnosis, there is a need for the development of pain-free, low-LOD, higher sensitivity and less-time-consuming methods. This review embraces various techniques (optical, mechanical, electrical, electrochemical, imaging, microwave, color based) and commercial techniques for BrC detection. The commercial options available to a future vision including AI and IoT techniques as a part of overcoming challenges for next-generation diagnostics of BrC are also included in the review. This review has furnished the working principle of techniques and works recently cited in the literature relevant to the above-mentioned methods. It renders all important information for scientists working in the field of BrC diagnostics.

## Data availability

No primary research results, software or code have been included, and no new data were generated or analysed as part of this review.

## Author contributions

Parikshana Mathur: writing – original draft, review and editing, visualization. Saakshi Dhanekar: conceptualization, supervision, writing – review and editing. B. D. Malhotra: writing – review and editing.

## Conflicts of interest

There are no conflicts to declare.

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

## Acknowledgements

The authors thank Shradha Suman Panda, Master's student in Medical Technologies Program in Smart Healthcare, IIT Jodhpur-AIIMS Jodhpur for helping in organizing a few topics for this manuscript.

## References

1 L. Jing, C. Xie, Q. Li, M. Yang, S. Li and H. Li, *et al.*, Electrochemical Biosensors for the Analysis of Breast Cancer Biomarkers: From Design to Application, *Anal. Chem.*, 2022, **94**(1), 269–296, Available from: **https://pubs.acs.org/doi/10.1021/acs.analchem.1c04475**.

2 Breast cancer [Internet], [cited 2024 Jul 15], Available from: **https://www.who.int/news-room/fact-sheets/detail/breast-cancer**.

3 I. Abrao Nemeir, J. Saab, W. Hleihel, A. Errachid, N. Jafferzic-Renault and N. Zine, The Advent of Salivary Breast Cancer Biomarker Detection Using Affinity Sensors, *Sensors*, 2019, **19**(10), 2373.

4 S. Afzal, M. Hassan, S. Ullah, H. Abbas, F. Tawakkal and M. A. Khan, Breast Cancer; Discovery of Novel Diagnostic Biomarkers, Drug Resistance, and Therapeutic Implications, *Front. Mol. Biosci.*, 2022, **9**, 783450.

5 F. Aminolroayaei, S. Shahbazi-Gahrouei, A. Khorasani and D. Shahbazi-Gahrouei, A Review of Imaging Methods and Recent Nanoparticles for Breast Cancer Diagnosis, *Information*, 2023, **15**(1), 10.

6 T. Mahmoudi, M. de la Guardia and B. Baradaran, Lateral flow assays towards point-of-care cancer detection: A review of current progress and future trends, *TrAC, Trends Anal. Chem.*, 2020, **125**, 115842.

7 Z. Yang, X. Mao, M. Zhu, S. Chen, Z. Gao and T. Jiang, *et al.*, Chinese expert consensus on flow cytometric detection of hematological malignant cells in tissue samples, *J. Natl. Cancer Cent.*, 2025, **5**(1), 28–37.

8 L. Vázquez-Iglesias, G. M. Stanfoca Casagrande, D. García-Lojo, L. Ferro Leal, T. A. Ngo and J. Pérez-Juste, *et al.*, SERS sensing for cancer biomarker: Approaches and directions, *Bioact. Mater.*, 2024, **34**, 248–268.

9 M. Sharifi, A. Hasan, F. Attar, A. Taghizadeh and M. Falahati, Development of point-of-care nanobiosensors for breast cancers diagnosis, *Talanta*, 2020, **217**, 121091.

10 A. Pulumati, A. Pulumati, B. S. Dwarakanath, A. Verma and R. V. L. Papineni, Technological advancements in cancer diagnostics: Improvements and limitations, *Cancer Rep.*, 2023, **6**(2), e1764, Available from: **https://onlinelibrary.wiley.com/doi/full/10.1002/cnr2.1764**.

11 S. Mansouri, Y. Alharbi, F. Haddad, S. Chabcoub, A. Alshrouf and A. A. Abd-Elghany, Electrical Impedance Tomography – Recent Applications and Developments, *J. Electr. Bioimpedance*, 2021, **12**(1), 50, Available from: **https://pmc.ncbi.nlm.nih.gov/articles/PMC8667811/**.

12 F. Achi, A. M. Attar and L. A. Ait, Electrochemical nanobiosensors for the detection of cancer biomarkers in real samples: Trends and challenges, *TrAC, Trends Anal. Chem.*, 2024, **170**, 117423.

13 A. Krilaviciute, J. A. Heiss, M. Leja, J. Kupcinskas, H. Haick and H. Brenner, *et al.*, Detection of cancer through exhaled breath: a systematic review, *Oncotarget*, 2015, **6**(36), 38643–38657, Available from: **https://www.oncotarget.com/article/5938/text/**.

14 P. Praveen and S. Lakshmi, *Cancer Detection Using MEMS Technology. 5th International Conference on Circuits, Control, Communication and Computing, I4C 2024*, 2024, 546–553.

15 S. V. Joshi, S. Sadeghpour, N. Kuznetsova, C. Wang and M. Kraft, Flexible micromachined ultrasound transducers (MUTs) for biomedical applications, *Microsyst. Nanoeng.*, 2025, **11**(1), 1–21, Available from: **https://www.nature.com/articles/s41378-024-00783-5**.

16 M. Aleixandre and M. C. Horrillo, Recent Advances in SAW Sensors for Detection of Cancer Biomarkers, *Biosensors*, 2025, **15**(2), 88, Available from: **https://www.mdpi.com/2079-6374/15/2/88/htm**.

17 S. Akgönüllü, E. Özgür and A. Denizli, Quartz Crystal Microbalance-Based Aptasensors for Medical Diagnosis, *Micromachines*, 2022, **13**(9), 1441, Available from: **https://www.mdpi.com/2072-666X/13/9/1441/htm**.

18 P. Jaglan, R. Dass and M. Duhan, Breast Cancer Detection Techniques: Issues and Challenges, *J. Inst. Eng. (India): B*, 2019, **100**(4), 379–386, Available from: **https://link.springer.com/article/10.1007/s40031-019-00391-2**.

19 K. Lee, Optical mammography: Diffuse optical imaging of breast cancer, *World J. Clin. Oncol.*, 2011, **2**(1), 64–72, Available from: **http://www.ncbi.nlm.nih.gov/pubmed/21603315**.

20 A. A. Abdul Halim, A. M. Andrew, M. N. Mohd Yasin, M. A. Abd Rahman, M. Jusoh and V. Veeraperumal, *et al.*, Existing and Emerging Breast Cancer Detection Technologies and Its Challenges: A Review, *Appl. Sci.*, 2021, **11**(22), 10753, Available from: **https://www.mdpi.com/2076-3417/11/22/10753/htm**.

21 K. M. Koczula and A. Gallotta, Lateral flow assays, *Essays Biochem.*, 2016, **60**(1), 111–120.

22 C. Liu, L. Yang, W. Zhang, D. Li, L. Li and H. Wang, *et al.*, A magnetic nanoparticle-based lateral flow immunochromatography assay for the rapid detection of fluoroquinolones in milk, *Eur. Food Res. Technol.*, 2021, **247**(10), 2645–2656, Available from: **https://link.springer.com/article/10.1007/s00217-021-03820-z**.

23 V. Shirshahi, S. N. Tabatabaei, S. Hatamie and R. Saber, Functionalized reduced graphene oxide as a lateral flow immuneassay label for one-step detection of Escherichia coli O157:H7, *J. Pharm. Biomed. Anal.*, 2019, **164**, 104–111.

24 L. Ye, X. Xu, A. Qu, L. Liu, C. Xu and H. Kuang, Quantitative assessment of the breast cancer marker HER2 using a gold nanoparticle-based lateral flow immunoassay, *Nano Res.*, 2024, **17**(6), 5452–5460, Available from: **https://link.springer.com/article/10.1007/s12274-024-6471-2**.


© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

25 K. Deng, Z. L. Yu, X. Hu, J. Liu, X. Hong and G. G. L. Zi, *et al.*, NIR-II fluorescent Ag2Se polystyrene beads in a lateral flow immunoassay to detect biomarkers for breast cancer, *Microchim. Acta*, 2023, **190**(12), 1–9, Available from: **https://link.springer.com/article/10.1007/s00604-023-06039-9**.

26 Y. Wang, M. A. Ali, E. K. C. Chow, L. Dong and M. Lu, An optofluidic metasurface for lateral flow-through detection of breast cancer biomarker, *Biosens. Bioelectron.*, 2018, **107**, 224–229.

27 J. P. Robinson, R. Ostafe, S. N. Iyengar, B. Rajwa and R. Fischer, Flow Cytometry: The Next Revolution, *Cells*, 2023, **12**(14), 1875.

28 H. K. Mishra, Clinical Applications of Flow Cytometry in Cancer Immunotherapies: From Diagnosis to Treatments, in *Signal Transduction Immunohistochemistry*, ed. A. E. Kalyuzhny, Methods in Molecular Biology, Humana, New York, NY, 2023, vol. 2593, DOI: **10.1007/978-1-0716-2811-9_6**.

29 K. Bhattacharyya, B. S. Goldschmidt and J. A. Viator, Detection and capture of breast cancer cells with photoacoustic flow cytometry, *J. Biomed. Opt.*, 2016, **21**(8), 087007, Available from: **https://pmc.ncbi.nlm.nih.gov/articles/PMC5005571/**.

30 W. Xu, Y. Zhang, D. Hou, J. Shen, J. Dong and Z. Gao, *et al.*, Quantitative SERS detection of multiple breast cancer miRNAs based on duplex specific nuclease-mediated signal amplification, *Anal. Methods*, 2023, **15**(24), 2915–2924, Available from: **https://pubs.rsc.org/en/content/articlehtml/2023/ay/d3ay00583f**.

31 L. Vázquez-Iglesias, G. M. Stanfoca Casagrande, D. García-Lojo, L. Ferro Leal, T. A. Ngo and J. Pérez-Juste, *et al.*, SERS sensing for cancer biomarker: Approaches and directions, *Bioact. Mater.*, 2024, **34**, 248–268.

32 V. Moisoiu, A. Stefancu, D. Gulei, R. Boitor, L. Magdo and L. Raduly, *et al.*, SERS-based differential diagnosis between multiple solid malignancies: breast, colorectal, lung, ovarian and oral cancer, *Int. J. Nanomed.*, 2019, **14**, 6165–6178.

33 V. Moisoiu, A. Socaciu, A. Stefancu, S. Iancu, I. Boros and C. Alecsa, *et al.*, Breast Cancer Diagnosis by Surface-Enhanced Raman Scattering (SERS) of Urine, *Appl. Sci.*, 2019, **9**(4), 806.

34 N. Gupta, S. Dhanekar, S. Chattopadhyay, S. S. Panda and H. Singh, An approach towards development of point of care diagnostics using ELISA, Proceedings of IEEE Sensors, 2021.

35 M. C. C. G. Carneiro, L. R. Rodrigues, F. T. C. Moreira and M. G. F. Sales, Paper-based ELISA for fast CA 15–3 detection in point-of-care, *Microchem. J.*, 2022, **181**, 107756.

36 J. W. Choi, B. I. Moon, J. W. Lee, H. J. Kim, Y. Jin and H. J. Kim, Use of CA15-3 for screening breast cancer: An antibody-lectin sandwich assay for detecting glycosylation of CA15-3 in sera, *Oncol. Rep.*, 2018, **40**(1), 145–154, Available from: **https://pubmed.ncbi.nlm.nih.gov/29749490/**.

37 J. Lee, B. Kim, B. Park, Y. Won, S. Y. Kim and S. Lee, Real-time cancer diagnosis of breast cancer using fluorescence lifetime endoscopy based on the pH, *Sci. Rep.*, 2021, **11**(1), 16864.

38 B. Kim, B. Park, S. Lee and Y. Won, GPU accelerated real-time confocal fluorescence lifetime imaging microscopy (FLIM) based on the analog mean-delay (AMD) method, *Biomed. Opt. Express*, 2016, **7**(12), 5055.

39 W. T. Dou, L. F. Liu, J. Gao, Y. Zang, G. R. Chen and R. A. Field, *et al.*, Fluorescence imaging of a potential diagnostic biomarker for breast cancer cells using a peptide-functionalized fluorogenic 2D material, *Chem. Commun.*, 2019, **55**(88), 13235–13238.

40 J. Eisermann, A. Kerth and D. Hinderberger, Dynamic self-assembly of ions with variable size and charge in solution, *RSC Adv.*, 2019, **9**(32), 18627–18640.

41 K. Glunde, Z. M. Bhujwalla and S. M. Ronen, Choline metabolism in malignant transformation, *Nat. Rev. Cancer*, 2011, **11**(12), 835–848.

42 U. Sharma and N. R. Jagannathan, In vivo MR spectroscopy for breast cancer diagnosis, *BJR Open*, 2019, **1**(1), 20180040.

43 A. G. V. Bitencourt, J. Goldberg, K. Pinker and S. B. Thakur, Clinical applications of breast cancer metabolomics using high-resolution magic angle spinning proton magnetic resonance spectroscopy (HRMAS 1H MRS): systematic scoping review, *Metabolomics*, 2019, **15**(11), 148.

44 A. Bitencourt, V. Sevilimedu, E. A. Morris, K. Pinker and S. B. Thakur, Fat Composition Measured by Proton Spectroscopy: A Breast Cancer Tumor Marker?, *Diagnostics*, 2021, **11**(3), 564.

45 M. A. Bilal Ahmadani, S. Bhatty, Z. ul Abideen, M. S. Yaseen, T. Laique and J. Malik, Imaging in Breast Cancer: Use of Magnetic Resonance Spectroscopy, *Cureus*, 2020, **12**(8), e9734.

46 J. C. Gómez-Cortés, J. J. Díaz-Carmona, J. A. Padilla-Medina, A. E. Calderon, A. I. B. Gutiérrez and M. Gutiérrez-López, *et al.*, Electrical Impedance Tomography Technical Contributions for Detection and 3D Geometric Localization of Breast Tumors: A Systematic Review, *Micromachines*, 2022, **13**(4), 496.

47 Q. Dong, Y. Zhang, Q. He, C. Xu and X. Pan, Image reconstruction method for electrical impedance tomography based on RBF and attention mechanism, *Comput. Electr. Eng.*, 2023, **110**, 108826.

48 J. Lee, S. Gweon, K. Lee, S. Um, K. R. Lee and K. Kim, *et al.*, A 9.6 mW/Ch 10 MHz Wide-bandwidth Electrical Impedance Tomography IC with Accurate Phase Compensation for Breast Cancer Detection, in *2020 IEEE Custom Integrated Circuits Conference (CICC)*, IEEE, 2020, pp. 1–4.

49 S. Hong, K. Lee, U. Ha, H. Kim, Y. Lee and Y. Kim, *et al.*, A 4.9 mΩ-sensitivity mobile electrical impedance tomography IC for early breast-cancer detection system, *IEEE J. Solid-State Circuits*, 2015, **50**(1), 245–257.

50 H. H. Wan, H. Zhu, C. C. Chiang, J. S. Li, F. Ren and C. T. Tsai, *et al.*, High sensitivity saliva-based biosensor in detection of breast cancer biomarkers: HER2 and CA15-3, *J. Vac. Sci. Technol., B:Nanotechnol. Microelectron.:Mater., Process., Meas., Phenom.*, 2024, **42**(2), 023202.

51 C. O'Brien, C. K. Khor, S. Ardalan and A. Ignaszak, Multiplex electrochemical sensing platforms for the

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

detection of breast cancer biomarkers, *Front. Med. Technol.*, 2024, **6**, 1360510.

52 A. A. Zare, H. Naderi-Manesh, S. M. Naghib, M. Shamsipur and F. Molaabasi, Label-free electrochemical cancer cell detection leveraging hemoglobin-encapsulated silver nanoclusters and Cu-MOF nanohybrids on a graphene-assisted dual-modal probe, *Sci. Rep.*, 2023, **13**(1), 21980.

53 C. Pothipor, J. Jakmunee, S. Bamrungsap and K. Ounnunkad, An electrochemical biosensor for simultaneous detection of breast cancer clinically related microRNAs based on a gold nanoparticles/graphene quantum dots/graphene oxide film, *Analyst*, 2021, **146**(12), 4000–4009.

54 P. Ranjan, M. Abubakar Sadique, S. Yadav and R. Khan, An Electrochemical Immunosensor Based on Gold-Graphene Oxide Nanocomposites with Ionic Liquid for Detecting the Breast Cancer CD44 Biomarker, *ACS Appl. Mater. Interfaces*, 2022, **14**(18), 20802–20812.

55 H. Ilkhani, A. Ravalli and G. Marrazza, Design of an Affibody-Based Recognition Strategy for Human Epidermal Growth Factor Receptor 2 (HER2) Detection by Electrochemical Biosensors, *Chemosensors*, 2016, **4**(4), 23.

56 M. Emami, M. Shamsipur, R. Saber and R. Irajirad, An electrochemical immunosensor for detection of a breast cancer biomarker based on antiHER2–iron oxide nanoparticle bioconjugates, *Analyst*, 2014, **139**(11), 2858–2866.

57 Q. A. M. Al-Khafaji, M. Harris, S. Tombelli, S. Laschi, A. P. F. Turner and M. Mascini, *et al.*, An Electrochemical Immunoassay for HER2 Detection, *Electroanalysis*, 2012, **24**(4), 735–742.

58 J. X. Zhang, C. L. Lv, C. Tang, A. J. Wang, L. P. Mei and P. Song, *et al.*, Sandwich-type ultrasensitive immunosensing of breast cancer biomarker based on core-shell Au@PdAg dog-bone-like nanostructures and Au@PtRh nanorods, *Sens. Actuators, B*, 2023, **382**, 133497.

59 E. Arkan, R. Saber, Z. Karimi and M. Shamsipur, A novel antibody–antigen based impedimetric immunosensor for low level detection of HER2 in serum samples of breast cancer patients via modification of a gold nanoparticles decorated multiwall carbon nanotube-ionic liquid electrode, *Anal. Chim. Acta*, 2015, **874**, 66–74.

60 A. Ravalli, C. G. da Rocha, H. Yamanaka and G. Marrazza, A label-free electrochemical affisensor for cancer marker detection: The case of HER2, *Bioelectrochemistry*, 2015, **106**, 268–275.

61 L. Chun, S. E. Kim, M. Cho and W. Choe, *et al.*, Electrochemical detection of HER2 using single stranded DNA aptamer modified gold nanoparticles electrode, *Sens. Actuators, B*, 2013, **186**, 446–450.

62 S. K. Arya, P. Zhurauski, P. Jolly, M. R. Batistuti, M. Mulato and P. Estrela, Capacitive aptasensor based on interdigitated electrode for breast cancer detection in undiluted human serum, *Biosens. Bioelectron.*, 2018, **102**, 106–112.

63 R. C. B. Marques, S. Viswanathan, H. P. A. Nouws, C. Delerue-Matos and M. B. González-García, Electrochemical immunosensor for the analysis of the breast cancer biomarker HER2 ECD, *Talanta*, 2014, **129**, 594–599.

64 S. Patris, P. De Pauw, M. Vandeput, J. Huet, P. Van Antwerpen and S. Muyldermans, *et al.*, Nanoimmunoassay onto a screen printed electrode for HER2 breast cancer biomarker determination, *Talanta*, 2014, **130**, 164–170.

65 C. Li, X. Qiu, D. keqin and Z. Hou, Electrochemical co-reduction synthesis of Au/ferrocene–graphene nanocomposites and their application in an electrochemical immunosensor of a breast cancer biomarker. Anal, *Methods*, 2014, **6**(22), 9078–9084.

66 H. Li, J. He, S. Li and A. P. F. Turner, Electrochemical immunosensor with N-doped graphene-modified electrode for label-free detection of the breast cancer biomarker CA 15-3, *Biosens. Bioelectron.*, 2013, **43**, 25–29.

67 S. Ge, X. Jiao and D. Chen, Ultrasensitive electrochemical immunosensor for CA 15-3 using thionine-nanoporous gold–graphene as a platform and horseradish peroxidase-encapsulated liposomes as signal amplification, *Analyst*, 2012, **137**(19), 4440.

68 W. Li, R. Yuan, Y. Chai and S. Chen, Reagentless amperometric cancer antigen 15-3 immunosensor based on enzyme-mediated direct electrochemistry, *Biosens. Bioelectron.*, 2010, **25**(11), 2548–2552.

69 C. Hong, R. Yuan, Y. Chai and Y. Zhuo, Ferrocenyl-doped silica nanoparticles as an immobilized affinity support for electrochemical immunoassay of cancer antigen 15-3, *Anal. Chim. Acta*, 2009, **633**(2), 244–249.

70 Y. Yang, Z. Zhong, H. Liu, T. Zhu, J. Wu and M. Li, *et al.*, Double-Layer Nanogold and Double-Strand DNA-Modified Electrode for Electrochemical Immunoassay of Cancer Antigen 15-3, *Electroanalysis*, 2008, **20**(24), 2621–2628.

71 L. C. Ward and S. Brantlov, Bioimpedance basics and phase angle fundamentals, *Rev. Endocr. Metab. Disord.*, 2023, **24**(3), 381–391, Available from: **https://pubmed.ncbi.nlm.nih.gov/36749540/**.

72 S. Mansouri, T. Alhadidi and M. B. Azouz, Breast Cancer Detection Using Low-Frequency Bioimpedance Device, *Breast Cancer: Targets Ther.*, 2020, **12**, 109–116, Available from: **https://www.dovepress.com/breast-cancer-detection-using-low-frequency-bioimpedance-device-peer-reviewed-fulltext-article-BCTT**.

73 V. Cappelletti, E. Iorio, P. Miodini, M. Silvestri, M. Dugo and M. G. Daidone, Metabolic Footprints and Molecular Subtypes in Breast Cancer, *Dis. Markers*, 2017, **2017**, 1–19.

74 P. Divyashree, H. A. Quadri, P. Dwivedi and S. Dhanekar, Demonstration of Intelligent Sensing by Nanosensors and use of Classification and Regression Models for Electronic Nose Applications, *IEEE Sens. J.*, 2024, **24**(21), 35423–35428.

75 H. Y. Yang, Y. C. Wang, H. Y. Peng and C. H. Huang, Breath biopsy of breast cancer using sensor array signals and machine learning analysis, *Sci. Rep.*, 2021, **11**(1), 1–9, Available from: **https://www.nature.com/articles/s41598-020-80570-0**.


© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

76 J. Liu, H. Chen, Y. Li, Y. Fang, Y. Guo and S. Li, *et al.*, A novel non-invasive exhaled breath biopsy for the diagnosis and screening of breast cancer, *J. Hematol. Oncol.*, 2023, **16**(1), 1–5, Available from: **https://jhoonline.biomedcentral.com/articles/10.1186/s13045-023-01459-9**.

77 O. Barash, W. Zhang, J. M. Halpern, Q. L. Hua, Y. Y. Pan and H. Kayal, *et al.*, Differentiation between genetic mutations of breast cancer by breath volatolomics, *Onco Targets Ther*, 2015, **6**(42), 44864–44876.

78 L. Lavra, A. Catini, A. Ulivieri, R. Capuano, L. Baghernajad Salehi and S. Sciacchitano, *et al.*, Investigation of VOCs associated with different characteristics of breast cancer cells, *Sci. Rep.*, 2015, **5**(1), 13246.

79 C. B. Ambrosone, Oxidants and Antioxidants in Breast Cancer, *Antioxid. Redox Signaling*, 2000, **2**(4), 903–917.

80 O. Herman-Saffar, Z. Boger, S. Libson, D. Lieberman, R. Gonen and Y. Zeiri, Early non-invasive detection of breast cancer using exhaled breath and urine analysis, *Comput. Biol. Med.*, 2018, **96**, 227–232.

81 L. D. De León-Martínez, M. Rodríguez-Aguilar, P. Gorocica-Rosete, C. A. Domínguez-Reyes, V. Martínez-Bustos and J. A. Tenorio-Torres, *et al.*, Identification of profiles of volatile organic compounds in exhaled breath by means of an electronic nose as a proposal for a screening method for breast cancer: a case-control study, *J. Breath Res.*, 2020, **14**(4), 046009, Available from: **https://iopscience.iop.org/article/10.1088/1752-7163/aba83f**.

82 M. J. Madou, *Fundamentals of Microfabrication and Nanotechnology, Three-Volume Set*, 2018, Available from: **https://www.taylorfrancis.com/books/mono/10.1201/9781315274164/fundamentals-microfabrication-nanotechnology-three-volume-set-marc-madou**.

83 A. V. G. K., G. Gogoi, B. Behera, S. Rila, A. Rangarajan and H. J. Pandya, RapidET: a MEMS-based platform for label-free and rapid demarcation of tumors from normal breast biopsy tissues, *Microsyst. Nanoeng.*, 2022, **8**(1), 1.

84 I. Acerbi, L. Cassereau, I. Dean, Q. Shi, A. Au and C. Park, *et al.*, Human breast cancer invasion and aggression correlates with ECM stiffening and immune cell infiltration, *Integr. Biol.*, 2015, **7**(10), 1120–1134.

85 E. Leikina, M. V. Mertts, N. Kuznetsova and S. Leikin, Type I collagen is thermally unstable at body temperature, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**(3), 1314–1318.

86 N. Ignatieva, O. Zakharkina, A. Dadasheva, A. Shekhter, A. Sviridov and V. Lunin, Transformation of the dermal collagen framework under laser heating, *J. Biophotonics*, 2019, **12**(12), e201960024.

87 A. Khosla, MEMS Based Eit Technology for Non-Invasive Breast Cancer Diagsnostics, *ECS Meeting Abstracts*, 2014, **02**(11), 690, Available from: **https://iopscience.iop.org/article/10.1149/MA2014-02/11/690**.

88 A. V. Anil, G. Gogoi, B. Behera, S. Rila, A. Rangarajan and H. J. Pandya, RapidET: a MEMS-based platform for label-free and rapid demarcation of tumors from normal breast biopsy tissues, *Microsyst. Nanoeng.*, 2022, **8**(1), 1–16, Available from: **https://www.nature.com/articles/s41378-021-00337-z**.

89 L. Arcarisi, L. Di Pietro, N. Carbonaro, A. Tognetti, A. Ahluwalia and C. De Maria, Palpreast—A New Wearable Device for Breast Self-Examination, *Appl. Sci.*, 2019, **9**(3), 381.

90 J. Jung, W. Lee, W. Kang, E. Shin, J. Ryu and H. Choi, Review of piezoelectric micromachined ultrasonic transducers and their applications, *J. Micromech. Microeng.*, 2017, **27**(11), 113001.

91 L. Li, J. Zheng, J. Chen, Z. Luo, Y. Su and W. Tang, *et al.*, Flexible Pressure Sensors for Biomedical Applications: From Ex Vivo to In Vivo, *Adv. Mater. Interfaces*, 2020, **7**(17), 2000743, Available from: **https://onlinelibrary.wiley.com/doi/full/10.1002/admi.202000743**.

92 C. Liu, C. Xue, B. Zhang, G. Zhang and C. He, The Application of an Ultrasound Tomography Algorithm in a Novel Ring 3D Ultrasound Imaging System, *Sensors*, 2018, **18**(5), 1332.

93 K. Chang, Y. Pi, W. Lu, F. Wang, F. Pan and F. Li, *et al.*, Label-free and high-sensitive detection of human breast cancer cells by aptamer-based leaky surface acoustic wave biosensor array, *Biosens. Bioelectron.*, 2014, **60**, 318–324.

94 W. Geng, Y. Liu, N. Yu, X. Qiao, M. Ji and Y. Niu, *et al.*, An ultra-compact acoustofluidic device based on the narrow-path travelling surface acoustic wave (np-TSAW) for label-free isolation of living circulating tumor cells, *Anal. Chim. Acta*, 2023, **1255**, 341138.

95 A. Karczmarczyk, K. Haupt and K. H. Feller, Development of a QCM-D biosensor for Ochratoxin A detection in red wine, *Talanta*, 2017, **166**, 193–197.

96 M. Bakhshpour, A. K. Piskin, H. Yavuz and A. Denizli, Quartz crystal microbalance biosensor for label-free MDA MB 231 cancer cell detection via notch-4 receptor, *Talanta*, 2019, **204**, 840–845.

97 M. Yılmaz, M. Bakhshpour, I. Göktürk, A. K. Pişkin and A. Denizli, Quartz Crystal Microbalance (QCM) Based Biosensor Functionalized by HER2/neu Antibody for Breast Cancer Cell Detection, *Chemosensors*, 2021, **9**(4), 80.

98 N. Köü, A. Köü, M. Duran, S. Simavl and N. Turhan, Comparison of standard mammography with digital mammography and digital infrared thermal imaging for breast cancer screening Meme kanseri taramasnda standart mamografi ile dijital mamografi ve dijital infrared termal görüntülemenin karlatrlmas, *J. Turk.-Ger. Gynecol. Assoc.*, 2010, **11**, 152–159.

99 Y. R. Parisky, A. Sardi, R. Hamm, K. Hughes, L. Esserman and S. Rust, *et al.*, Efficacy of Computerized Infrared Imaging Analysis to Evaluate Mammographically Suspicious Lesions, *AJR, Am. J. Roentgenol.*, 2003, **180**(1), 263–269.

100 S. Ekici and H. Jawzal, Breast cancer diagnosis using thermography and convolutional neural networks, *Med. Hypotheses*, 2020, **137**, 109542.

101 R. Resmini, L. Silva, A. S. Araujo, P. Medeiros, D. Muchaluat-Saade and A. Conci, Combining Genetic Algorithms and SVM for Breast Cancer Diagnosis Using Infrared Thermography, *Sensors*, 2021, **21**(14), 4802.

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

102 S. Tello-Mijares, F. Woo and F. Flores, Breast Cancer Identification via Thermography Image Segmentation with a Gradient Vector Flow and a Convolutional Neural Network, *J. Healthc. Eng.*, 2019, **2019**, 1–13.

103 L. V. Wang and H. I. Wu, Definitions of Optical Properties, *Biomedical Optics: Principles and Imaging*, 2007, pp. 343–346, Available from: **https://www.wiley.com/en-us/Biomedical+Optics%3A+Principles+and+Imaging-p-9780471743040**.

104 E. Y. Chae, H. H. Kim, S. Sabir, Y. Kim, H. Kim and S. Yoon, *et al.*, Development of digital breast tomosynthesis and diffuse optical tomography fusion imaging for breast cancer detection, *Sci. Rep.*, 2020, **10**(1), 13127.

105 J. E. McGuinness, M. L. Altoe, A. Marone, L. E. Franks, S. M. Lee and H. K. Kim, *et al.*, Diffuse optical tomography breast imaging measurements are modifiable with pre-surgical targeted and endocrine therapies among women with early stage breast cancer, *Breast Cancer Res. Treat.*, 2021, **189**(1), 297–304, Available from: **https://link.springer.com/article/10.1007/s10549-021-06320-6**.

106 Y. Wang, Y. Wang, S. Li, S. Li, Y. Wang and Q. Yan, *et al.*, Compact fiber-free parallel-plane multi-wavelength diffuse optical tomography system for breast imaging, *Opt. Express*, 2022, **30**(5), 6469–6486, Available from: **https://opg.optica.org/viewmedia.cfm?uri=oe-30-5-6469&seq=0&html=true**.

107 D. N. Elsheakh, R. A. Mohamed, O. M. Fahmy, K. Ezzat and A. R. Eldamak, Complete Breast Cancer Detection and Monitoring System by Using Microwave Textile Based Antenna Sensors, *Biosensors*, 2023, **13**(1), 87.

108 E. C. Fear, P. M. Meaney and M. A. Stuchly, *Microwaves for breast cancer detection? IEEE Potentials*, 2003, 221, pp. 12–18.

109 L. Wang, Microwave Imaging and Sensing Techniques for Breast Cancer Detection, *Micromachines*, 2023, **14**(7), 1462, Available from: **https://www.mdpi.com/2072-666X/14/7/1462/htm**.

110 M. T. Islam, M. Z. Mahmud, M. T. Islam, S. Kibria and M. Samsuzzaman, A Low Cost and Portable Microwave Imaging System for Breast Tumor Detection Using UWB Directional Antenna array, *Sci. Rep.*, 2019, **9**(1), 15491.

111 IEEE Xplore, Full-Text PDF: [Internet], [cited 2024 May 16], Available from: **https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7384493**.

112 Hello - The Blue Box Biomedical Solutions, S.L. [Internet], [cited 2024 May 17], Available from: **https://thebluebox.ai/**.

113 BRCA1 elisa kit | Human Breast Cancer Susceptibility Protein 1 ELISA Kit-AAC37594.1.

114 ab108633 Cancer Antigen CA15-3 Human ELISA Kit Instructions for Use.

115 HercepTest Kits | Agilent [Internet], [cited 2024 May 17], Available from: **https://www.agilent.com/en/product/pharmdx/herceptest-kits**.

116 Xpert® Breast Cancer STRAT4 [Internet], [cited 2024 May 17], Available from: **https://www.cepheid.com/en-GB/tests/oncology-human-genetics/xpert-breast-cancer-strat4.html**.

117 G. Jian, S. Zhu, X. Yuan, S. Fu, N. Yang and C. Yan, *et al.*, Biodegradable triboelectric nanogenerator as a implantable power source for embedded medicine devices, *NPG Asia Mater.*, 2024, **16**(1), 12.

118 S. Solanki, A. K. Gupta, U. Saha, A. V. Krasnoslobodtsev, R. K. Gupta and B. D. Malhotra, Triboelectric Nanogenerator-based smart biomedical sensors for healthcare, *Sustain. Energy Technol. Assessments*, 2023, **57**, 103233.

119 B. Gil, S. Anastasova and G. Z. Yang, Low-powered implantable devices activated by ultrasonic energy transfer for physiological monitoring in soft tissue via functionalized electrochemical electrodes, *Biosens. Bioelectron.*, 2021, **182**, 113175.

120 S. S. Rao, G. G. Bushnell, S. M. Azarin, G. Spicer, B. A. Aguado and J. R. Stoehr, *et al.*, Enhanced Survival with Implantable Scaffolds That Capture Metastatic Breast Cancer Cells In Vivo, *Cancer Res.*, 2016, **76**(18), 5209–5218.

121 A. R. K. Sasikala, A. R. Unnithan, R. G. Thomas, S. W. Ko, Y. Y. Jeong and C. H. Park, *et al.*, Multifaceted Implantable Anticancer Device for Potential Postsurgical Breast Cancer Treatment: A Single Platform for Synergistic Inhibition of Local Regional Breast Cancer Recurrence, Surveillance, and Healthy Breast Reconstruction, *Adv. Funct. Mater.*, 2018, **28**(8), 1704793.

122 R. Jalloul, H. K. Chethan and R. Alkhatib, A Review of Machine Learning Techniques for the Classification and Detection of Breast Cancer from Medical Images, *Diagnostics*, 2023, **13**(14), 2460.

123 A. Onasanya and M. Elshakankiri, Smart integrated IoT healthcare system for cancer care, *Wireless Networks*, 2021, **27**(6), 4297–4312.

124 M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon and W. Zhou, Breast Cancer Detection in the IOT Health Environment Using Modified Recursive Feature Selection, *Wirel. Commun. Mob. Comput.*, 2019, **2019**, 1–19.

125 S. Salvi and A. Kadam, Breast Cancer Detection Using Deep learning and IoT Technologies, *J. Phys.: Conf. Ser.*, 2021, **1831**, 012030.

126 P. Gamble, R. Jaroensri, H. Wang, F. Tan, M. Moran and T. Brown, *et al.*, Determining breast cancer biomarker status and associated morphological features using deep learning, *Commun. Med.*, 2021, **1**(1), 14.

127 M. Sadeghi, S. Sadeghi, S. M. Naghib and H. R. Garshasbi, A Comprehensive Review on Electrochemical Nano Biosensors for Precise Detection of Blood-Based Oncomarkers in Breast Cancer, *Biosensors*, 2023, **13**(4), 481.

128 S. Augustine, P. Kumar and B. D. Malhotra, Amine-Functionalized $MoO_3$ @RGO Nanohybrid-Based Biosensor for Breast Cancer Detection, *ACS Appl. Bio Mater.*, 2019, **2**(12), 5366–5378.

129 S. Augustine, A. G. Joshi, B. K. Yadav, A. Mehta, P. Kumar and V. Renugopalakrishanan, *et al.*, An emerging nanostructured molybdenum trioxide-based biocompatible

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 May 2025. Downloaded on 5/26/2025 6:28:07 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

sensor platform for breast cancer biomarker detection, *MRS Commun.*, 2018, **8**(3), 668–679, Available from: **https://link.springer.com/article/10.1557/mrc.2018.182**.

130 M. A. Ali, K. Mondal, C. Singh, B. Dhar Malhotra and A. Sharma, Anti-epidermal growth factor receptor conjugated mesoporous zinc oxide nanofibers for breast cancer diagnostics, *Nanoscale*, 2015, **7**(16), 7234–7245, Available from: **https://pubs.rsc.org/en/content/articlehtml/2015/nr/c5nr00194c**.

131 R. Arvidsson and S. F. Hansen, Environmental and health risks of nanorobots: an early review, *Environ. Sci.: Nano*, 2020, **7**(10), 2875–2886.

132 H. Bakherad, F. Ghasemi, M. Hosseindokht and H. Zare, Nanobodies; new molecular instruments with special specifications for targeting, diagnosis and treatment of triple-negative breast cancer, *Cancer Cell Int.*, 2022, **22**(1), 245.

133 H. Patel, K. Parekh, L. Fernel Gamarra, J. Bustamante Mamani, A. A. da Hora and A. M. Figueiredo Neto, In vitro evaluation of magnetic fluid hyperthermia therapy on breast cancer cells using monodispersed Mn0.5Zn0.5Fe2O4 nanoflowers, *J. Magn. Magn. Mater.*, 2023, **587**, 171275.

134 M. Pumera, Electrochemically powered self-propelled electrophoretic nanosubmarines, *Nanoscale*, 2010, **2**(9), 1643.

135 Z. Xu, R. Ni and Y. Chen, Targeting breast cancer stem cells by a self-assembled, aptamer-conjugated DNA nanotrain with preloading doxorubicin, *Int. J. Nanomed.*, 2019, **14**, 6831–6842.

136 B. Xie, X. Yang, R. Zhang, J. Guo, Z. Chen and Y. He, Hollow and Porous Fe3C-NC Nanoballoons Nanozymes for Cancer Cell H2O2 Detection, *Sens. Actuators, B*, 2021, **347**, 130597.

137 M. K. Hameed, J. B. M. Parambath, M. T. Gul, A. A. Khan, Y. Park and C. Han, *et al.*, Arylated gold nanostars aided SERS study of breast cancer cells, *Appl. Surf. Sci.*, 2022, **583**, 152504.

138 P. Ji, X. Wang, J. Yin, Y. Mou, H. Huang and Z. Ren, Selective delivery of curcumin to breast cancer cells by self-targeting apoferritin nanocages with pH-responsive and low toxicity, *Drug Delivery*, 2022, **29**(1), 986–996.

139 A. N. Neagu, T. Jayaweera, K. Weraduwage and C. C. Darie, A Nanorobotics-Based Approach of Breast Cancer in the Nanotechnology Era, *Int. J. Mol. Sci.*, 2024, **25**(9), 4981.

140 F. Lebre, N. Chatterjee, S. Costa, E. Fernández-de-Gortari, C. Lopes and J. Meneses, *et al.*, Nanosafety: An Evolving Concept to Bring the Safest Possible Nanomaterials to Society and Environment, *Nanomaterials*, 2022, **12**(11), 1810.

141 B. Kaur, S. Kumar and B. K. Kaushik, Recent advancements in optical biosensors for cancer detection, *Biosens. Bioelectron.*, 2022, **197**, 113805.

# A Minimalist MOS-based Meminductor Emulator Circuit

Rahul Kumar Gupta
Electronics and Communication Engineering
Delhi Technological University,
Delhi, India
rahulkumargupta_2k21phdec08@dtu.ac.in

Mahipal Singh Choudhry
Electronics and Communication Engineering
Delhi Technological University,
Delhi, India
msc_1976@yahoo.com

Varun Saxena
School of Engineering
Jawaharlal Nehru University
Delhi, India
varunsaxena@jnu.ac.in

2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT) | 979-8-3315-4335-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/DICCT64131.2025.10986325

***Abstract*— A simple grounded meminductor emulator (MIE) based on a MOSFET configuration is presented in this paper. With six MOSFETs and two capacitors, the proposed circuit may efficiently create pinched hysteresis loops (PHL) at frequencies up to 100 KHz. The dependability of the design is confirmed by thorough analyses that include PVT assessments, which refer to modeling chips at various process, voltage, and temperature corners and non-volatility testing. Simulations were performed with 180 nm CMOS technology characteristics in CST (Cadence Spectre Tool). The consumption of power is a few microwatts in the proposed MIE. A comparative analysis of existing designs illustrates the benefits of the proposed MIE circuit.**

**Keywords— Meminductor, hysteresis loop, nonvolatile, incremental, decremental, memelement, grounded.**

## I. INTRODUCTION

Electrical and electronic circuits have traditionally relied on capacitors (C), resistors (R), and inductors (L) as their fundamental building blocks. Professor Leon Chua first proposed the theory of memristors as a fourth circuit element in 1971 and 1976 [1,2]. Many years later, in 2008, Stanley Williams's team at HP Labs achieved a significant advancement in circuit theory by successfully creating the first physical memristor [3]. Building on this breakthrough, scientists went on to build memcapacitors and meminductors by applying the idea of memory to the other essential parts [4]. The relationship of different electrical terminologies related to fundamental elements is shown in Fig. 1. The emergence of memristors [1-4, 16, 17, 18], memcapacitors [6], and meminductors, opened up new avenues for the design of advanced circuits and systems [6]. Unlike memristors and memcapacitors, which have seen some commercial availability and potential application, such as crossbar arrays [28-30].

These emulators provide researchers with a practical means to examine the theoretical properties and applications of meminductors without constraints of physical limitations [7-8]. Numerous approaches have been explored in the design of meminductor emulators, leading to a wide variety of circuit topologies [9]. These designs often leverage commercially available components, such as OTAs-operational transconductance amplifiers [19, 22, 24, 27], current conveyors [19-20], and multipliers [7, 20-21, 24-27], VDTA- voltage differential transconductance amplifier [23], differential-voltage current conveyor [24], voltage-difference buffered amplifier [25], along with passive components like resistors and capacitors [10-11]. Some designs incorporate memristors [11] for meminductor implementation. Flux-controlled emulators rely on the magnetic flux linkage to define the meminductor's behavior [9], while charge-controlled emulators utilize the electric charge as the control parameter [7]. The design of meminductor emulators also considers the terminal configurations, distinguishing between grounded [11-12, 19, 21, 22, 24, 26] and floating emulators [20, 23, 25].

The development of accurate and efficient meminductor emulators has opened up numerous possibilities across diverse fields. These emulators are crucial for investigating the fundamental behavior of meminductors and exploring their potential in various applications [12] such as chaotic oscillators [13], neuromorphic circuits [12-13], high-frequency filters [9-10], and amplitude modulation [8], etc.

Even with great advancement in meminductor emulator design, numerous difficulties still exist [14]. Emulators' accuracy, especially at higher frequencies, has to be improved more yet. Some emulator designs' intricacy can make their useful implementation difficult [10–11]. Moreover, wide acceptance depends on the creation of strong and dependable emulators immune to process fluctuations and temperature changes [10]. Development of simpler, more accurate, and strong emulators capable of running over larger frequency ranges should be the main emphasis of the next studies. Exploration of new circuit topologies and active components might result in major developments in this field [12]. Moreover essential to provide consistent and dependable performance evaluation is the development of standardized testing techniques for meminductor emulators [14]. At last, investigating the possibilities of digital meminductor emulators—such as FPGAs—could have benefits in terms of programmability and flexibility [14-15]. Limiting their applicability in real-world applications, existing meminductor emulators can suffer from complex circuit topologies, greater component counts, and susceptibility to process, voltage, and temperature (PVT) fluctuations.

The presented research paper goes through the following outlines such as Section II focuses on the proposed meminductor emulator schematic diagram and mathematical analysis, Subsequently, Section III presents the different analyses through simulation using the 180nm GPDK

979-8-3315-4335-8/25/$31.00 ©2025 IEEE



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:54:11 UTC from IEEE Xplore. Restrictions apply.

technology. The comparison is discussed in Section IV with the existing meminductor emulator circuits. Section V concludes the proposed work and the future scope.

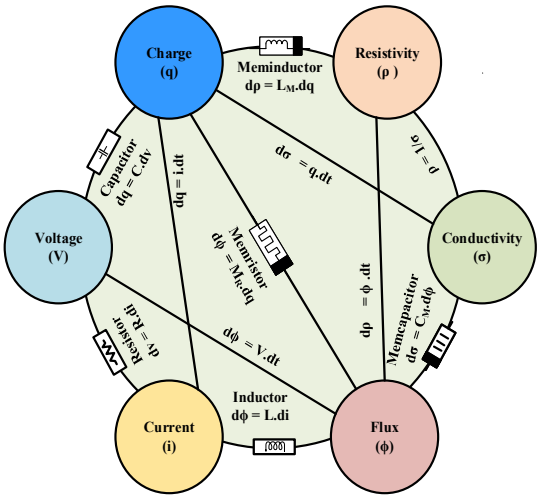



Fig. 1. Relationship between terms related to fundamental elements.

## II. PROPOSED MEMINDUCTOR EMULATOR

This research introduces a new and simple meminductor emulator circuit, as illustrated in Fig. 2, comprising six MOSFETs and two capacitors. Among these components, the transistors $M_1$ and $M_2$ function as common source (CS) amplifier in which $M_2$ drives the transistor, and $M_1$ functions as a current source load controlled by $V_{C1}$. $M_3$ and $M_4$ also act as a CS amplifier in which $V_{C2}$ is applied as an input to transistor ($M_4$) and transistor ($M_3$) is the current source load controlled by an external biasing voltage $V_B$. Similarly, $M_5$ and $M_6$ also represent the CS stage. In the third stage of the common source amplifier, $V_{C1}$ is applied to the $M_6$ transistor as input, and $M_5$ is the current source load that $V_{C2}$ controls. The aspect ratio (W/L) is illustrated in Table I.

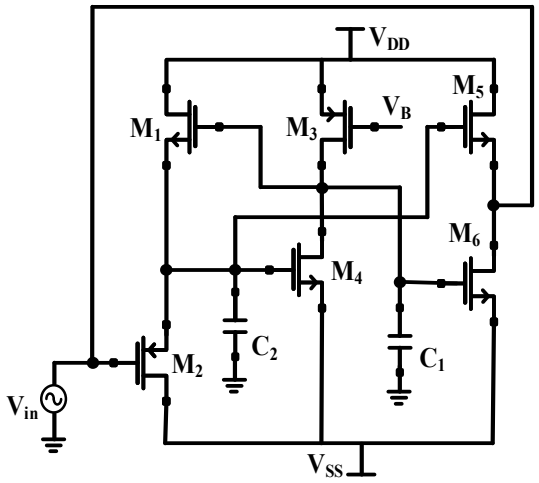



Fig. 2. Circuit diagram of proposed meminductor emulator.

**TABLE I**
**W/L RATIO OF USED MOSFETS**

| MOSFETs | Type | W/L Ratio (*μm*) |
|---|---|---|
| $M_1$, $M_3$ | N-type MOS, P-type MOS | 4 μm/ 0.18 μm |
| $M_{2\text{-}4}$ | N-type MOS | 2 μm/ 0.18 μm |
| $M_{5\text{-}6}$ | N-type MOS | 6 μm/ 0.18 μm |

## III. SIMULATION RESULTS

The simulation part of the proposed MIE is discussed in this section. CVS tool with 180 nm GPDK technology illustrates every observation. Supply voltages $V_{DD}$ = 0.9V and $V_{SS}$ = -0.9 V and using $C_1$ and $C_2$ values of 500 pF and 100 pF, respectively, the proposed MIE characteristic graphs were plotted using bias voltages $V_B$ = -0.8V.

The objective is to observe the transient characteristics of the proposed MIE. The transient response for the meminductive behavior of this MIE is tracked with a sinusoidal input signal ($V_{in}$ = 400 mV and $f_{in}$ = 10 KHz). Fig. 3 illustrates the phase lag between flux (φ) and input current, hence displaying the inductive character of the proposed MIE. Figs. 3 and 4 show the waveform of input current (i), flux(φ), and voltage (V), for the proposed MIE.

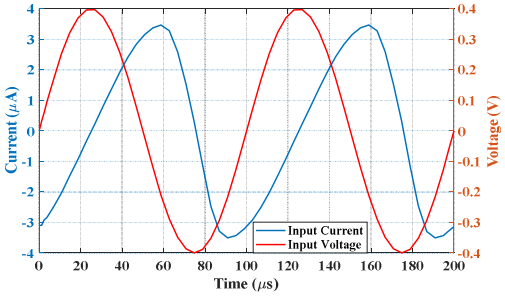



Fig. 3. Transient behaviour of MIE (Input voltage and input current).

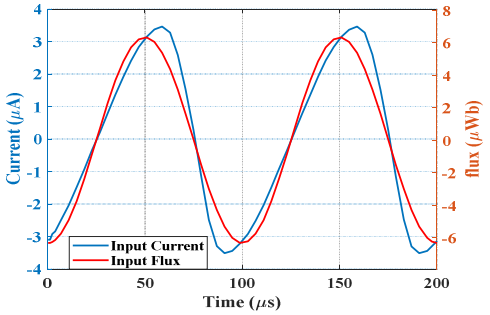



Fig. 4. Transient behaviour of MIE (Integration of input voltage as an input flux and input current).

Non-volatile behavior of semiconductors is among their most significant features. It is expected that the meminductance will vary accordingly throughout the ON duration of a pulse signal and remain constant in its OFF duration. We observe this nonvolatile behavior using a pulse input signal ($V_1$ =100 mV, $V_2$ = 0 mV, $T_{ON}$ =5 μs, and an $T_{OFF}$ = 5 μs). The ratio of flux (φ) and the input current (i), Fig. 4 shows the meminductance. Fig. 5 makes clear that the meminductance stays constant throughout the OFF time of the input pulse and rises linearly during the ON time. This



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:54:11 UTC from IEEE Xplore. Restrictions apply.

observation supports the non-volatile nature of proposed MIE.

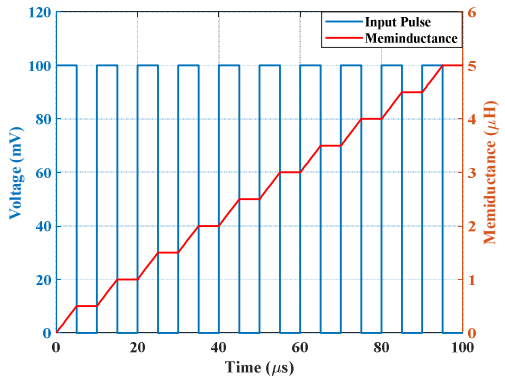



Fig. 5. Non-volatile characteristics of proposed MIE circuit.

Under a sinusoidal signal, the proposed MIE also shows a key characteristic known as a pinched hysteresis loop (PHL) in the flux (φ) against the current (i). The PHL curves are displayed in Fig. 6 to Fig. 11 for the input sinusoidal signal with the different frequencies and an amplitude of 400 mV. The dumbbell-shaped PHL curve with zero crossover supports even more the meminductive behavior of the proposed MIE. PHL curves over the range of 100Hz to 100 KHz, MIE are shown in Fig. 6 to Fig. 11. PHLs are also observed at some high frequencies over the range of 10 KHz to 100 KHz, as shown in Fig. 8 to Fig. 11. We can observe the non-ideal properties of the proposed MIE circuit and shifting of pinching at high frequencies.

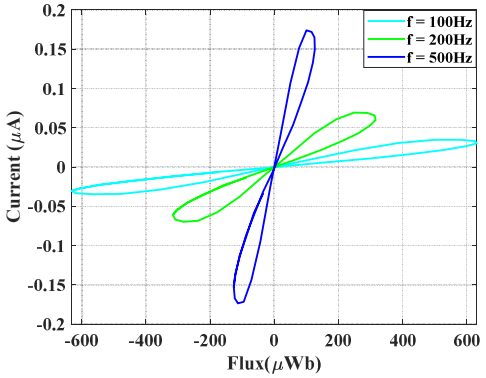



Fig. 6. PHL curve ($f_{in}$: 100 Hz to 500 Hz, $V_{in}$ =400mV).

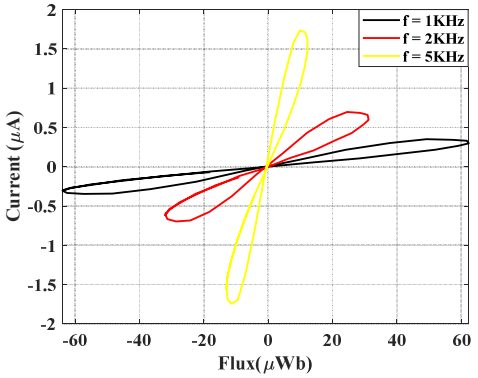



Fig. 7. PHL curve ($f_{in}$: 1 KHz to 5 KHz, $V_{in}$ = 400mV).

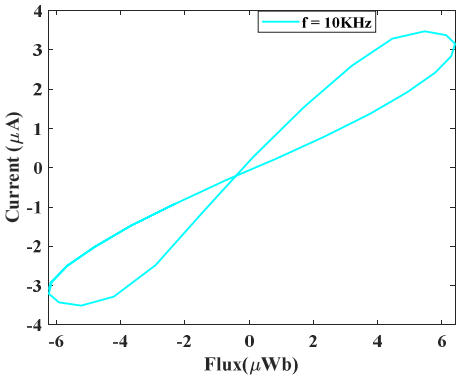



Fig. 8. PHL curve ($f_{in}$ =10 KHz, $V_{in}$ = 400 mV).

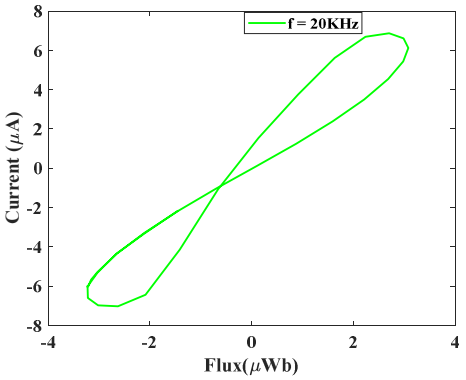



Fig. 9. PHL curve ($f_{in}$ =20 KHz, $V_{in}$ = 400 mV).

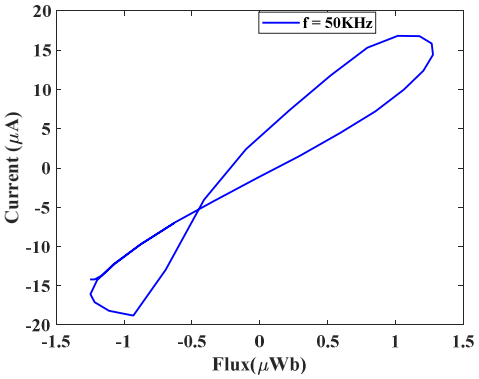



Fig. 10. PHL curve ($f_{in}$ =50 KHz, $V_{in}$ = 400 mV).

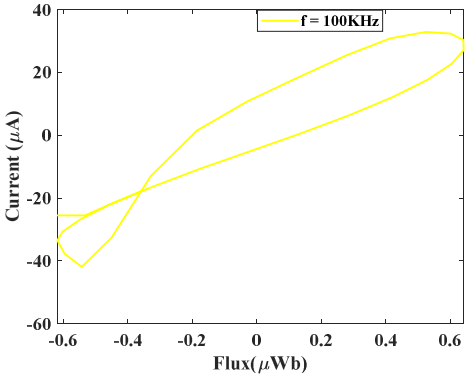



Fig. 11. PHL curve ($f_{in}$ =100 KHz, $V_{in}$ = 400 mV).

Many applications depend on capacitive tuning, hence the circuit's capacitance tunability has to be investigated. Variations in capacitance affect the PHL curve, therefore



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:54:11 UTC from IEEE Xplore. Restrictions apply.

establishing the capacitance tuning of the circuit. The PHL curves of the proposed MIE with input signal ($f_{in}$ =20 KHz, $V_{in}$ = 400 mV) at various values of $C_2$ and $C_1$ are examined. Fig. 13 shows the PHL curves with zero crossing for variations in capacitance $C_2$ between 50 pF and 200 pF. These graphs show the really good capacitance tuning ability of the proposed MIE. Likewise, Fig. 12 shows the zero crossing PHL curves generated with a variation in capacitance $C_1$ between 100 pF and 1nF.

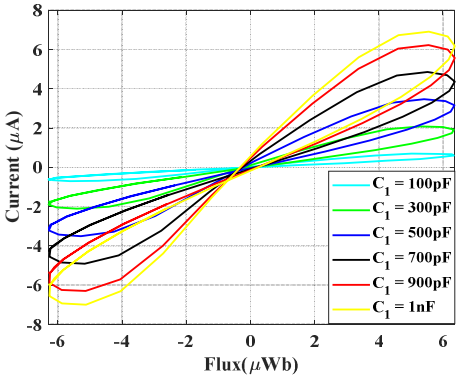



Fig. 12. PHL curve ($C_1$: 100pF to 1nF).

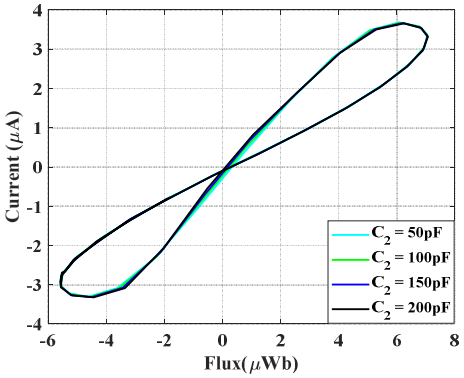



Fig. 13. PHL curve ($C_2$: 50pF to 200pF).

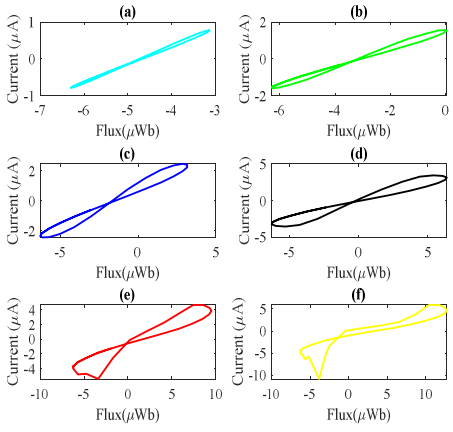



Fig. 14. PHL curve ($V_{in}$ : 300 mV to 800 mV).

Electronic tunability is another important feature of proposed MIE in reaction to changes in applied bias voltage or current. Fig. 15 illustrates the PHL curve over a range of $V_B$ values from -300 mV to -1 V. It illustrates the electronic tunability characteristic of the proposed MIE. With the range of 300 mV to 800 mV at a frequency of 10 KHz, Fig. 14 displays the PHL curve of the proposed MIE circuit for various input supply voltages.

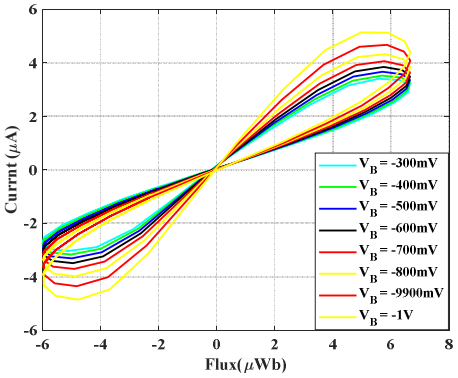



Fig. 15. PHL curve ($V_B$: -300 mV to -1 V).

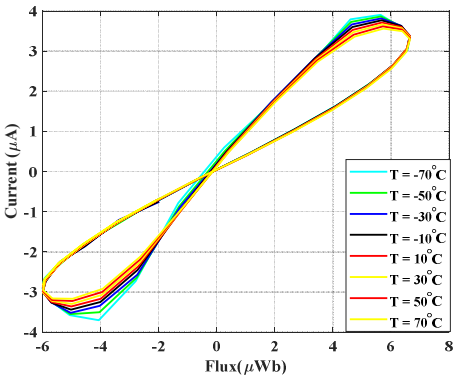



Fig. 16. PHL curve (-70$^0$C to 70$^0$C).

We have investigated the reaction of the proposed MIE for temperature fluctuations between -70°C and +70°C. The PHL curve is shown in Fig. 16 together with the expected meminductor behavior across the specified temperature range. Consistent shapes in the PHL curves supports the proposed MIE efficient operation as a meminductor emulator.

## IV. COMPARISON AND DISCUSSION

Following this, a comprehensive comparison was undertaken to assess the proposed meminductor emulator circuit about existing counterparts [19-27] consisting of different parameters such as active devices, MOSFET counts, which is higher in [19, 20, 23, 26], passive components, operating frequency, tunability [19, 23, 24, 25] or non tunable [20, 21, 22, 26], and power consumption which is higher with the more number of MOSFETs [24, 27].

The most disadvantage of the existing meminductor emulator circuits is complexity. The proposed MIE has the following advantages:

1. There is no active building block used.
2. This is a very simple circuit.
3. Power consumption is much less (in μW).
4. There is no multiplier used in the circuit.
5. Only six MOSFETs are used with two capacitors.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:54:11 UTC from IEEE Xplore. Restrictions apply.

The consumption of power is few microwatts because of no use of active building blocks. The comparison is outlined in Table II.

## V. Conclusion

This work aims to replicate a MIE circuit in an alternative manner using six MOSFETs and two capacitors with external biasing. The proposed circuit provides a more simple and efficient design than previously proposed circuits. Simulation results reveal that the circuit performs satisfactorily. Moreover, the study evaluates the impact of temperature changes, therefore confirming the durability of the circuit in several conditions. The key outputs of this work are lower component count, more design simplicity, and consistent performance under several situations. These qualities make the proposed MIE a strong component for future applications in signal processing, neuromorphic computing, and programmable analog circuits. To verify the practicality of the circuit, experimental validation will be required for further research. Further improvements in real electrical systems could boost power efficiency and extend the frequency range, thereby making it even more beneficial.


## ACKNOWLEDGEMENT

The authors appreciate Delhi Technological University giving access to the simulation tools needed for the suggested work to be carried out. The author also thanks his supervisors for their ongoing encouragement.


**TABLE II**
**COMPARISON OF EXISTING MEMRISTORS WITH PROPOSED MEMRINDUCTOR EMULATORS TAKING DIFFERENT PARAMETERS**

| Ref | Active Building Blocks | MOSFETs | RLCs | Multiplier less | $M_R$ | Frequency | Electronic Tunable | Consumed Power | Floating (F) /Grounded (G) |
|---|---|---|---|---|---|---|---|---|---|
| [21] | CCII = 2, OTA = 1 | 33 | R-1, C-2, L-0 | Yes | 0 | 300 KHz | Yes | NA | Grounded |
| [22] | CCII = 1, OPAMP = 8 | 31 | R-9, C-2, L-0 | No | 0 | 400 Hz | No | NA | Floating |
| [23] | OPAMP = 7 | - | R-13, C-2, L-0 | No | 0 | 300 Hz | No | NA | Grounded |
| [24] | OTA = 1 | 12 | R-1, C-1, L-1 | Yes | 0 | 500 KHz | No | NA | Grounded |
| [25] | VDTA = 2 | 64 | R-0, C-2, L-0 | Yes | 0 | 1.5 MHz | Yes | NA | Floating |
| [26] | DVCC = 1, OTA = 2 | 34 | R-1, C-2, L-0 | No | 0 | 10 MHz | Yes | 120uW | Grounded |
| [27] | VDBA = 1, CDBA = 1 | 36 | R-1, C-2, L-0 | Yes | 0 | 2 MHz | Yes | NA | Floating |
| [28] | OPAMP = 2 | - | R-3, C-1, L-0 | No | 1 | 5 KHz | No | NA | Grounded |
| [29] | MOOTA = 1, OTA = 1 | 54 | R-1, C-2, L-0 | No | 0 | 1 MHz | Yes | 3.87mW | Floating |
| Proposed Work | - | 6 | R-0, C-2, L-0 | Yes | 0 | 100 KHz | Yes | 1.19uW | Grounded |

## References


[1] L. Chua, "Memristor-The missing circuit element," in IEEE Transactions on Circuit Theory, vol. 18, no. 5, pp. 507-519, September 1971.

[2] L. O. Chua and Sung Mo Kang, "Memristive devices and systems," in Proceedings of the IEEE, vol. 64, no. 2, pp. 209-223, Feb. 1976.

[3] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found", nature, 453(7191), pp.80-83, 2008.

[4] M. Di Ventra, Y. V. Pershin and L. O. Chua, "Circuit Elements With Memory: Memristors, Memcapacitors, and Meminductors," in Proceedings of the IEEE, vol. 97, no. 10, pp. 1717-1724, Oct. 2009.

[5] Y. Babacan. "An operational transconductance amplifier-based memcapacitor and meminductor." Electrica 18, no. 1, pp. 36-38, 2018.

[6] Q. Zhao, C. Wang, & X. Zhang, "A universal emulator for memristor, memcapacitor, and meminductor and its chaotic circuit. Chaos" An Interdisciplinary Journal of Nonlinear Science, 29(1), 2019.

[7] P. K. Sharma, R. K. Ranjan, F. Khateb and M. Kumngern, "Charged Controlled Mem-Element Emulator and Its Application in a Chaotic System," in IEEE Access, vol. 8, pp. 171397-171407, 2020.

[8] P. Kumar, & S. K. Paul. "High-frequency tunable grounded & floating incremental-decremental meminductor emulator and application." arXiv e-prints, 2022, arXiv-2205.

[9] M. K. Khan, N. Bhuwal, S. Sagar and D. Gupta, "Meminductor Emulator via Flux Approach for Wide Frequency Range Applications," 2024 28th International Symposium on VLSI Design and Test (VDAT), Vellore, India, pp. 1-6, 2024.

[10] N. Bhuwal, K. Solanki, M. K. Majumder and D. Gupta, "Flux Controlled Grounded Meminductor Emulator Using Single DVCCTA," 2024 37th International Conference on VLSI Design and 2024 23rd International Conference on Embedded Systems (VLSID), Kolkata, India, pp. 107-112, 2024

[11] A. Gupta, S. K. Rai, & M. Gupta. "Grounded meminductor emulator using operational amplifier-based generalized impedance converter and its application in high pass filter." International Journal of Electrical and Electronics Research 10, no. 3 pp. 496-500, 2022.

[12] J. Vista and A. Ranjan, "High Frequency Meminductor Emulator Employing VDTA and its Application," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 10, pp. 2020-2028, Oct. 2020.

[13] N. Raj, R. K. Ranjan, F. Khateb and M. Kumngern, "Mem-Elements Emulator Design With Experimental Validation and Its Application," in IEEE Access, vol. 9, pp. 69860-69875, 2021.

[14] O. M. Afolabi,, V. A. Adeyemi, E.T. Cuautle, & J. C. Nuñez-Perez. "FPGA Realization of a Fractional-Order Model of Universal Memory Elements." Fractal and Fractional 8, no. 10, 605, 2024.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:54:11 UTC from IEEE Xplore. Restrictions apply.

[15] T. H. Vu, T. D. Vu, & et. al.. "FQsun: A Configurable Wave Function-Based Quantum Emulator for Power-Efficient Quantum Simulations." arXiv preprint arXiv:2411.04471, 2024.

[16] P. Srivastava, R. K. Gupta, R.K. Sharma, and R. K. Ranjan, "MOS-only memristor emulator". Circuits, Systems, and Signal Processing, 39, pp.5848-5861, 2020.

[17] R. K. Gupta, M. S. Choudhry, V. Saxena, and S. Taran, "A single MOS-memristor emulator circuit", Circuits, Systems, and Signal Processing, 43(1), pp.54-73, 2024.

[18] R. K. Gupta, M. Joshi, & et. al., "A high-frequency compact memristor emulator circuit and its applications as wave shaping and generation circuit" Chaos, Solitons & Fractals, 192, p.115964, 2025.

[19] S. K. Rai, B. Aggarwal, and V. Singroha, "Meminductor emulators using off-the-shelf active blocks with application in chaotic oscillator." Integration 96, 102165, 2025.

[20] M. P. Sah, R.K. Budhathoki, C Yang, and H. Kim, H., "Mutator-based meminductor emulator for circuit applications", Circuits, Systems, and Signal Processing, 33, pp.2363-2383, 2014.

[21] F. Yuan, Fang, and Y. Li. "A chaotic circuit constructed by a memristor, a memcapacitor and a meminductor." Chaos: An Interdisciplinary Journal of Nonlinear Science 29, no. 10, 2019.

[22] Y. Babacan, "An operational transconductance amplifier-based memcapacitor and meminductor", Electrica, 18(1), pp.36-38, 2018.

[23] K. Bhardwaj, K. and M. Srivastava, "New electronically adjustable implement emulator for realizing the behavior of fully-floating meminductor and memristor". Microelectronics Journal, 114, p.105126, 2021.

[24] A. Raj, S. Singh, and P. Kumar, "Electronically tunable high-frequency single output OTA and DVCC-based meminductor". Analog Integrated Circuits and Signal Processing, 109, pp.47-55, 2021.

[25] N. Yadav, S. K. Rai, and R. Pandey "An electronically tunable meminductor emulator and its application in chaotic oscillator and adaptive learning circuit", Journal of Circuits, Systems and Computers, 32(02), p.2350031, 2023.

[26] P. B. Petrović, "A new electronically controlled floating/grounded meminductor emulator based on a single MO-VDTA" Analog Integrated Circuits and Signal Processing, 110(1), pp.185-195, 2022.

[27] D. Ersoy and F. Kaçar, "Electronically Charge-Controlled Tunable Meminductor Emulator Circuit With OTAs and Its Applications," in IEEE Access, vol. 11, pp. 53290-53300, 2023.

[28] F. Lalchhandama, "Scalable Synthesis of Logic Functions Using Memristor Crossbar Array" Doctoral dissertation, IIT Kharagpur", 2022

[29] N. M. G., F. Lalchhandama, K. Datta and I. Sengupta, "Modelling and Simulation of Non-Ideal MAGIC NOR Gates on Memristor Crossbar," 2018 8th International Symposium on Embedded Computing and System Design (ISED), Cochin, India, pp. 124-128, 2018.

[30] D. N. Yadav, P L Thangkhiew, & et. al., "Improving Self-Fault-Tolerance Capability of Memristor Crossbar Using a Weight-Sharing Approach." Rn 1, V1, 2024.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:54:11 UTC from IEEE Xplore. Restrictions apply.

IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

**Traitement du Signal**
Vol. 42, No. 2, April, 2025, pp. 593-606
Journal homepage: http://iieta.org/journals/ts

# An Efficient Attention-Based LSTM Framework for Blood Glucose Level Prediction

Sunny Arora[1*], Shailender Kumar[1], Pardeep Kumar[2]

[1] Department of Computer Science Engineering, Delhi Technological University, Delhi 110042, India
[2] Department of Computer Science, Warwick Manufacturing Group (WMG), University of Warwick, Coventry CV4 7AL, UK

Corresponding Author Email: sunnyarora_2k18phdco@dtu.ac.in

Copyright: ©2025 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.420201

**Received:** 28 April 2024
**Revised:** 4 November 2024
**Accepted:** 25 January 2025
**Available online:** 30 April 2025

***Keywords:***
*diabetes, blood glucose, prediction, Long Short-Term Memory (LSTM), attention, RMSProp optimizer*

**ABSTRACT**

Predicting blood glucose is highly significant for patients with diabetes to manage their condition efficiently. Deep learning (DL) approaches have demonstrated great potential in blood glucose prediction modeling. By leveraging time-series data from continuous glucose monitoring (CGM) devices, this model can capture complicated temporal dependencies and patterns in glucose dynamics. DL based blood glucose prediction models learn from insulin dosages, past glucose readings, physical activity levels, meal intake, and other related features to forecast future blood glucose levels with maximum accuracy. Thus, this study presents an Optimal Attention-based Long Short-Term Memory for Blood Glucose Level Prediction (OALSTM BGLP) model. Firstly, it employs Min-Max scaling to normalize the input data, ensuring consistent and meaningful comparisons across different features. Additionally, the model generates time series data for multiple forecasting horizons, including 15, 30, 45, and 60-minute(m) intervals, enabling flexible and dynamic predictions to accommodate various planning and decision-making needs. Moreover, the OALSTM-BGLP technique uses the ALSTM model, which incorporates an attention mechanism to selectively concentrate on relevant information within the input sequence while capturing long-term dependencies. This attention mechanism permits the method to effectively extract salient features from the input data, enhancing its predictive capabilities. Furthermore, the model is optimized using the RMSProp optimizer, which adjusts the rate of learning dependent on the magnitude of recent gradients, facilitating efficient training and convergence. The performance evaluation of the developed technique on the OhioT1DM dataset shows its promising performance over recent state-of-the-art methods.

## 1. INTRODUCTION

Diabetes mellitus is a metabolic disorder that disrupts the body's ability to regulate blood glucose (BG) levels. Beta cells in the pancreas secrete insulin, an endocrine hormone that regulates glucose utilization [1]. Type 1 diabetes is characterized by the loss of these insulin-producing beta cells, leading to insulin deficiency [2]. As a result, individuals with diabetes must continuously monitor their blood glucose levels, regulate them through consistent insulin administration, and make informed decisions about their medical regimen (e.g., once daily or before each meal) to maintain normal blood sugar level [3]. The primary objective in diabetes management is to optimize insulin dosages to prevent hyper/ hypo glycemia [4]. This is difficult due to the multitude of factors, including lifestyle, mental state, stress, and physical exercise, that also affect insulin intake, nutrition, and glucose levels [4]. Despite various advancements in diabetes observation, continuous glucose monitoring (CGM) remains aggressive. CGM devices are capable of providing an individual's glycemic condition at an assumed time. A positive recognition could dramatically expand the everyday behavior of diabetes by the patients themselves [5].

To simplify and accomplish the management of type 1 diabetes, a continual growth in automated procedures is quite essential, despite the advancements obtained thus far. In this way, commerce-precise BG devices might be game-altering [6]. These devices can provide early warnings about potential glycemic actions, enabling the implementation of either automatic or non-automated preventive measures. Moreover, these mechanisms are a criterion for the arrival of a closed-loop artificial pancreas as the present vision for the final automatic management of type 1 diabetes [7]. Hybrid modelling, data-driven techniques, and physical techniques are utilized for forecasting blood glucose levels. The data-driven method explores the present and historical values of diabetes management-linked variables to plan prospective BG excursions [8].

Generally, there are three primary categories of time-series prediction methods: classical time-series prediction, machine learning (ML), and deep learning (DL). The research reported various studies that used classical time series and ML based regression models, NNs, and DL methodologies for BG-level predictions. Eren-Oruklu et al. [9] designed an adaptive univariate autoregressive moving average (ARMA) model with fixed model orders solely based on CGM data. Meanwhile, Turksoy et al. [10] developed a multivariate ARMA with exogenous inputs (ARMAX) model, which used

exogenous variables such as glucose and insulin while excluding meal information. Given the time-varying non-stationarity of CGM data, Yang et al. proposed an ARIMA model using an algorithm that adaptively determines model orders and simultaneously updates model parameters [4]. They found that their model outperformed the adaptive univariate and ARIMA models. As more research was done, Georga et al. [11] suggested SVR and SVR with feature selection using RF and RReliefF. Hamdi et al. [12] also proposed the DE-SVR for BG-level predictions.

Researchers found that ML-based classical time series struggle to predict BG dynamics due to their assumption of linear data, sensitivity to non-stationary CGM data, difficulty in capturing complex patterns, susceptibility to noise and outliers, and inability to adapt to changing conditions. However, even though SVR and SVR with kernel functions can handle non-linearity, the limitation of machine learning methods is how the input data is represented, which has an effect on how well the predictions work.

Motivated by the excellent modelling capabilities of artificial neural networks (ANNs) for nonlinear and non-stationary phenomena, several studies have employed ANN strategies to forecast BG levels. DL differs from traditional ANNs in the number of hidden layers, their interconnections, and their ability to abstract inputs. These intricate models with increased complexity outperformed conventional shallow NNs. These advanced DL based algorithms can autonomously discern patterns from data. While feed-forward ANNs fared better at predicting BG levels, traditional RNNs have discovery constraints due to the vanishing/exploding gradient problem, which is a major drawback of using these networks.

The incorporation of memory cells and forget gates into traditional RNNs has rectified this issue in Long Short-Term Memory (LSTM) networks.

Recent studies examined RNNs and LSTMs for enhanced accuracy in BG level prediction. Fox et al. [13] employed a RNN with Gated Recurrent Unit (GRU) cells, whereas Sun et al. [14] implemented a Bi-LSTM based RNN for BG prediction. Nevertheless, the precision attained by cutting-edge models for actual patient data is insufficient, rendering these methods potentially inapplicable for diabetes care. Conventional LSTMs sometimes forfeit essential information throughout extended sequences due to the restricted capacity of their memory cells to preserve data from distant time steps. The discovery of the attention mechanism allowed researchers to address long sequencing issues and concentrate on the most pertinent segments of the sequence. Traditional LSTM models may dilute sequential information over long time spans, potentially leading the model to overlook subtle yet influential glucose patterns. The attention method enables our model to concentrate on significant previous CGM readings that may more substantially impact future glucose levels.

This study offers an Optimal Attention-Based Long Short-Term Memory for BG Level Prediction (OALSTM-BGLP) Model. At first, it employs Min-Max Scaling to normalize the input data, certifying consistent and meaningful contrasts across dissimilar features. Furthermore, the OALSTM-BGLP approach utilizes the ALSTM technique, which includes an attention mechanism to selectively concentrate on related data within the input sequence while taking into account long-term dependencies. Besides, the method has been optimized by employing the RMSProp optimizer, which adjusts the rate of learning based on the magnitude of recent gradients, facilitating efficient training and convergence. The suggested methodology outperforms state-of-the-art methods in a performance evaluation conducted on the OhioT1DM dataset.

## 2. LITERATURE REVIEW

For the purpose of studying and detecting diabetes mellitus, Zhang et al. [15] proposed AHDHS-Stacking, an EL framework. The two components—a feature selector (FS) and an optimiser of basic learner groupings—utilize the harmony search (HS) approach. An adaptive hyperparameter technique was used to speed up the iterative process, and the average performance of each base learner was used as an aim of FS to create the full performance. Yang et al. [16] introduced a method for abnormal blood glucose detection using an autonomous neighborhood parameter selection strategy, which is based on Adaptive Density-Based Spatial Clustering of Applications with Noise. To make up for missing blood glucose values, Yang et al. [16] proposed a feature-engineered imputation technique. Lastly, global and local spatial-temporal properties are extracted from sequence data using a temporal multi-head attention model. In the study by Butt et al. [17], a multi-layer LSTM-based RNN is used to predict BG levels in people with type 1 diabetes. The BG level is being predicted within a forecast limit using the created model.

Langarica et al. [18] suggested the Input and State Recurrent Kalman Network (ISRKN) as a new deep learning-based method for predicting glucose levels based on probabilities . In the latent space of a DNN, this network combines an input and a state Kalman filter. This lets secure calculations of later distributions and the distribution of uncertainty through Kalman calculations happen. Song et al. [19] studied a new combined ML technique termed Bagging-ABC-ELM. The artificial bee colony (ABC) system achieved the optimum input biases and weights of the ELM method. The bagging model was utilized to improve the constancy of the system and achieve greater performance than ELM. Ye et al. [20] suggested using an electric nose (E-Nose) device equipped with a metal oxide (MOX) gas sensor array as a new method to quantitatively recognize and study BG levels by determining breath biomarkers. Progressive ML algorithms have investigated and suggested exact predictions of the BG level based on the dimensions of 41 members over 10 days. Zhu et al. [21] designed an IoMT-enabled wearable device to implement the embedded model, a new DL based model that employs an attention-based evidential RNN. The device consists of a low-power, low-cost system on a chip that uses Bluetooth connectivity and edge computing to detect predictive hypoglycemia and make real-time BG predictions. Also created were desktop and cloud platforms for data storage and model fine-tuning, as well as an app for smartphones to display BG trajectories and forecasts. Lu and Song [22] suggested a new DL based hybrid technique to forecast levels of BG for 30 min of prediction horizon, that combined stacked Bi-GRU-based RNN, multi-layer perceptron, and Arora et al. [23] designed convolutional recurrent connection for 60 min aggregated CGM data for 60min prediction horizon and achieved 30.51 RMSE.

## 3. THE PROPOSED MODEL

In this study, we have presented a new OALSTM-BGLP methodology. The foremost intention of the OALSTM-BGLP

technique comprises different kinds of sub-processes involved in min-max normalization, ALSTM-based prediction process, and RMSProp optimizer-based parameter tuning. Figure 1 illustrates the workflow of the OALSTM-BGLP methodology.

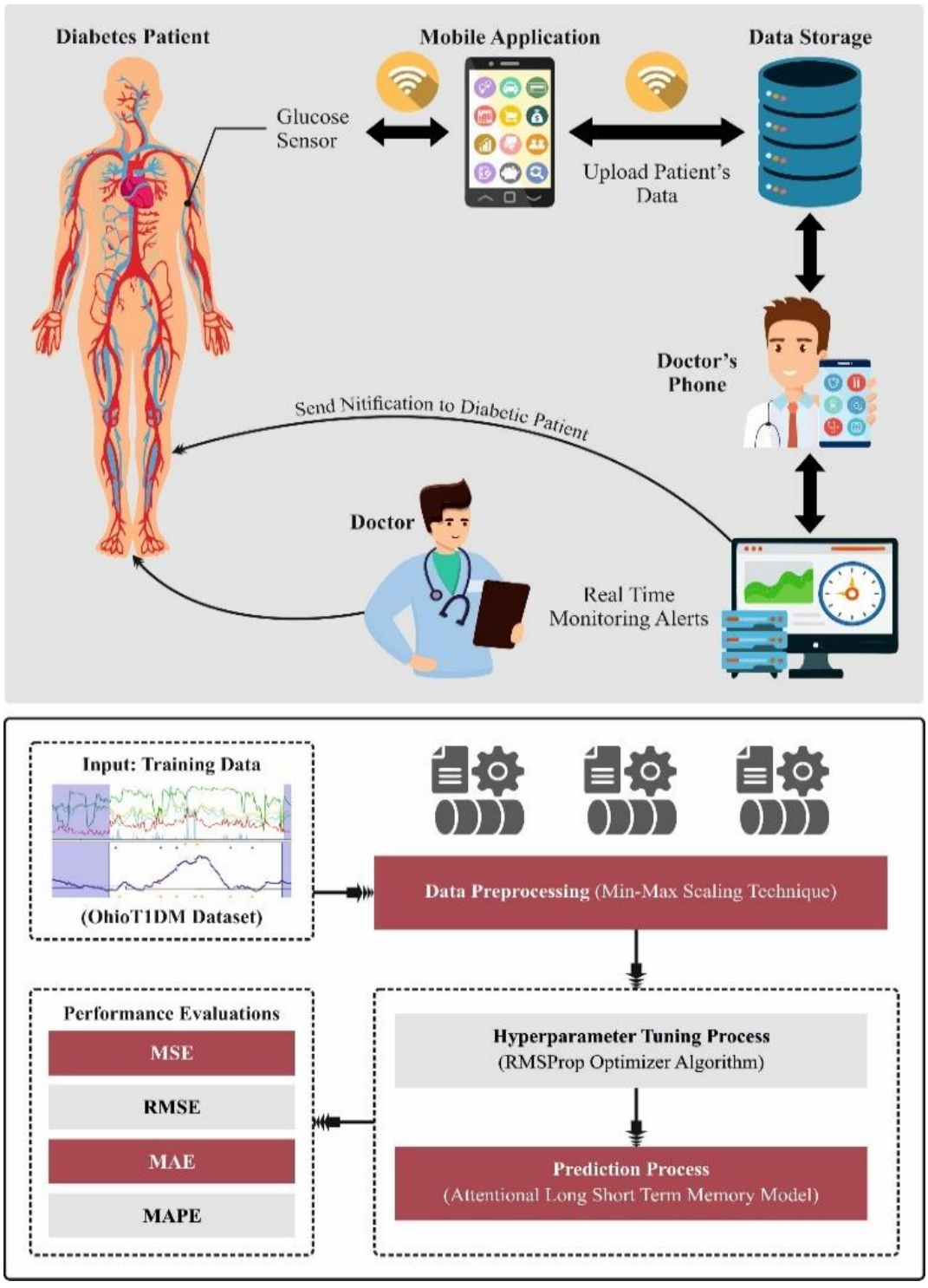



**Figure 1.** Workflow of the proposed OALSTM-BGLP methodology

### 3.1 Min-max normalization

Firstly, the OALSTM-BGLP method employs Min-Max scaling to normalize the input. Min-Max normalization is a popular data conversion method to retain the sensitive features in the data [24]. By doing so, 8 weeks real time CGM data for six patients from OhioT1DM [25] is normalized by the use of min-max normalization based on consideration of maximum and minimum values in the data. This method carries out a linear conversion on the original data. It is especially suitable for the classification task and is used in several applications namely neural networks, artificial intelligence, clustering, nearest neighbor classification, and so on.

The goal is to normalize the original data D into a preserved form of data D' that fulfills the privacy requirement with a least possible loss with high privacy information. This technique focuses mainly on data conversion through the min-max normalization to change the original data. All the attributes in a data are normalized by mounting the value such that they fall in the minimum range within [0.0-1.0]. $V_i$ of

attribute A from the range of [$min_i$-$max_i$] to [$newmin_A$-$newmax_A$].

$$V_i' = \frac{V_i min_A}{max_A - min_A}(\text{new}max_A\text{-new}min_A)+\text{new}min_A \quad (1)$$

where, $V_i$ indicates the new computed value. The relationship among the original values is preserved through the min-max normalization method.

### 3.2 ALSTM-based prediction process

For the prediction process, the OALSTM-BGLP technique uses the ALSTM model. In 1997, the LSTM network was initially developed [26]. Owing to its exclusive design framework, LSTM is appropriate for handling and forecasting significant actions with actual long intervals in time series. LSTM is nothing but a gated recursive neural network (GRNN) that inserts a gating device to switch the data spread in the NN based on RNN. For normal NN, the motive why performing gradient explosion is clarified below: there is a non-linear relation among the exterior condition at the latter stage and the exterior condition at this stage, and this link is assigned in every time step. The GRNN was able to resolve the problem of gradient explosion by raising the linear requirement among them. Generally, in LSTM, 3 control gate units contain namely output gate $o_t$ forget gate $f_t$ and input gate $i_t$. The gates of forget and input are the bases for LSTM to recall the dependency of long-term.

Input gate ($i_t$) defines how much data about the existing state wants to be kept in the interior state:

$$i_t = \sigma(U_i h_{t-1} + W_i x_t + b_i) \quad (2)$$

where, $\sigma$ denote the function of logistics, $W_i$ and $U_i$ represents the weight matrix; $b_i$ refers to the bias term. $h_{t-1}$ refers to the output of memory block at time $t$-1 and $x_t$ signify the input vector at time $t$.

Forget gate ($f_t$) defines how much data from the previous wants to be rejected.

$$f_t = \sigma\left(U_f h_{t-1} + W_f x_t + b_f\right) \quad (3)$$

where, $W_f$ and $U_f$ indicate the weight matrix; $b_f$ signifies the bias.

Output gate ($o_t$) fixes how much data from the interior state wants to output to the exterior state at an existing stage.

$$o_t = \sigma(U_o h_{t-1} + W_o x_t + b_o) \quad (4)$$

where, $W_o$ and $U_o$ denote the weight matrix; $b_o$ refers to the bias term.

The foremost steps of the LSTM prediction method are mentioned below.

(1) At first, the exterior state at the previous stage ($h_{t-1}$) and the input vector at the existing stage ($x_t$) have been employed to compute candidate state ($\tilde{C_t}$) and 3 gates;

(2) Next, merge the input gate ($i_t$) and forget gate ($f_t$) to upgrade the interior state ($C_t$) of the existing stage;

(3) Finally, dependent upon the output gate ($o_t$) the interior state data was transmitted to the exterior state ($h_t$).

The abovementioned steps can recognize the prediction of the LSTM technique recognized in Keras utilizing TensorFlow.

LSTM is capable of efficiently resolving the issue of gradient explosion [27]. At the procedure of training, overfitting will basis the test set to collapse. To resolve this issue, a dropout layer was inserted to enhance the framework of LSTM. Generally, the LSTM places the NN input vector to fixed measurement, which has reliable results while dealing with lower sizes. If the size of the input parameter is comparatively big, it will disturb the solution of the method owing to the size explosion issue. To concentrate on the prominent parameter, this research presents an attention mechanism (AM), which is recognized by preserving the LSTM intermediate output encoding for the input series.

Most attention techniques trust an Encoding-Decoding structure [28]. The encoding procedure of the original codec method produces a vector of intermediate in the technique of Seq2Seq that is utilized to keep the data of the original series. However, the vector is fixed. If the input original series length is extensive, then this vector can able to store restricted data which bounds the accepting capability of the system. Use an AM to pause the restraints of the original codec technique on a fixed vector. Figure 2 represents the infrastructure of ALSTM.

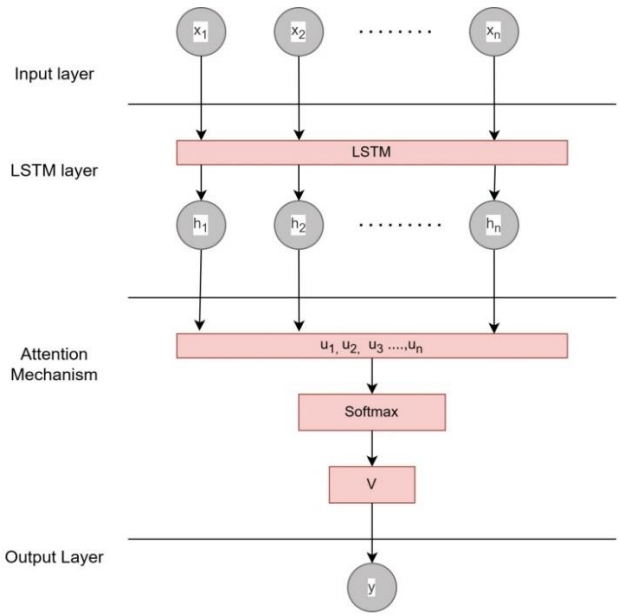



**Figure 2.** Structure of ALSTM

The AM is primarily executed in the subsequent steps [26]. In LSTM, the output [$h_1$, $h_2$, …, $h_n$] is changed non-linearly to get [$u_1$, $u_2$, $u_3$, … $u_n$]. When predicting BG levels using CGM data with an Attention LSTM, more attention is typically given to the most relevant past glucose readings that strongly influence the current prediction. The AM dynamically allocates higher weights to time steps in the input sequence where patterns or trends are most critical for making accurate predictions.

The significance of each intermediate time step may be represented by the attention weight matrix $\alpha_1$, $\alpha_2$, $\alpha_3$, …, $\alpha_n$, which is generated by the AM. The weight matrix ($W_k$) can signify the prominence of the intermediate state. Lastly, a weighted sum has been executed for a parameter of input and weight to get the encoded vector $V$. The output $y$ is acquired by decode as per the encode vector $V$. The complete formulation is mentioned below:

$$u_k = \tanh(W_k h_k + b_k) \quad (5)$$

$$\alpha_k = \frac{\exp(u_k^T u_s)}{\sum_{k=1}^{n} \exp(u_k^T u_s)} \quad (6)$$

$$V = \sum_k^n \alpha_k h_k \quad (7)$$

The weight matrix is represented by $W_k$, a bias term or offset is denoted as $b_k$, normalised attention weights are represented by $\alpha_k$, and the attention matrix for the time series CGM data is randomly initialised as $u_s$.

### 3.3 Hyperparameter tuning using RMSProp optimizer

Eventually, the parameter tuning process was executed by using the RMSProp optimizer to boost the ALSTM model performance. Optimizer is a hyperparameter required for training DL networks. They are widely classified into dual for non-convex optimizer issues, such as adaptive rate of learning optimization techniques, and non-adaptive rate of learning (classical SGD). Even though there is a second-order optimizer technique for a convex problem, the considerable variation among convex and non-convex optimizer issues for convex optimizer problem has one global goal, while the non-convex optimizer issue has more than one local goal. The objective is on first-order optimizer algorithms for non-convex problems. This is due to a non-convex optimization problem that is the most widespread in NN study. The loss surface defines the difficulty and complexity of finding the global optima. Therefore, an optimizer is employed for reducing the cost function which is evaluated by the cross-entropy, whereas the loss function calculates the error which shows the model efficiency. RMSprop is a gradient-based optimization method used in NN model training. A gradient of complex functions such as NN models tends to explode or vanish as the data transmits via the function. RMSprop was designed as a stochastic method for learning the mini-batch model.

RMSprop addresses the abovementioned problem through the moving average of the squared gradient for normalizing the gradients. This balances the step size (momentum), increases the step for a smaller gradient to prevent vanishing, and decreases the step for a larger gradient to prevent exploding. Geoffrey Hinton developed RMSProp, which is like an algorithm of gradient descent with motion [29]. RMSProp attempts to solve AdaGrad's entirely reducing rate of learning by utilizing a squared gradient moving average that uses the scale of current gradient descent for standardization. So, with the rate of learning upsurge, this method utilized might travel in a straight way with faster converge of higher steps.

$$E[g^2]_t = 0.9E[g^2]_{t+1} + 0.1g_t^2 \quad (8)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{(1-\gamma)g_{t-1}^2 + \gamma g_t + e}} \cdot g_t \quad (9)$$

where, $g_t$ represents the moving average of squared gradients. $\gamma$ denotes the term of decay which has a range from zero to one. In this work, the RMSProp optimizer is used to define the hyperparameter concerned in the ALSTM method. Below, we define the objective function that is used to measure the MSE.

$$MSE = \frac{1}{T}\sum_{j=1}^{L}\sum_{i=1}^{M}\left(y_j^i - d_j^i\right)^2 \quad (10)$$

Here, $M$ and $L$ represent the resultant values of layer and data respectively, whereas $y_j^i$ and $d_j^i$ denote the attained and desired sizes for $j^{th}$ unit from the resultant layer of the system at time $t$ respectively.

## 4. RESULTS ANALYSIS

The literature defines and uses a number of statistical criteria to quantify the performance of BGL predictions. In this work, we used the Mean Square Error (MSE), Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), and the Mean Absolute Percentage Error (MAPE). These measurements were prevalent in the literature, particularly for predictive tasks and specifically for forecasting BG levels.

The RMSE is a reliable measure of the prediction's accuracy, since it permits the error to be expressed in units identical to the predicted quantity. Here, the error is defined as the disparity between the predicted and observed BGL values at every time point where a BGL measurement exists within the prediction time frame [30].

Consider a sequence of BG levels (denoted y) and its expected measurements ŷ with a length equal to n. The RMSE is expressed as:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (11)$$

In most cases, the scale of the error is defined in percentage terms using the MAPE, which is given by:

$$\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (12)$$

In this section, the performance analysis of the OALSTM-BGLP technique takes place on the OhioT1DM dataset.

The results are inspected under varying time horizons. Table 1 and Figure 3 represent the blood glucose prediction results of the OALSTM-BGLP technique under 15m prediction horizon (PH). These obtained result values exhibit that the OALSTM-BGLP technique displays effectual prediction under varying patient IDs. With patient ID of 540, the OALSTM-BGLP technique obtains MSE of 43.1915, RMSE of 6.5720, MAE of 3.5361, and MAPE of 0.0235. Next, based on patient ID of 552, the OALSTM-BGLP method gets MSE of 12.1690, RMSE of 3.4884, MAE of 2.3258, and MAPE of 0.0185. Also, with patient ID of 584, the OALSTM-BGLP algorithm gains MSE of 65.3590, RMSE of 8.0845, MAE of 3.6239, and MAPE of 0.0244. Meanwhile, based on patient ID of 596, the OALSTM-BGLP system achieves MSE of 18.5014, RMSE of 4.3013, MAE of 2.5968, and MAPE of 0.0201.

Table 2 and Figure 4 examine the BG prediction outcomes of the OALSTM-BGLP system at 30m PH. These experimental results showed that the OALSTM-BGLP method demonstrates effective prediction with varying patient IDs. According to patient ID 540, the OALSTM-BGLP system gains MSE of 34.5652, RMSE of 5.8792, MAE of 3.1300, and MAPE of 0.0207. Moreover, with patient ID of 552, the OALSTM-BGLP system obtains MSE of 12.0182, RMSE of 3.4667, MAE of 2.2911, and MAPE of 0.0188. Similarly, based on patient ID of 584, the OALSTM-BGLP technique achieves MSE of 67.5320, RMSE of 8.2178, MAE of 3.6693, and MAPE of 0.0252. Besides, with patient ID of 596, the OALSTM-BGLP algorithm offers MSE of 18.4980, RMSE of 4.3009, MAE of 2.5721, and MAPE of 0.0195, correspondingly.

**Table 1.** Prediction using OALSTM-BGLP under 15m PH

| Patient ID | MSE | RMSE | MAE | MAPE |
|---|---|---|---|---|
| 540 | 43.1915 | 6.5720 | 3.5361 | 0.0235 |
| 544 | 18.1007 | 4.2545 | 2.7271 | 0.0195 |
| 552 | 12.1690 | 3.4884 | 2.3258 | 0.0185 |
| 567 | 61.1774 | 7.8216 | 2.8162 | 0.0220 |
| 584 | 65.3590 | 8.0845 | 3.6239 | 0.0244 |
| 596 | 18.5014 | 4.3013 | 2.5968 | 0.0201 |

**Table 2.** Prediction outcomes of OALSTM-BGLP at 30m PH

| Patient ID | MSE | RMSE | MAE | MAPE |
|---|---|---|---|---|
| 540 | 34.5652 | 5.8792 | 3.1300 | 0.0207 |
| 544 | 18.5069 | 4.3020 | 2.5005 | 0.0174 |
| 552 | 12.0182 | 3.4667 | 2.2911 | 0.0188 |
| 567 | 60.1172 | 7.7535 | 3.2886 | 0.0272 |
| 584 | 67.5320 | 8.2178 | 3.6693 | 0.0252 |
| 596 | 18.4980 | 4.3009 | 2.5721 | 0.0195 |

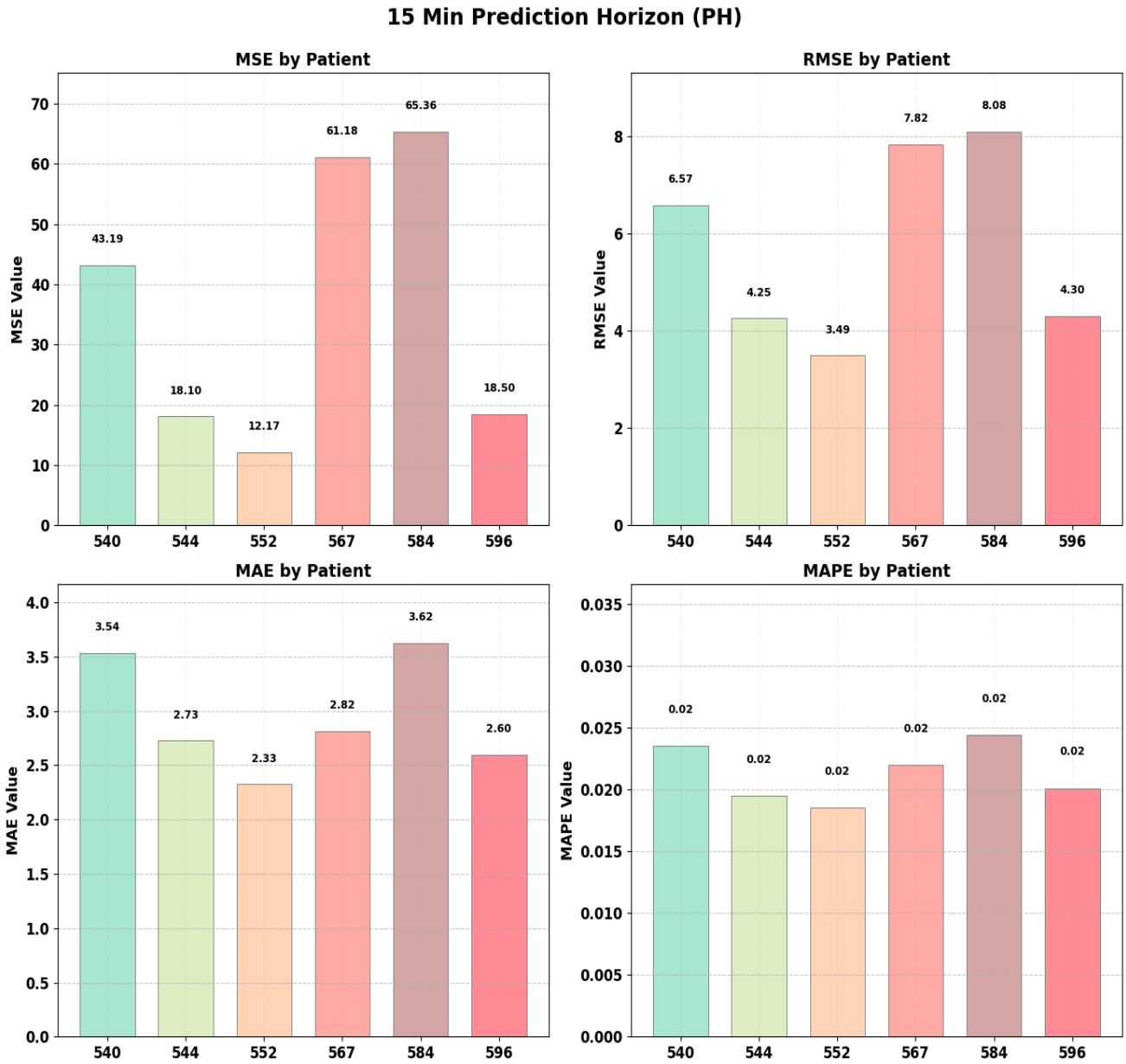



**Figure 3.** Prediction results of the OALSTM-BGLP model under 15m PH (a-b) MSE and RMSE (c-d) MAE and MAPE

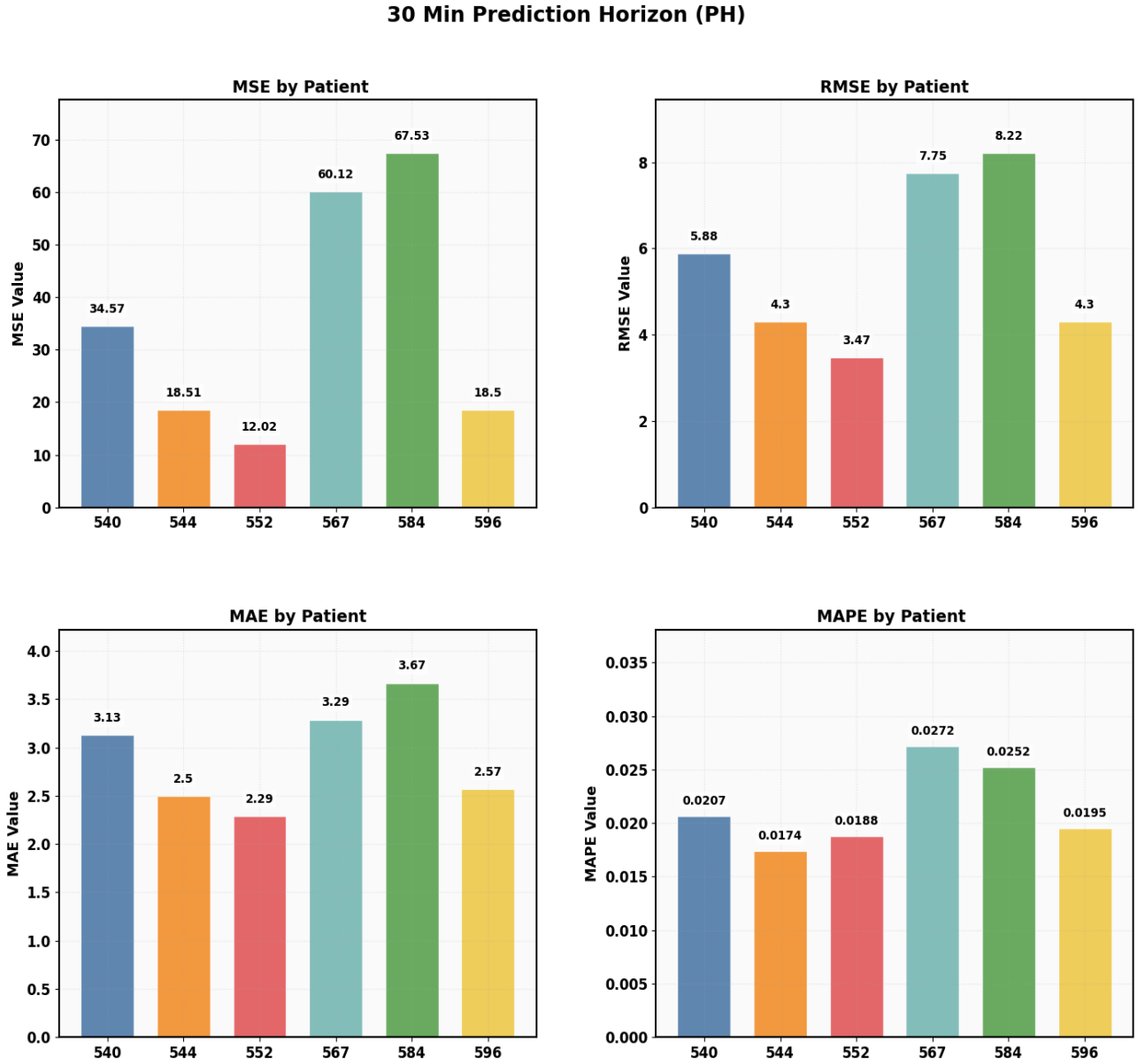



**Figure 4.** Prediction from the OALSTM-BGLP system at 30m PH (a-b) MSE and RMSE (c-d) MAE and MAPE

Table 3 and Figure 5 display the blood glucose prediction analysis of the OALSTM-BGLP method with 45m PH. These experimentation outcomes indicated that the OALSTM-BGLP algorithm exhibits proficient prediction at varying patient IDs. Based on patient ID 540, the OALSTM-BGLP system provides MSE of 36.2906, RMSE of 6.0242, MAE of 3.1425, and MAPE of 0.0212**.**

Additionally, with patient ID of 552, the OALSTM-BGLP approach gets MSE of 12.1974, RMSE of 3.4925, MAE of 2.3315, and MAPE of 0.0189. Likewise, based on patient ID of 584, the OALSTM-BGLP method acquires MSE of 68.5057, RMSE of 8.2768, MAE of 3.8021, and MAPE of 0.0260. Finally, with patient ID of 596, the OALSTM-BGLP system accomplishes MSE of 18.1017, RMSE of 4.2546, MAE of 2.5421, and MAPE of 0.0194.

Table 4 and Figure 6 demonstrate the blood glucose prediction outcomes of the OALSTM-BGLP system under 60m PH. These achieved findings denoted that the OALSTM-BGLP algorithm shows successful prediction with varying patient IDs. For the patient ID of 540, the OALSTM-BGLP method gains MSE of 39.2007, RMSE of 6.2610, MAE of 3.3479, and MAPE of 0.0227. Meanwhile, with patient ID of 552, the OALSTM-BGLP approach gets MSE of 14.0294, RMSE of 3.7456, MAE of 2.5598, and MAPE of 0.0209. Moreover, based on patient ID of 584, the OALSTM-BGLP technique provides MSE of 68.3196, RMSE of 8.2656, MAE of 3.6609, and MAPE of 0.0253. In conclusion, with patient ID of 596, the OALSTM-BGLP method achieves MSE of 20.3714, RMSE of 4.5135, MAE of 2.6548, and MAPE of 0.0202.

Table 5 and Figure 7 present a brief average result of the OALSTM-BGLP technique on varying time horizon. These experimentation findings implied that the OALSTM-BGLP technique exhibits better predictive outcomes under all horizon time intervals. With horizon time of 15m, the OALSTM-BGLP technique offers MSE of 36.4165, RMSE of 5.7537, MAE of 2.9377, and MAPE of 0.0213. Additionally, based on horizon time of 30m, the OALSTM-BGLP method gives MSE of 35.2062, RMSE of 5.6533, MAE of 2.9086, and MAPE of 0.0215. Followed by horizon time of 45m, the OALSTM-BGLP algorithm provides MSE of 36.7045, RMSE of 5.7611, MAE of 2.9926, and MAPE of 0.0222. At last, with horizon time of 60m, the OALSTM-BGLP method gains MSE of 37.5874, RMSE of 5.9081, MAE of 3.0980, and MAPE of

0.0231, respectively.

Figure 8 compares actual against predicted values of the OALSTM-BGLP system using the training dataset. The figure reported that the OALSTM-BGLP method properly predicts the blood glucose levels. It is also noticed that the predicted blood glucose levels by the OALSTM-BGLP technique are much nearer to actual values.

Figure 9 compares the actual and predicted values of the OALSTM-BGLP method on the testing dataset. This figure pointed out that the OALSTM-BGLP system appropriately predicts blood glucose levels. This can be perceived that the predicted blood glucose levels by the OALSTM-BGLP algorithm is highly close to actual values.

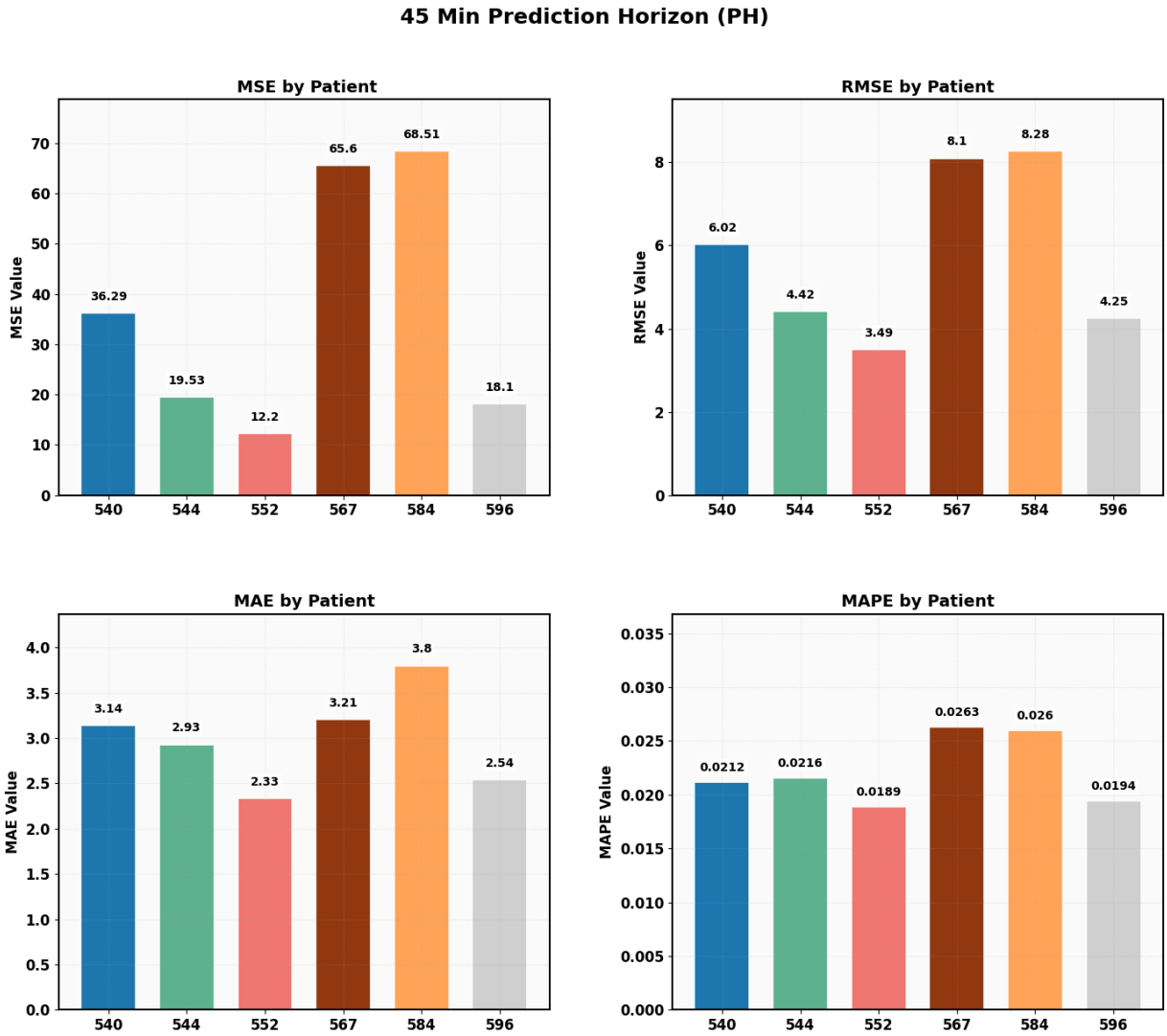



**Figure 5.** Prediction of the OALSTM-BGLP model under 45m PH (a-b) MSE and RMSE (c-d) MAE and MAPE

**Table 3.** Prediction outcomes of the OALSTM-BGLP method on 45-min PH

| Patient ID | MSE | RMSE | MAE | MAPE |
|---|---|---|---|---|
| 540 | 36.2906 | 6.0242 | 3.1425 | 0.0212 |
| 544 | 19.5290 | 4.4192 | 2.9300 | 0.0216 |
| 552 | 12.1974 | 3.4925 | 2.3315 | 0.0189 |
| 567 | 65.6025 | 8.0995 | 3.2077 | 0.0263 |
| 584 | 68.5057 | 8.2768 | 3.8021 | 0.0260 |
| 596 | 18.1017 | 4.2546 | 2.5421 | 0.0194 |

**Table 4.** Prediction outcomes of the OALSTM-BGLP method at 60m PH

| Patient ID | MSE | RMSE | MAE | MAPE |
|---|---|---|---|---|
| 540 | 39.2007 | 6.2610 | 3.3479 | 0.0227 |
| 544 | 25.2215 | 5.0221 | 3.1173 | 0.0225 |
| 552 | 14.0294 | 3.7456 | 2.5598 | 0.0209 |
| 567 | 58.3816 | 7.6408 | 3.2471 | 0.0268 |
| 584 | 68.3196 | 8.2656 | 3.6609 | 0.0253 |
| 596 | 20.3714 | 4.5135 | 2.6548 | 0.0202 |

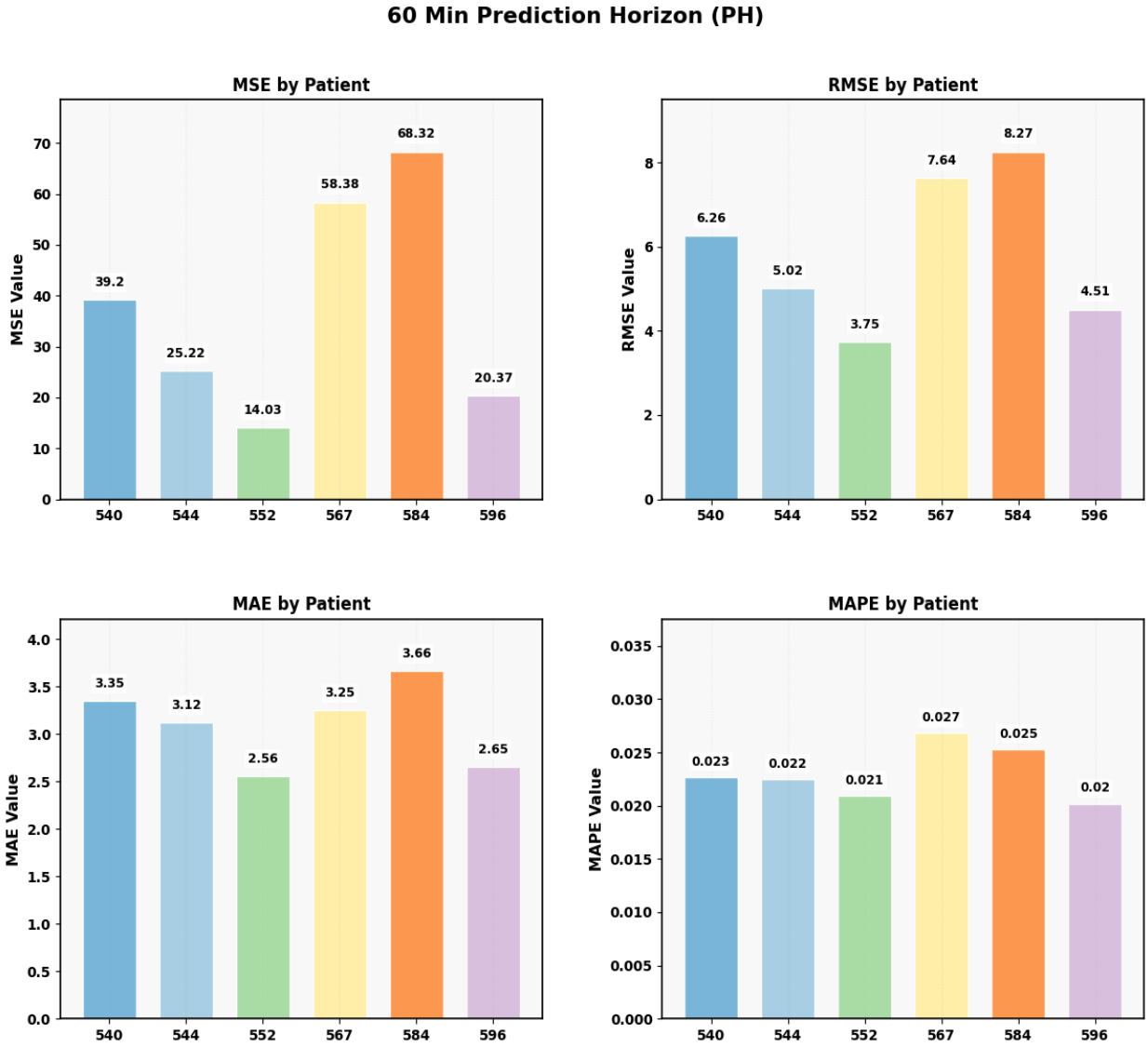



**Figure 6.** Prediction outcome of the OALSTM-BGLP model under 60 PH (a-b) MSE and RMSE (c-d) MAE and MAPE

**Table 5.** Average PH of OALSTM-BGLP method under varying PH

| PH(m) | MSE | RMSE | MAE | MAPE |
|---|---|---|---|---|
| 15 | 36.4165 | 5.7537 | 2.9377 | 0.0213 |
| 30 | 35.2062 | 5.6533 | 2.9086 | 0.0215 |
| 45 | 36.7045 | 5.7611 | 2.9926 | 0.0222 |
| 60 | 37.5874 | 5.9081 | 3.0980 | 0.0231 |

Figure 10 presents the prediction concentration results of the OALSTM-BGLP technique under varying time horizons. These results pointed out that the OALSTM-BGLP technique accomplishes better performance with enhanced prediction concentration results under all-time horizons.

The results demonstrate low error values for the numerical metrics MSE, RMSE, MAE, and MAPE, indicating a significant enhancement in prediction accuracy. Precision is crucial for effective diabetes management, as accurate predictions minimise the discrepancy between predicted and actual glucose levels. This approach enhances decision-making speed by reducing errors, and aiding individuals and healthcare professionals in effectively predicting and managing hypo- or hyperglycaemic episodes. This enhances patient safety by avoiding overcorrections and facilitates personalised treatment approaches that are adapted to individual glucose variability. This increased accuracy enables healthcare providers and patients to make informed decisions about insulin dosage, dietary changes, and lifestyle modifications. Thus, low error metrics indicate the model's robustness.

Numerical metrics are useful for evaluating model performance; however, they do not always provide a comprehensive understanding of real-world impact. Clarke's Error Grid Analysis (CEGA) is commonly utilised to evaluate the clinical reliability of blood glucose level predictions [31]. CEGA categorizes predictions into five regions: A, B, C, D, and E, concerning the clinical outcome of insulin dosing based on the predicted BG level. The most unfavorable scenario (D and E regions) involves an excessively high BG level prediction, which may result in hypoglycemia, a critical emergency condition. Consequently, we can categorise the identical absolute numerical error into different regions based on the actual BGL range and the sign of the error. A predicted

value is considered 'clinically acceptable' if it falls within either the A or B region. CEGA modifies the grid such that the 'accurate' domain matches the 'clinically acceptable' categorisation. We have generated Clarke error grids for prediction horizons at 15m, 30m, 45m, and 60m to analyze the clinical implications of the OALSTM-BGLP model. Figure 10 prominently displays the CEGA at 15m, 30m, 45m, and 60m prediction horizons, the majority of the data points also lie in region A, reinforcing the reliability and clinical relevance of our model's predictions across different time intervals. This consistent performance highlights the potential of the OALSTM-BGLP model in real-world applications for BG level prediction.

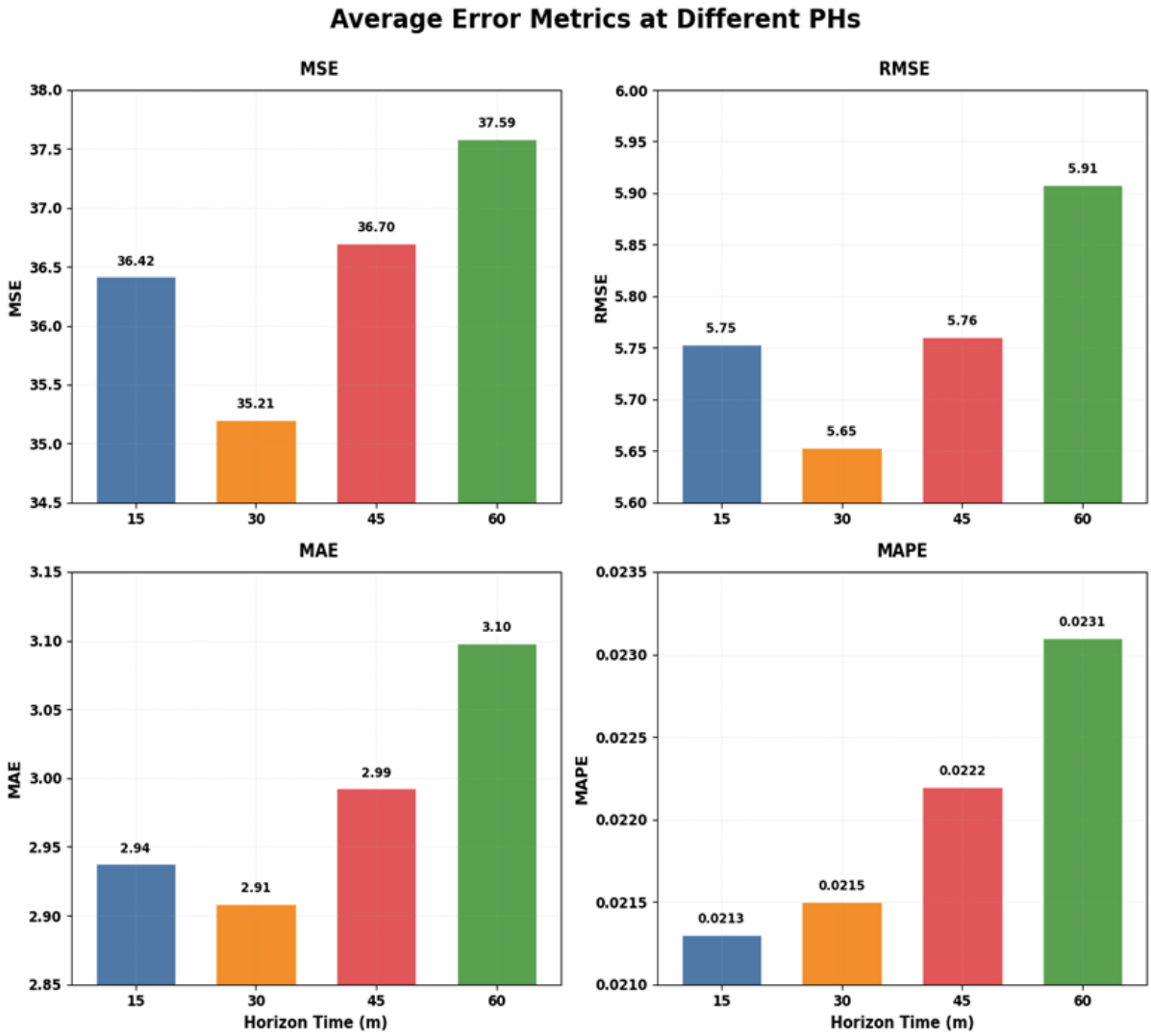



**Figure 7.** Average PH of OALSTM-BGLP model (a-b) MSE and RMSE (c-d) MAE and MAPE

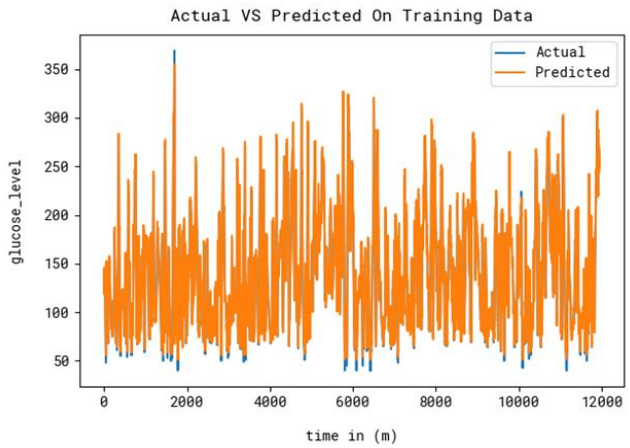



**Figure 8.** Comparison of actual and predicted glucose values using the OALSTM-BGLP model on the training dataset

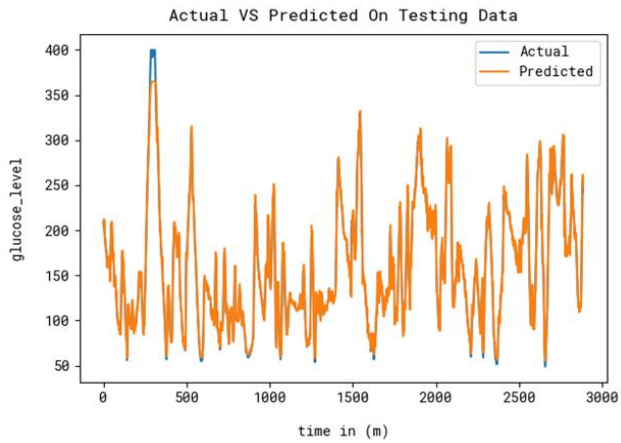



**Figure 9.** Comparison of actual and predicted values of the OALSTM-BGLP model under testing dataset

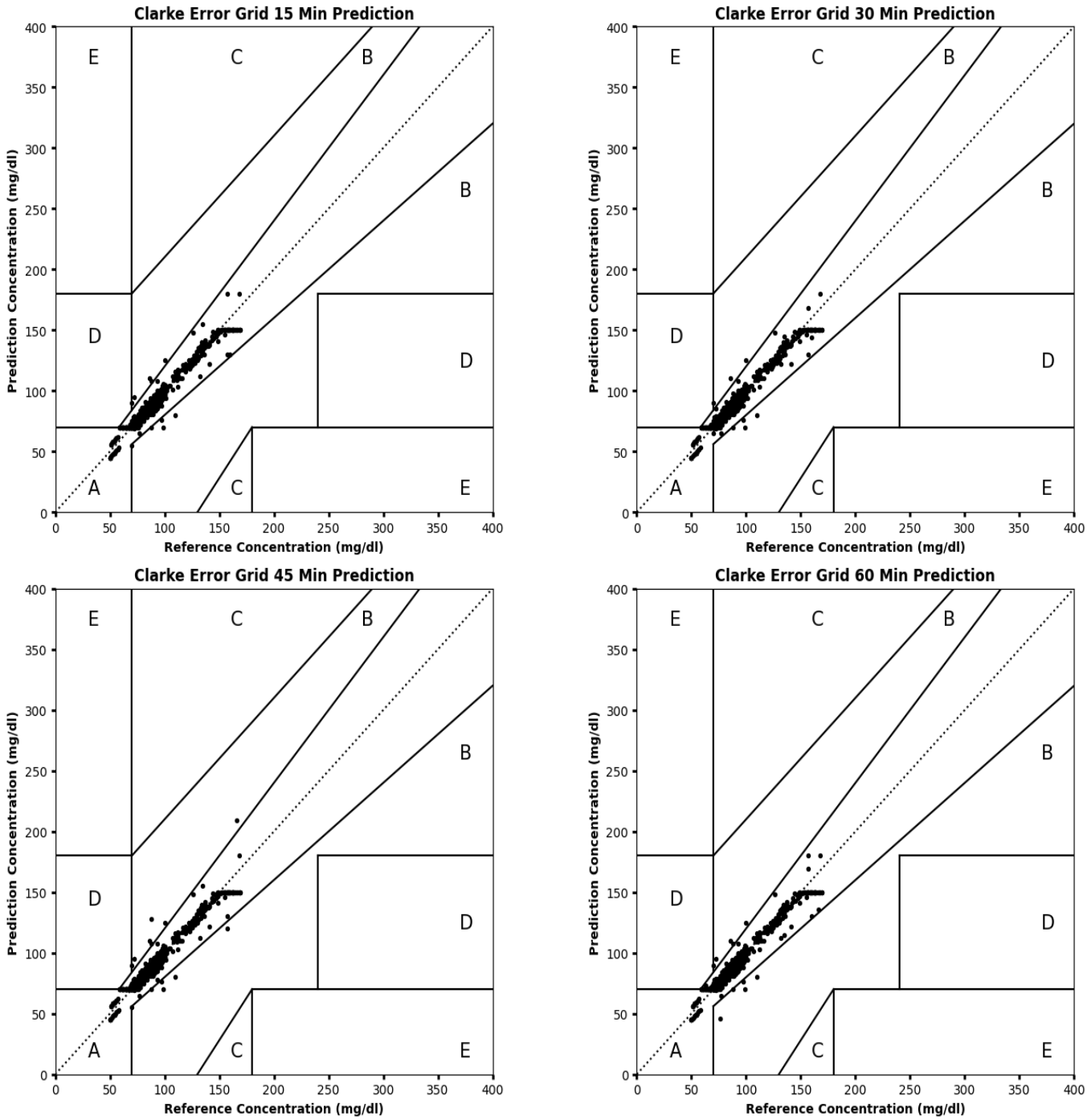



**Figure 10.** Clarke error grid analysis at different prediction horizons: (a) 15m, (b) 30m, (c) 45m, and (d) 60m

## 5. DISCUSSION

The new OALSTM-BGLP model is designed for predicting BG levels by leveraging historical CGM data across four critical prediction intervals: 15, 30, 45, and 60 minutes. The selection of time intervals for BG level prediction is determined by clinical significance and the properties of the data. Glucose levels exhibit rapid fluctuations, particularly during mealtime. Shorter intervals, such as 15 minutes, are clinically significant for monitoring rapid glucose fluctuations, facilitating prompt interventions for insulin adjustments. While intermediate intervals (30–60 m) provide valuable insights into broader trends, they are beneficial for general glucose management, including meal planning, physical activity, and insulin administration. Additionally, the sampling frequency of CGM data (every 5 m) influences the selected PH, balancing model accuracy with real-time applicability. This approach ensures the model aligns with both clinical needs and data constraints.

Our results demonstrate that the proposed model outperformed various recent DL-based approaches in terms of predictive accuracy. While direct comparison with existing studies is feasible for the 30m and 60m predictions, a noticeable gap exists in the recent literature for 15m and 45m horizons. Table 6 illustrates its performance in comparison to new models tested on various versions of the OhioT1DM dataset.

**Table 6.** Comparison of recent literature utilizing the OhioT1DM Dataset with the OALSTM-BGLP model

| Study | Model | Patients | PHs | RMSE [mg/dL] |
|---|---|---|---|---|
| [32] | RNN | 6 | 30m | 22.43 |
| | | | 60m | 36.39 |
| [33] | Deep Residual Forecasting Architecture | 6 | 30m | 18.2 |
| | | | 60m | 31.7 |
| [34] | GluNet | 6 | 30m | 19.28 to 22.93 and |
| | | | 60m | 31.83 to 39.53 |
| [35] | Multilayer CRNN | 6 | 30m | 17.45 |
| | | | 60m | 33.67 |
| [36] | DCNN, Seq-to-Seq LSTM, MLR, BRC | 12 | 30m | 17.52 |
| | | | 60m | 24.58 |
| [21] | E3NN | 12 | 30m | 18.92 |
| | | | 60m | 32.54 |
| [30] | LSTM+WaveNet+GRU | 12 | 30m | 21.90, |
| | | | 45m | 29.12, |
| | | | 60m | 35.10 |
| [37] | Autonomous Channel DL Framework | 12 | 30m | 16.77 to 21.08 and |
| | | | 60m | 28.36 to 35.22 |
| **Proposed** | **OALSTM-BGLP** | **6** | **15m** | **5.75** |
| | | | **30m** | **5.65** |
| | | | **45m** | **5.76** |
| | | | **60m** | **5.91** |

Martinsson et al. [32] proposed an approach based on RNNs trained in an end-to-end fashion that required nothing but the glucose level history for the patient and predicted the BG on 30 m and 60 m PHs. The approach achieved the RMSE values of 22.43 and 36.39 for 30 and 60m PHs, respectively. Falcone et al. used data from six patients from an updated version of the OhioT1DM dataset in their study to suggest a new deep residual time series forecasting architecture. This architecture modified N-BEATS by incorporating a bidirectional LSTM network into each block in place of fully connected layers. The method included supplementary input variables, including bolus insulin and carbohydrates, and introduced three additional loss terms. The proposed deep residual forecasting method achieved average RMSEs of 18.2 for the 30 m glucose forecasting interval and 31.7 for the 60 m interval. Li et al. [34] introduced GluNet, a glucose forecasting method, utilising data from six patients in the Ohio T1DM 2018 dataset and employing deep neural networks. The findings indicated that GluNet attained RMSE values between 19.28 mg/dl and 22.93 mg/dl for 30 m PH, whereas for 60 m PH, RMSE values varied from 31.83 mg/dl to 39.53 mg/dl. Freiburghaus et al. [35] introduced a multilayer CRNN that employs multivariate data from six patients in the OhioT1DM dataset. The RMSE obtained was 17.45 for a 30 m PH and 33.67 for a 60 m PH.

Zhang et al. [36] designed four different models, including deep learning algorithms and regression models, and tested how well they could predict BG levels within 30 or 60m PH using data from 12 people with OhioT1DM. The two deep learning models are the DCNN and the Seq-to-Seq LSTM model. The two regression models are a MLR model and a BRC model. The Seq-to-Seq LSTM model exhibited superior performance in 30m ahead predictions with a 17.52 Average. RMSE, whereas the MLR model excelled in 60m-ahead predictions with a 24.58 avg. RMSE. Zhu et al. [21] proposed a GRU-based RNN model, E3NN, that incorporates an attention mechanism and evidential regression. The model was developed and evaluated on three datasets, including OhioT1DM (12 subjects). The model scored RMSE 18.92 on 30 m PH and 32.54 on 60 m PH. Dudukcu et al. [30] proposed a fusion of LSTM, WaveNet, and GRU models. Experiments were done with data from 12 patients with OhioT1DM to make predictions on 30m, 45m, and 60m PH, as well as 21.90, 29.12, and 35.10 Avg. RMSE was achieved, respectively. Yang et al. [37] developed an autonomous channel deep learning framework. Researchers tested the framework on data from 12 patients in the clinical OhioT1DM dataset. The study's experiments revealed an RMSE range of 16.775 to 21.085 on 30 m PH and 28.36 to 35.22 on 60 m PH.

In this study, we predict BG levels at 15, 30, 45, and 60 m into the future, offering a more granular and comprehensive analysis compared to most recent studies, which primarily focus on 30 and 60 m horizons. By including shorter (15 m) and intermediate (45 m) prediction intervals, our approach captures more immediate and nuanced glucose dynamics, which are critical for timely interventions and effective diabetes management. Our RMSE results demonstrate robust performance across all horizons.

## 6. CONCLUSIONS

In this study, we have introduced the OALSTM-BGLP methodology, which combines Min-Max normalization, an attention-based LSTM model, and RMSProp optimizer-based parameter tuning for accurate BG level prediction. The OALSTM-BGLP technique initially normalises the input data through Min-Max scaling, facilitating consistent and meaningful comparisons among different features. The attention mechanism in the ALSTM model enables the model to concentrate on the most pertinent information within the input sequence, while also effectively capturing long-term dependencies. This enhances the model's capacity to extract significant features, thereby improving predictive accuracy. The RMSProp optimiser modifies the learning rate in accordance with the magnitude of recent gradients, thereby enhancing training efficiency and accelerating convergence. The assessment of the proposed method using the OhioT1DM dataset indicates its superior performance, surpassing recent state-of-the-art techniques regarding accuracy. The enhanced predictive capability can substantially assist in real-time diabetes management, enabling patients and healthcare providers to make timely and informed decisions. Potential

real-world applications encompass integration into personalised diabetes management systems, mobile health applications, and continuous glucose monitoring devices, where it can deliver actionable insights to avert hyper/hypoglycemic events.

Future research could look into the inclusion of additional variables, including insulin dosage, physical activity, and meal intake, to improve prediction accuracy. Furthermore, examining the model's incorporation into wearable devices for continuous, real-time monitoring could facilitate advancements in automated diabetes management. Future research should examine the model's performance across diverse patient populations to verify its robustness and applicability in different clinical contexts.

## REFERENCES


[1] Antoniou, M., Mateus, C., Hollingsworth, B., Titman, A. (2024). A systematic review of methodologies used in models of the treatment of diabetes mellitus. Pharmacoeconomics, 42(1): 19-40. https://doi.org/10.1007/s40273-023-01312-4

[2] Yothapakdee, K., Charoenkhum, S., and Boonnuk, T. (2022). Improving the efficiency of machine learning models for predicting blood glucose levels and diabetes risk. Indonesian Journal of Electrical Engineering and Computer Science, 27(1): 555-562. http://doi.org/10.11591/ijeecs.v27.i1.pp555-562

[3] Li, L., Ma, Y., Jiang, J., Liu, Z., Ye, Z., Liu, S., Pu, C., Chen, C.S., Wan, Y. (2023). Machine learning models for blood glucose prediction in patients with Diabetes Mellitus: A systematic review and network meta-analysis. JMIR Medical Informatics, 11: e47833. https://doi.org/10.2139/ssrn.4401684

[4] Arora, S., Kumar, S., Kumar, P. (2023). Multivariate models of blood glucose prediction in type1 diabetes: A survey of the state-of-the-art. Current Pharmaceutical Biotechnology, 24(4): 532-552. https://doi.org/10.2174/1389201023666220603092433

[5] Ahmed, A., Aziz, S., Abd-alrazaq, A., Farooq, F., Househ, M., Sheikh, J. (2023). The effectiveness of wearable devices using artificial intelligence for blood glucose level forecasting or prediction: Systematic review. Journal of Medical Internet Research, 25: e40259. https://doi.org/10.2196/40259

[6] Saha, P., Marouf, Y., Pozzebon, H., Guergachi, A., Keshavjee, K., Noaeen, M., Shakeri, Z. (2024). Predicting time to diabetes diagnosis using random survival forests. MedRxiv. https://doi.org/10.1101/2024.02.03.24302304

[7] Khadem, H., Nemat, H., Elliott, J., Benaissa, M. (2023). Blood glucose level time series forecasting: nested deep ensemble learning lag fusion. Bioengineering, 10(4): 487. https://doi.org/10.3390/bioengineering10040487

[8] Langarica, S., Rodriguez-Fernandez, M., Núñez, F., Doyle III, F.J. (2023). A meta-learning approach to personalized blood glucose prediction in type 1 diabetes. Control Engineering Practice, 135: 105498. https://doi.org/10.1016/j.conengprac.2023.105498

[9] Eren-Oruklu, M., Cinar, A., Quinn, L. (2010). Hypoglycemia prediction with subject-specific recursive time-series models. Journal of Diabetes Science and Technology, 4(1): 25-33. http://doi.org/10.1177/193229681000400104

[10] Turksoy, K., Bayrak, E.S., Quinn, L., Littlejohn, E., Cinar, A. (2013). Multivariable adaptive closed-loop control of an artificial pancreas without meal and activity announcement. Diabetes Technology and Therapeutics, 15(5): 386-400. http://doi.org/10.1089/dia.2012.0283

[11] Georga, E.I., Protopappas, V.C., Polyzos, D., Fotiadis, D.I. (2015). Evaluation of short-term predictors of glucose concentration in type 1 diabetes combining feature ranking with regression models. Medical and Biological Engineering and Computing, 53(12): 1305-1318. http://doi.org/10.1007/s11517-015-1263-1

[12] Hamdi, T., Ben Ali, J., Di Costanzo, V., Fnaiech, F., Moreau, E., Ginoux, J.M. (2018). Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. Biocybernetics and Biomedical Engineering, 38(2): 362-372. http://doi.org/10.1016/j.bbe.2018.02.005

[13] Fox, I., Ang, L., Jaiswal, M., Pop-Busui, R., Wiens, J. (2018). Deep multi-output forecasting: learning to accurately predict blood glucose trajectories. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp.1387-1395. http://doi.org/10.1145/3219819.3220102

[14] Sun, Q., Jankovic, M.V., Bally, L., Mougiakakou, S.G. (2018). Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network. In 2018 14th Symposium on Neural Networks and Applications (NEUREL), pp. 1-5. http://doi.org/10.1109/NEUREL.2018.8586990

[15] Zhang, Z., Lu, Y., Ye, M., Huang, W., Jin, L., Zhang, G., Ge, Y., Baghban, A., Zhang, Q., Wang, H., Zhu, W. (2024). A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis. Journal of King Saud University-Computer and Information Sciences, 36(1): 101873. https://doi.org/10.1016/j.jksuci.2023.101873

[16] Yang, G., Liu, S., Li, Y., He, L. (2023). Short-term prediction method of blood glucose based on temporal multi-head attention mechanism for diabetic patients. Biomedical Signal Processing and Control, 82: 104552. https://doi.org/10.1016/j.bspc.2022.104552

[17] Butt, H., Khosa, I., Iftikhar, M.A. (2023). Feature transformation for efficient blood glucose prediction in type 1 diabetes mellitus patients. Diagnostics, 13(3): 340. https://doi.org/10.3390/diagnostics13030340

[18] Langarica, S., Rodriguez-Fernandez, M., Doyle, F.J., Núñez, F. (2023). A probabilistic approach to blood glucose prediction in type 1 diabetes under meal uncertainties. IEEE Journal of Biomedical and Health Informatics, 27(10): 5054-5065. https://doi.org/10.1109/JBHI.2023.3309302

[19] Song, S., Wang, Q., Zou, X., Li, Z., Ma, Z., Jiang, D., Fu, Y., Liu, Q. (2023). High-precision prediction of blood glucose concentration utilizing Fourier transform Raman spectroscopy and an ensemble machine learning algorithm. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 303: 123176. https://doi.org/10.1016/j.saa.2023.123176

[20] Ye, Z., Wang, J., Hua, H., Zhou, X., Li, Q. (2022). Precise detection and quantitative prediction of blood glucose level with an electronic nose system. IEEE Sensors Journal, 22(13): 12452-12459. https://doi.org/10.1109/JSEN.2022.3178996

[21] Zhu, T., Kuang, L., Daniels, J., Herrero, P., Li, K., Georgiou, P. (2022). IoMT-enabled real-time blood glucose prediction with deep learning and edge computing. IEEE Internet of Things Journal, 10(5): 3706-3719. https://doi.org/10.1109/JIOT.2022.3143375
[22] Lu, X., Song, R. (2022). A hybrid deep learning model for blood glucose prediction. In 2022 IEEE 11th Data Driven Control and Learning Systems Conference (DDCLS), pp. 1037-1043. https://doi.org/10.1109/DDCLS55054.2022.9858348
[23] Arora, S., Kumar, S., Kumar, P. (2025). Forecasting blood glucose level using convolutional recurrent connection. Sādhanā, 50: 16. https://doi.org/10.1007/s12046-024-02657-y
[24] Raju, K.S., Govardhan, A., Rani, B.P., Sridevi, R., Murty, M.R. (2020). Proceedings of the Third International Conference on Computational Intelligence and Informatics: ICCII 2018. Springer Nature.
[25] Marling, C., Bunescu, R. (2020). The OhioT1DM dataset for blood glucose level prediction: Update 2020. CEUR Workshop Proceedings, 2675: 71-74.
[26] Kang, Q., Chen, E.J., Li, Z.C., Luo, H.B., Liu, Y. (2023). Attention-based LSTM predictive model for the attitude and position of shield machine in tunneling. Underground Space, 13: 335-350. https://doi.org/10.1016/j.undsp.2023.05.006
[27] Pathak, D., Kashyap, R. (2023). Neural correlate-based e-learning validation and classification using convolutional and long short-term memory networks. Traitement du Signal, 40(4): 1457-1467. https://doi.org/10.18280/ts.400414
[28] Li, C., Cai, Y., Li, Y., Zhang, P. (2024). Fusion of dual sensor features for fall risk assessment with improved attention mechanism. Traitement du Signal, 41(1): 73-83. https://doi.org/10.18280/ts.410106
[29] Yaqub, M., Feng, J., Zia, M.S., Arshid, K., Jia, K., Rehman, Z.U., Mehmood, A. (2020). State-of-the-art CNN optimizer for brain tumor segmentation in magnetic resonance images. Brain Sciences, 10(7): 427. https://doi.org/10.3390/brainsci10070427
[30] Dudukcu, H.V., Taskiran, M., Yildirim, T. (2021). Blood glucose prediction with deep neural networks using weighted decision level fusion. Biocybernetics and Biomedical Engineering, 41(3): 1208-1223. https://doi.org/10.1016/j.bbe.2021.08.007
[31] Ali, J.B., Fnaiech, F., Saad, N., Najeh, I., Fnaiech, N., Burrus, N. (2018). Continuous blood glucose level prediction of type 1 diabetes based on artificial neural network. Biocybernetics and Biomedical Engineering, 38(4): 828-840. https://doi.org/10.1016/j.bbe.2018.08.002
[32] Martinsson, J., Schliep, A., Eliasson, B., Mogren, O. (2020). Blood glucose prediction with variance estimation using recurrent neural networks. Journal of Healthcare Informatics Research, 4: 1-18. https://doi.org/10.1016/j.bbe.2018.06.005
[33] Rubin-Falcone, H., Fox, I., Wiens, J. (2020). Deep residual time-series forecasting: Application to blood glucose prediction. KDH@ECAI, 20: 105-109.
[34] Li, K., Liu, C., Zhu, T., Herrero, P., Georgiou, P. (2019). GluNet: A deep learning framework for accurate glucose forecasting. IEEE Journal of Biomedical and Health Informatics, 24(2): 414-423. https://doi.org/10.1109/JBHI.2019.2929398
[35] Freiburghaus, J., Rizzotti, A., Albertetti, F. (2020). A deep learning approach for blood glucose prediction of type 1 diabetes. In Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with the 24th European Conference on Artificial Intelligence (ECAI 2020), pp. 131-135. https://ceur-ws.org/Vol-2675/paper23.pdf.
[36] Zhang, M., Flores, K.B.,Tran, H.T. (2021). Deep learning and regression approaches to forecasting blood glucose levels for type 1 diabetes. Biomedical Signal Processing and Control, 69: 102923. https://doi.org/10.1016/j.bspc.2021.102923
[37] Yang, T., Yu, X., Ma, N., Wu, R., Li, H. (2022). An autonomous channel deep learning framework for blood glucose prediction. Applied Soft Computing, 120: 108636. https://doi.org/10.1016/j.asoc.2022.108636

## NOMENCLATURE

| | |
|---|---|
| BG | Blood Glucose |
| BRC | Bidirectional Reservoir Computing |
| CGM | Continuous Glucose Monitoring |
| DL | Deep Learning |
| DCNN | Dilated Convolutional Neural Network |
| ML | Machine Learning |
| PH | Prediction Horizon |
| NN | Neural Networks |
| RNN | Recurrent Neural Network |
| GRNN | Gated Recurrent Neural Network |
| LSTM | Long Short Term Memory |
| AM | Attention Mechanism |
| ALSTM | Attention based Long Short Term Memory |
| SGD | Stochastic Gradient Descent |
| MSE | Mean Square Error |
| RMSE | Root Mean Square Error |
| MAE | Mean Absolute Error |
| MAPE | Mean Absolute Percentage Error |
| MLR | Multiple Linear Regression |

### Greek Symbols

| | |
|---|---|
| $\alpha$ | Weight |
| $\sigma$ | Function of Logistics |
| $\Upsilon$ | Decay Term |
| $\eta$ | Learning Rate |
| $\theta$ | Parameter value w.r.t time |

### Subscripts

| | |
|---|---|
| A | Attribute |
| t | Time |

Copyright of Traitement du Signal is the property of International Information & Engineering Technology Association (IIETA) and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Physica Scripta

INCLUSIVE PUBLISHING TRUSTED SCIENCE

PURPOSE-LED PUBLISHING™

PAPER

# An energy and cost-effective realization of high-performance single-layer vedic multiplier in quantum-dot cellular automata

To cite this article: Hemanshi Chugh and Sonal Singh 2025 *Phys. Scr.* **100** 065112

View the article online for updates and enhancements.

You may also like

- Comparative Analysis of Vedic Multiplier using Various Adder Architectures
Jatin Yadav, Anupam Kumar, Shaik Shareef et al.
- Area and power efficient FIR filter design in quantum cellular automata using competent adder
Kamaraj Arunachalam, Divya Bharathi Jaya Baskar and Vishnu Chithra Ramaraj
- Design of High Performance ALU Using Vedic Mathematics
V Swathi, Kavitha Panduga and Gurrala Shiva Kumari

This content was downloaded from IP address 14.139.251.98 on 20/05/2025 at 08:57

Physica Scripta

PAPER

# An energy and cost-effective realization of high-performance single-layer vedic multiplier in quantum-dot cellular automata

**Hemanshi Chugh and Sonal Singh**

Department of Electronics and Communication Engineering, Delhi Technological University, Bawana Road, Delhi, India

**E-mail:** sonalsingh@dtu.ac.in

**RECEIVED**
20 December 2024

**REVISED**
24 April 2025

**ACCEPTED FOR PUBLICATION**
30 April 2025

**PUBLISHED**
16 May 2025

**Keywords:** vedic multiplier, QCA, average output polarization (AOP)

## Abstract

Quantum-dot cellular automata (QCA) is a promising nanotechnology offering numerous advantages over traditional transistor-dependent technology, featuring lower energy requirements and faster operation. This work delves into developing an optimized QCA-based 2-bit Vedic multiplier essential in arithmetic operations. The suggested multiplier is designed using cell interaction-based fault-tolerant half-adder circuit and majority gates within a co-planar architecture. An extensive defect analysis underscores the half-adder circuit's resilience for larger circuits. The suggested design is comprehensively evaluated using the QCADesigner-E and QCAPro software on several design parameters, such as area, quantum cost, complexity and energy dissipation. Furthermore, the impact of temperature on output polarization is thoroughly examined. The simulation results revealed remarkable improvements in several key circuit metrics. The proposed 2-bit multiplier exhibits a notable 46.2% decrease in latency and a remarkable 90.7% reduction in quantum cost, with reduced crossovers compared to the closest reported design. The suggested 4-bit Vedic multiplier also outperforms several existing QCA-based multipliers, offering a highly efficient and compact design. These significant enhancements are attributed to the optimized cell-interaction mechanism in QCA technology, showcasing its potential for efficient and high-performance circuit design, as underscored in this research's findings. The superior metrics and reduced fabrication complexity of the suggested multiplier facilitate its practical implementation in various nanoelectronics applications.

## 1. Introduction

The persistent pursuit of miniaturization in semiconductor technology has driven the continuous scaling of transistors, leading to the advent of nano-scale electronics. However, on the nanoscale, traditional CMOS technology encounters inherent physical limitations due to the decreasing gate length. These limitations lead to several short-channel effects, including an increased leakage current, reduced transconductance, and increased susceptibility to noise. These detrimental effects undermine the fundamental characteristics of transistors, leading to device instability, increased power consumption, and hindered performance [1]. In response to these challenges, alternative nanotechnologies such as spin electron transistors, nanowire transistors, and quantum-dot cellular automata (QCA) have been proposed. The QCA technology has gained recognition as an innovative and energy-efficient technology offering significant benefits at the nanoscale including enhanced device density and operational switching frequencies [2]. In turn, to ensure robustness and reliability, we have employed multiple simulation tools such as QCA-Designer v2.0.3, QCA Designer-E v2.2, and QCAPro to validate our designs.

Recent developments in QCA have seen the integration of highly efficient arithmetic operation designs, including adders and multipliers [3–5]. The optimization of multipliers primarily revolves around reducing the carry propagation delay. Concurrently with these advancements, researchers have explored various implementation strategies for Vedic multipliers (VM) in QCA technology, such as sutras-based designs like 'Urdhva Tiryakbhyam (UT)', 'Nikhilam' and 'Anurupayana' sutra. Initially in 2016, Chudasama and Sasamal [6] realized a multi-layer 4×4 VM in QCA, employing the carry-save adder technique to reduce the computation

© 2025 IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

time. Their simulation results demonstrated significant improvements over the conventional 4×4 Wallace and Dadda multipliers, achieving 30%, 60% and 50% reduction in cell density, footprint and latency respectively. In 2019, Pooja *et al* [7] introduced a layered architecture of 2-bit VM incorporating 291 standard cells, and footprint of 0.4 $\mu m^2$. Simultaneously, a 4-bit VM employing 2-bit multipliers, half adders (HA), and full adders was also presented comprising of 1959 cells in a footprint of 2.39 $\mu m^2$. It is noteworthy that the use of multiple adders contributed to an increase in both cell count and area requirements. Further, in 2019, Gassoumi *et al* [8] presented a novel FIR filter design utilizing a co-planar Vedic multiplier. This VM design exhibits 1683 cells within a footprint of 8.42 $\mu m^2$. However, this design exhibited a significant propagation of 15 clock cycles(CC). To address this limitation, Gassoumi *et al* [9] further proposed a low-power Multiply and Accumulate Unit (MAC) design based on half adders and full adders (FA), enabling the construction of co-planar n-bit Vedic multipliers using the USE (universal, Scalable, and Efficient) clocking method. The presented 2-bit VM reduces cell count and area requirements, utilizing 726 cells in a footprint of 2.66 $\mu m^2$. Further, a 4-bit VM was presented, which required 15,673 cells in footprint of 74.68 $\mu m^2$, and incurs a significant latency of 264 clock zones. In 2020, a 2 × 2 VM utilizing adders based on cell interaction was presented by Safoev and Jeon [10], involving 194 cells in a footprint of 0.17 $\mu m^2$, with 2.25 CC delay. Their continued efforts led to the development of a 4 × 4 VM employing 1726 cells. Further, Ali Newaz Bahar *et al* [11] proposed a coplanar implementation of an n-bit (n = 4 to 64) serial-parallel multiplier using a modified E-shaped XOR gate, achieving a significant reduction in cell count and area. Additionally, a comprehensive energy dissipation analysis was presented. Several squarer circuits have also been proposed in previous literature. More recently, in 2021, Angshuman Khan *et al* [12] presented two co-planar Vedic squarer circuit designs using the UT sutra, exhibiting marked improvements in terms of cell density, footprint, latency and power dissipation relative to prior designs. Specifically, design 1 utilizes 135 cells and occupies 0.20 $\mu m^2$ of space, with a footprint usage of 21.85% and 1.75 CC delay, while the further optimized design 2 requires 64 cells in 0.08 $\mu m^2$ and latency of 0.75 CC.

However, in 2019, Bhoi *et al* [13] also designed a single-layer 2-bit Vedic squarer, employing basic gates and using the 'Yavadunam' method. Their design efficiently used 77 standard cells, resulting in a footprint of 0.17 $\mu m^2$ with a propagation of four CC, reflecting a notable decrease in cell and density requirements. Subsequently, in 2024, Seyed-Sajad Ahmadpour *et al* [14] introduced an energy-efficient quantum multiplier designed for low-power IoT systems, leveraging reversible logic. Meanwhile, Rajender Daggula and Vasudeva Bevara [15] present an optimal design of a 4-bit VM featuring 2,940 cells with a footprint of 7.33 $\mu m^2$ achieving a 90% improvement over CMOS circuits in terms of power efficiency. A comparative summary of existing multiplier designs is presented in table 1. The domain of Vedic multiplication in QCA technology is presently in its exploration stages, characterized by a noticeable absence of comprehensive performance analyses. This phase underscores the extensive possibilities yet to be explored within the realm of Vedic multipliers. In response to these identified gaps, this study seeks to overcome current limitations by presenting a detailed investigation of a novel 2-bit VM design. The proposed VM utilizes half adders and majority gates in its novel architecture, and its energy consumption is evaluated using QCAPro and QCADesigner-E (QDE). The novelty lies not only in the architectural intricacies but also in the potential for enhanced computational speed and efficiency. The key objectives of this research are outlined as below:

- Design and physical verification of the HA circuit using kink energy calculations.
- Holistic fault analysis of a QCA-based HA circuit against single cell omission, double cell omission, and additional cell deposition defect.
- Implementation of an optimized single-layer 2-bit VM based on the UT technique from Vedic mathematics followed by the 4-bit VM implementation.
- Detailed energy dissipation analyses using QCA Designer-E and QCAPro tools.
- Examine the effects of temperature on the average output polarization for the QCA based HA and 2-bit Vedic multiplier circuit.
- Detailed QCA-cost calculation and relevant comparisons to assess the design efficiency of the circuit.

## 2. Essentials of QCA and vedic multiplication

### 2.1. Concepts of QCA technology

The fundamental building block of QCA technology is the QCA cell, a square nano-scale structure incorporating four quantum dots and two mobile electrons that can shift positions between adjacent dots [17, 18]. Due to the

**Table 1.** Review of existing multiplier implementations.

| Refrences | Technique | Tool | Merit | Limitation |
|---|---|---|---|---|
| [9] | Coplanar implementation using the USE clocking scheme | QCADesigner, QCAPro | Reliability evaluation at different temperatures | Bulky architecture |
| [10] | Multilayer 4-bit VM implementation | QCADesigner 2.0.3 | Low complexity | The multilayer approach presents fabrication challenges |
| [11] | Serial parallel multiplier using E-Xor gate | QCADesigner, QCAPro | Reduced cell count, area and cost | Complexity |
| [12] | Coplanar Vedic squarer using UT sutra | QCADesignerE 2.2, QCAPro | High speed and low power consumption | Limited scope as it focuses specifically on squarer circuits |
| [13] | Coplanar Vedic squarer using Yavadunam sutra | QCADesigner | High performance | Energy dissipation not discussed |
| [14] | Parity preserving reversible Vedic multiplier | Not mentioned | Reduced quantum cost | No simulation presented |
| [15] | Coplanar VM using UT sutra | QCADesigner-E | Low power circuit | High cost |
| [16] | Reversible adders and multipliers | QCADesignerE 2.2 | Ultra-low power circuit | No simulation of multiplier |

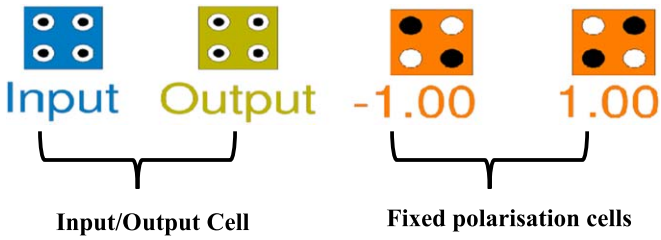



**Figure 1.** Basic components of QCA.

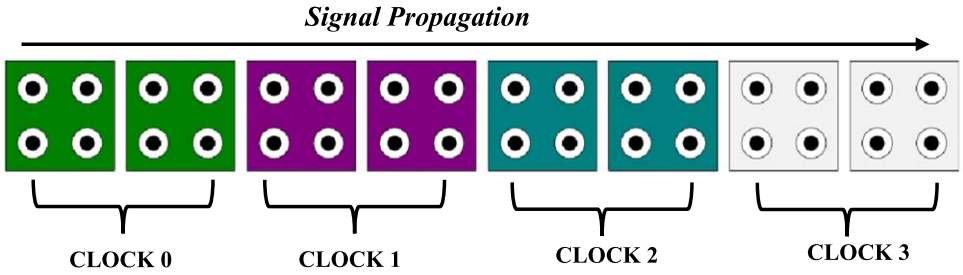



**Figure 2.** Signal transmission in QCA.

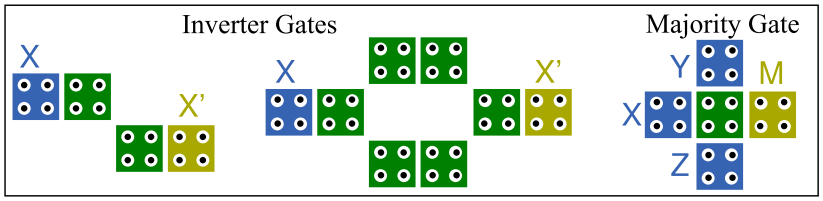



**Figure 3.** Gates in QCA technology.

Coulomb interaction, electrons experience mutual repulsion and therefore are located diagonally next to each other in the square cell. Depending on the diagonal occupied by the electrons, each individual cell exists in one of the two configurations, representing cell polarity +1 and −1, as depicted in figure 1. In an effort to execute logical operations and signal transmission, semiconductor-based QCA systems employ a four-phase clocking scheme with a 90° phase shift between each phase [19].

QCA circuits utilize four clock signals - 0, 1, 2, *and*3 further divided into 4 zones - switch, hold, release, and relax [20]. Cells operating in distinct clock zones are indicated by different colors as illustrated in figure 2. The cells within the circuit are synchronized with these clocks to maintain the directional flow. Information that traverses through all four zones will experience a latency of one clock cycle. To enable computational operations, QCA cells are organized into gates, primarily featuring Inverter(INV) gates and Majority gates (MG). The inverter gate is the fundamental element in digital circuits and is depicted in two forms in figure 3. Further, the 3-input QCA MG employing five QCA cells, with three cells serving as inputs, central cell as the driver cell, and the remaining one functioning as the output are also displayed in figure 3 [21]. The output of the MG is determined by the following expression:

$$M(X, Y, Z) = XY + YZ + XZ \tag{1}$$

In QCA technology, wire crossings primarily fall into two categories: co-planar and multi-layer crossovers (as shown in figure 4). Rotated co-planar crossovers employ an inverter chain or rotated cell sequence with cells rotating at 45° angles to create wire junctions on the same layer [22]. Logical co-planar crossings, however, are governed by specific clock phase rules. These rules prohibit wires with the same clock phase from crossing each other, but allow crossings between 90° wires in different clocks, such as clock zero with clock two and clock one with clock three [23]. On the other hand, multi-layer crossovers utilize a multi-level stereoscopic structure that incorporates several layers [24]. They employ 90° wires to establish a bridge-type structure, with distinct

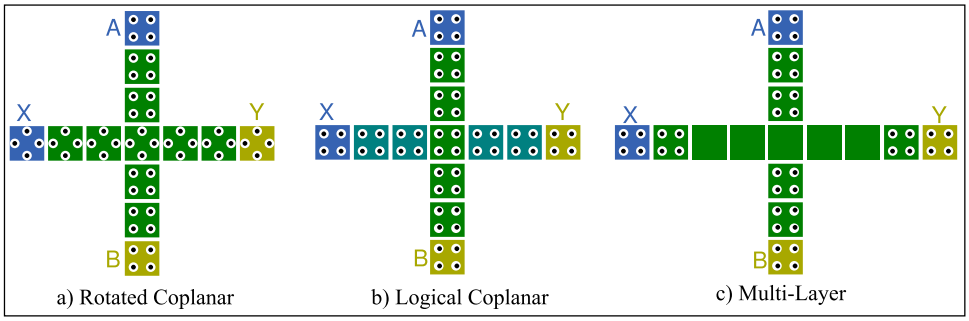



**Figure 4.** QCA Crossovers.

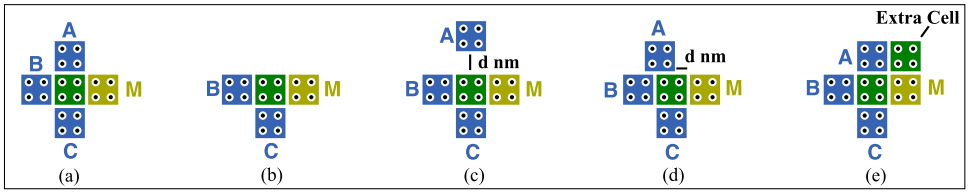



**Figure 5.** QCA Defects, (a) Ideal MG, (b) Missing cell defect, (c) Displaced cell defect, (d) Misaligned cell defect, (e) Extra cell defect.

functions assigned to the lower, middle, and upper layers. The choice between co-planar and multi-layer crossovers is contingent on considerations like area efficiency and design complexity. Multi-layer crossovers are more space-efficient but may entail higher costs and intricate designs, making co-planar crossovers a practical choice for specific applications.

In recent times, there have been several initiatives aimed to address the QCA-based manufacturing issues which primarily arise during the chemical synthesis and deposition phases. It has been observed that faults related to cell misplacement are more prevalent during the deposition phase [25]. This is largely attributed to the compact sizes of QCA cells and the critical requirement for extremely precise cell arrangement, necessitating high levels of accuracy in the manufacturing process [26]. These defects significantly impact the performance of QCA logical devices, compromising their reliability and functionality. Therefore, it is essential to design fault-tolerant QCA circuits to ensure their robust implementation. A fault-free MG gate is used as a reference in figure 5(a) to depict the categorization of various possible deposition defects [26]:

(i) *Cell Omission (Missing Cell):* Cell omission refers to a situation when a QCA cell is moved from its initial place during the manufacturing process. When compared to the fault-free original device, this defect appears due to a missing cell, interfering with the device's intended functionality. An example of this defect is shown in figure 5(b).

(ii) *Cell Displacement:* This occurs when a QCA cell loses its original direction and deviates from its intended placement, resulting in altered distances between the defective cell and its neighboring cells. This defect is demonstrated in figure 5(c) where input A is displaced erroneously by d nm.

(iii) *Cell Misalignment:* It is another common defect in QCA manufacturing, where cells are misaligned from their planned locations causing the defective cell to be closer to some cells while moving away from others. Figure 5(d) showcases an example of the misaligned cell A by d nm.

(iv) *Extra (Additional) Cell:* When an extra cell is inadvertently deposited on the substrate, it leads to additional cell defect which disrupts the intended arrangement of QCA cells and has detrimental effects on the overall functionality of the device. Figure 5(e) demonstrates a scenario where an extra cell is added mistakenly.

### 2.2. Concepts of vedic multiplier

This section delves into the Urdhva-Tiryakbhyam (UT) algorithm, a core technique in Vedic multiplication based on vertical and crosswise operations. Renowned for its versatility for enabling parallel computation of partial products and sums, this algorithm supports both decimal and binary multiplication with enhanced

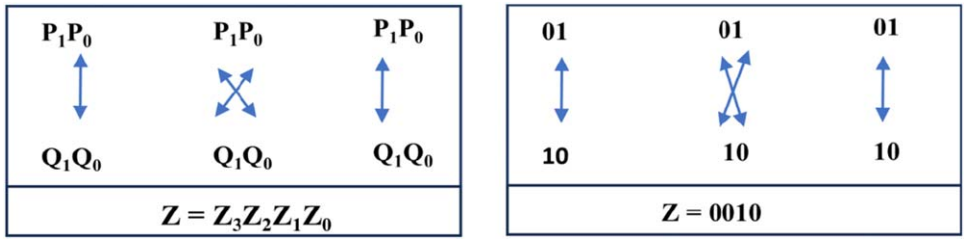



**Figure 6.** Line diagram of 2-bit multiplication.

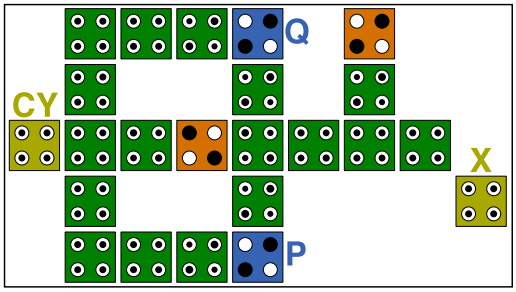



**Figure 7.** Proposed QCA-based half adder design.

efficacy [10]. The algorithm breaks down operands P and Q as $P_1x + P_0$ and $Q_1x + Q_0$ respectively; yielding the following equation [10]:

$$P_1Q_1x^2 + (P_1Q_0 + P_0Q_1)x + P_0Q_0 \tag{2}$$

The UT technique is based on the following principles[10]:

- The least significant bit is the linear multiplication of $P_0$ and $Q_0$.
- The cross products of $P_1$ with $Q_0$ and $P_0$ with $Q_1$ along with the sum of these results forms the multiplier of $x$.
- The most significant bit of $x^2$ indicates the linear multiplication of $P_1$ and $Q_1$.

The Vedic UT algorithm optimizes the multiplication process for operands $P$ and $Q$, breaking it into manageable steps as illustrated in figure 6. The steps combine to produce a four-bit result $Z$ [$Z_3Z_2Z_1Z_0$], efficiently using logic gates for multiplication and addition. The line diagram of this process is also shown, assuming operands such as 01 and 10, respectively, which yields an output Z as 0010 (figure 6.

## 3. Proposed architectures

### 3.1. Optimized HA

In QCA, most half-adder designs are realized through equation-based architectures, combining inverter and MG gates. The initial introduction of a novel single-layer QCA based HA circuit showcasing distinctive features is introduced in [27]. Significantly, it uses a single clock phase and excludes the necessity for any interconnects by incorporating only one INV and MG gate. The proposed configuration with sum (X) and carry (CY) outputs encompasses only 23 cells. The arrangement of the circuit is illustrated in figure 7, and its functioning can be represented as shown in equation (3):

$$X = EX - OR(P, Q) \qquad CY = MG(P, Q) \tag{3}$$

The design ensures optimal and efficient operation with a latency of 0.25 clock zones. Two notable advantages of this design include its simplicity and a reduced cell count. Furthermore, the lack of cross-connections in the layout contributes to a decrease in manufacturing expenses [27].

In the present study, this HA design serves as the foundation for implementing the 2-bit VM. The ongoing efforts involve conducting fault analysis to thoroughly explore the integration and performance of HA design in this context.

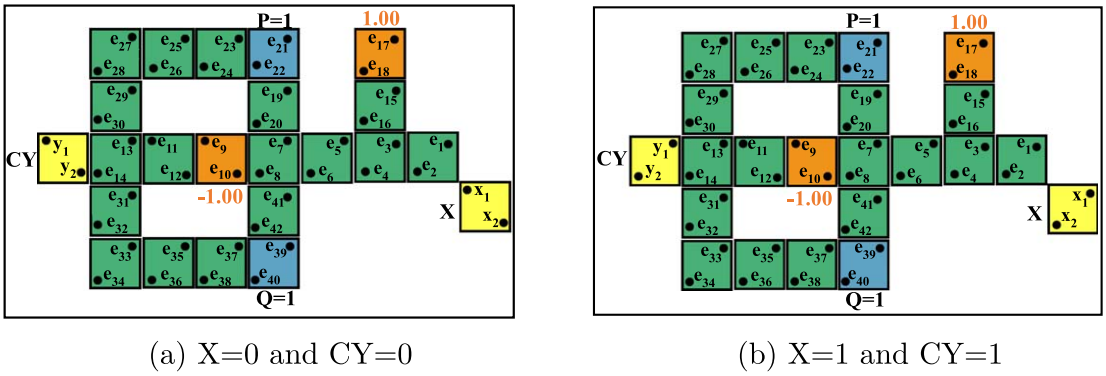



**Figure 8.** The cellular network lattice of the proposed HA.

*3.1.1. Physical verification of ha operation using kink energy*

When designing QCA circuits, it's crucial to understand that the polarization state of a cell isn't solely influenced by the input signal; rather, it intricately depends on the collective dynamics of its neighboring cells [28]. This principle tends to be overlooked in design processes, giving rise to sneaky noise paths. These unintended interactions between adjacent cells can lead to unwanted information flow within the QCA framework [28, 29]. While these spurious noise channels may seem negligible in smaller QCA designs, their impact grows exponentially as the design scales in size and complexity,resulting in functional failures. This phenomenon is related to the concept of kink energy, which represents the potential energy between electrons in QCA cells. The cumulative energy of a cell is derived from the kink energies of its neighboring cells, with the strength of each interaction diminishing as the distance increases. In accordance with this, the noise channels persist as long as multiple quantum cells coexist within each other's radius of influence. These paths can disrupt the intended operation of logically sound QCA building blocks, posing a challenge to the hierarchical composition of larger QCA designs.

To substantiate the physical validity of the proposed QCA-based HA design, our assumptions include uniform quantum cell dimensions (18 × 18nm) and a 2 nm proximity between adjacent cells. Graphically, QCA cells are depicted as squares, with electron positions denoted by circles within each cell [30]. To ensure stability, the arrangement of electrons within each cell is meticulously optimized to minimize their potential energy. The energy ($U$) among two charges a and b is calculated in *Joules* using equation (4)

$$U_{ab} = \frac{1}{4\pi\pi_0\epsilon_r}\frac{q_a q_b}{|r_a - r_b|} \tag{4}$$

where $\epsilon_r$ is the relative permittivity, $q_a$ and $q_b$ are the charges of a and b cells separated by a distance of $|r_a - r_b|$. By replacing all standard values, $U_{ab}$ can be written as shown in equation (5)

$$U_{ab} = \frac{9 \times 10^9 \times (1.6)^2 \times 10^{-38}}{r_{ab}} = \frac{23.04 \times 10^{-29}}{r_{ab}} \tag{5}$$

The total potential energy ($U_T$) is determined by summing the potential energies calculated using equation (6) [28]:

$$U_T = \sum_{x=1}^{n} U_{ab} \tag{6}$$

The key principle of this approach is to determine the state with the minimum potential energy [29]. Since the HA has two inputs, it would have four different input combinations, resulting in a large number of calculations. To simplify the analysis, only the gate accuracy for the input mode (P, Q) = (1, 1) is considered in figure 8. The kink energy between electrons in the output cells X($x_1$ and $x_2$) and CY($y_1$, $y_2$) with all other electrons is calculated and presented in tables 2 and 3. These tables analyze different scenarios based on the assumed values for the output sum and carry signals. However, table 2 examines the situation where the output sum and carry are assumed to be zero.

Similarly, table 3 analyzes the scenarios with the output sum and carry assumed to be one, respectively. The kink energy values in the corresponding tables can be compared to determine the correct values for the outputs X and CY, as a stable state has less energy. For the output CY, analyzing the values reveals that one is the correct value. Similarly, for the output X, the correct value is zero.

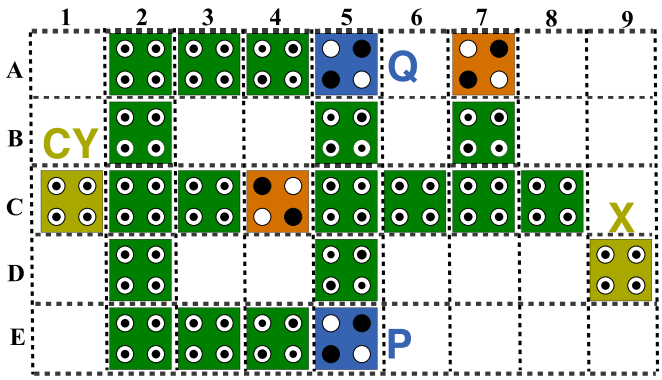



**Figure 9.** Layout of the fault-tolerant HA design.

**Table 2.** Computation of kink energy among all electrons of figure 8(a).

| For electron $x_1$ | For electron $x_2$ |
|---|---|
| $U_{x_1} = \sum_{i=1}^{42} U_{ix_1} = 9.86 \times 10^{-20} J$ | $U_{x_2} = \sum_{i=1}^{42} U_{ix_2} = 10.75 \times 10^{-20} J$ |
| $U_{x_1x_2} = U_{x_1} + U_{x_2} = (9.86 + 10.75) \times 10^{-20} = 20.61 \times 10^{-20} J$ | |
| For electron $y_1$ | For electron $y_2$ |
| $U_{y_1} = \sum_{i=1}^{42} U_{iy_1} = 14.74 \times 10^{-20} J$ | $U_{y_2} = \sum_{i=1}^{42} U_{iy_2} = 30.15 \times 10^{-20} J$ |
| $U_{y_1y_2} = U_{y_1} + U_{y_2} = (14.74 + 30.15) \times 10^{-20} = 44.89 \times 10^{-20} J$ | |

**Table 3.** Computation of kink energy among all electrons of figure 8(b).

| For electron $x_1$ | For electron $x_2$ |
|---|---|
| $U_{x_1} = \sum_{i=1}^{42} U_{ix_1} = 13.66 \times 10^{-20} J$ | $U_{x_2} = \sum_{i=1}^{42} U_{ix_2} = 9.76 \times 10^{-20} J$ |
| $U_{x_1x_2} = U_{x_1} + U_{x_2} = (13.66 + 9.76) \times 10^{-20} = 23.42 \times 10^{-20} J$ | |
| For electron $y_1$ | For electron $y_2$ |
| $U_{y_1} = \sum_{i=1}^{42} U_{iy_1} = 26.25 \times 10^{-20} J$ | $U_{y_2} = \sum_{i=1}^{42} U_{iy_2} = 15.14 \times 10^{-20} J$ |
| $U_{y_1y_2} = U_{y_1} + U_{y_2} = (26.25 + 15.14) \times 10^{-20} = 41.4 \times 10^{-20} J$ | |

### *3.1.2. Reliability assessment and defect analyses of the proposed HA*

- *Resilience to Single-Cell Omission Defect:* The suggested HA circuit is examined for its resilience against the omission of single-cell defect (SCOD) with the exception of the input, output, and fixed cells. To uniquely distinguish each cell within the circuit, a grid-based placement approach is utilized, as illustrated in figure 9. The columns and rows of the matrix are labeled with natural numbers (1, 2, 3, etc.) and alphabet letters respectively. Table 4 shows that the X output demonstrates a fault tolerance ability of 58.8%, with 10 out of 17 instances producing fault-free outputs. However, the CY output achieves a fault tolerance ability of 70.58%, with only 12 out of 17 instances generating correct outputs. These results demonstrate the resilience of the HA circuit with a substantial majority of cases producing accurate outputs.
- *Resilience to Double-Cell Omission Defect:* Imperfect manufacturing processes can lead to the absence of multiple cells in a circuit. To assess the adder's resilience to this defect, the impact of double-cell omission (DCOD) is analysed on the circuit's sum and carry outputs. Table 5 reveals that the X output exhibits a fault tolerance of 71.43%, while the CY output achieves a fault tolerance of 35.7%. These findings demonstrate the proposed adder's robustness against double-cell omission faults, with a significant majority of cases producing

**Table 4.** Assessment of SCOD on Proposed HA.

| Cell | Sum | Carry | Cell | Sum | Carry |
|---|---|---|---|---|---|
| A2 | ✓ | × | A3 | ✓ | ✓ |
| A4 | ✓ | ✓ | B2 | ✓ | × |
| B5 | × | ✓ | B7 | × | ✓ |
| C2 | ✓ | × | C3 | ✓ | ✓ |
| C5 | × | ✓ | C6 | × | ✓ |
| C7 | × | ✓ | C8 | × | ✓ |
| D2 | ✓ | × | D5 | × | ✓ |
| E2 | ✓ | × | E3 | ✓ | ✓ |
| E4 | ✓ | ✓ | | | |

**Table 5.** Assessment of DCOD on Proposed HA.

| Cells | Sum | Carry | Cells | Sum | Carry |
|---|---|---|---|---|---|
| A2,A3 | ✓ | × | A3,A4 | ✓ | × |
| A2,B2 | ✓ | × | B2,C2 | ✓ | × |
| B5,C5 | × | ✓ | B7,C7 | × | ✓ |
| C2,D2 | ✓ | × | C2,C3 | ✓ | × |
| C5,C6 | × | ✓ | C7,C8 | × | ✓ |
| C5,D5 | × | ✓ | D2,E2 | ✓ | × |
| E2,E3 | ✓ | × | E3,E4 | ✓ | × |

accurate outputs. Consequently, the proposed adder can be considered highly fault-tolerant in the presence of double-cell omission defect.

- *Resilience to Cell Displacement/Misalignment Defect:*This defect occurs when the displacement or misalignment of a cell exceeds a critical threshold ($d_p$), affecting the circuits functionality [25].To ensure the robustness of HA circuit, the displacement of the device cells along four possible directions: North, South, East, and West is analyzed and presented in table 6. Here, $d_p$ represents the the maximum permissible displacement of a cell that does not affect the circuit's behavior. If a displacement or misalignment in a given direction is not possible, it is denoted by a '–'. Notably, a permissible displacement of $\geqslant$500 nm suggests that the corresponding cell displacement has a negligible impact on the circuit's behavior. The results indicate that the HA circuit has a good tolerance against these defects.
- *Resilience to Extra Cell Deposition Defect:*The strength of the Coloumbic interaction between cells diminishes rapidly with increase in the distance. The radius of effect (RoE) is defined as the circular area surrounding a cell's center within which it exerts influence on its neighbors.To assess the proposed half-adder's resilience against such defects, additional cells labeled as E1-E29 are placed around the HA circuit, as displayed in figure 10. The impact of each extra cell is individually analyzed (ECDD), and the observed outputs are summarized in table 7. The sum output produced correct results for an impressive 24 of 29 possible scenarios, demonstrating an exceptional 82.75% resilience against the deposition of the extra cells. Remarkably, carry exhibits unwavering fault tolerance, achieving 100% accuracy in all cases.

Therefore, cells placed in close proximity can affect each other's polarization state. During the deposition process, extra cells may be introduced alongside those required for the circuit's proper operation. If these extra cells are positioned too close, they can degrade the circuit's performance.

Further, table 8 offers an extensive comparison of fault tolerance ratios among different HA circuits, including those designed by Majeed *et al* [31] and Safoev *et al* [32], which were examined for cell omission and cell addition defects. The table highlights the superior fault tolerance percentage of the proposed HA design compared to recent designs across all fault types for both outputs, as outlined below:

- For the sum output, it is evident that the proposed approach exhibits superior performance across all fault types compared to both references. Particularly, for SCOD, the proposed HA demonstrates a modest improvement of approximately 2.93% compared to [31], while achieving a remarkable 203.61% improvement compared to [32]. Similarly, for DCOD and ECDD, the proposed method showcases substantial

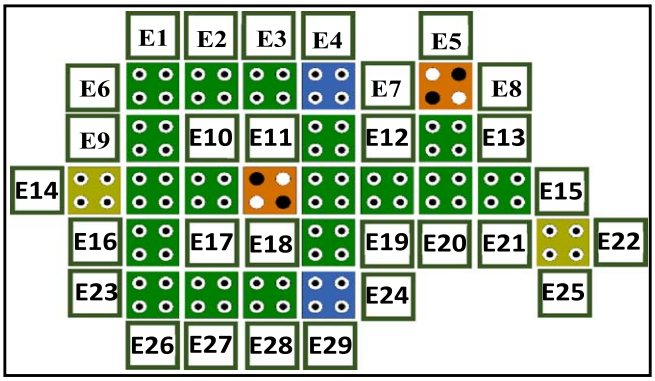



**Figure 10.** HA design with additional cell deposition.

**Table 6.** Assessment of cell displacement/misalignment defect for proposed HA circuit.

| | Displacement ($d_p$) | | Misalignment ($d_p$) | |
|---|---|---|---|---|
| Cell | North | South | East | West |
| A2 | $\leqslant$ 3nm | — | — | $\leqslant$ 5nm |
| A3 | $>$ 500nm | $\leqslant$ 16nm | — | — |
| A4 | $>$ 500nm | $\leqslant$ 16nm | — | — |
| Q | $\leqslant$ 3nm | — | $\leqslant$ 7nm | — |
| B2 | — | — | $\leqslant$ 4nm | $\leqslant$ 4nm |
| B5 | — | — | $\leqslant$ 1.2nm | $\leqslant$ 5nm |
| B7 | — | — | $\leqslant$ 1.5nm | $\leqslant$ 1.5nm |
| CY | $\leqslant$ 3nm | $\leqslant$ 3nm | — | — |
| C3 | $>$ 500nm | $>$ 500nm | — | — |
| C6 | $<$ 1nm | $<$ 1nm | — | — |
| C7 | — | $<$ 1nm | — | — |
| C8 | $\leqslant$ 1nm | $\leqslant$ 7nm | $<$ 10nm | — |
| D2 | — | — | $\leqslant$ 4nm | $\leqslant$ 4nm |
| D5 | — | — | $\leqslant$ 1.2nm | $\leqslant$ 5nm |
| X | $\leqslant$ 10nm | — | — | $\leqslant$ 10nm |
| E2 | — | $\leqslant$ 3nm | — | $\leqslant$ 5 nm |
| E3 | $>$ 500nm | $\leqslant$ 16nm | — | — |
| E4 | $>$ 500nm | $\leqslant$ 16nm | — | — |
| P | — | $\leqslant$ 3nm | $\leqslant$ 7nm | — |

**Table 7.** Assessment of ECDD on proposed HA.

| Cells | Sum | Carry | Cells | Sum | Carry |
|---|---|---|---|---|---|
| E1 | ✓ | ✓ | E16 | ✓ | ✓ |
| E2 | ✓ | ✓ | E17 | ✓ | ✓ |
| E3 | ✓ | ✓ | E18 | ✓ | ✓ |
| E4 | ✓ | ✓ | E19 | ✓ | ✓ |
| E5 | ✓ | ✓ | E20 | × | ✓ |
| E6 | ✓ | ✓ | E21 | × | ✓ |
| E7 | ✓ | ✓ | E22 | ✓ | ✓ |
| E8 | ✓ | ✓ | E23 | ✓ | ✓ |
| E9 | ✓ | ✓ | E24 | ✓ | ✓ |
| E10 | ✓ | ✓ | E25 | ✓ | ✓ |
| E11 | ✓ | ✓ | E26 | ✓ | ✓ |
| E12 | × | ✓ | E27 | ✓ | ✓ |
| E13 | × | ✓ | E28 | ✓ | ✓ |
| E14 | ✓ | ✓ | E29 | ✓ | ✓ |
| E15 | × | ✓ | | | |

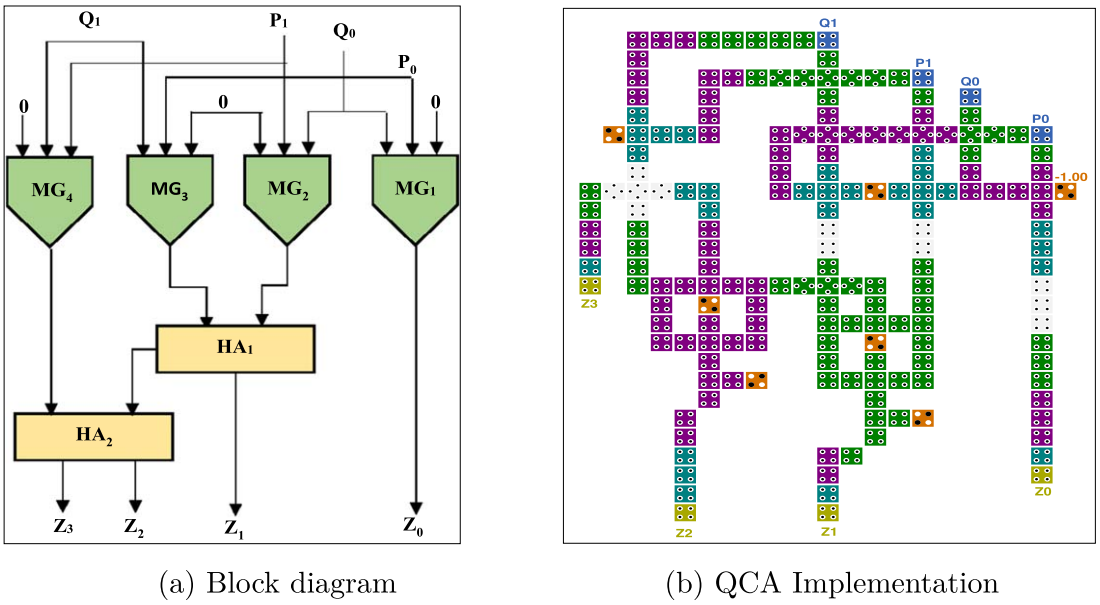



**Figure 11.** The proposed 2-Bit VM.

**Table 8.** Fault resilience assessment of QCA based Half Adder circuits.

| HA Designs | SCOD | | DCOD | | ECDD | |
|---|---|---|---|---|---|---|
| | Sum | Carry | Sum | Carry | Sum | Carry |
| [31] | 57.14 % | 28.57 % | 66.66 % | 33.33 % | 68.18 % | 63.63 % |
| [32] | 23.53 % | 76.47 % | 25 % | 75 % | 76.47 % | 88.24 % |
| Proposed | 58.8% | 70.6% | 71.43% | 35.7% | 82.75% | 100% |

enhancements of approximately 24.15% and 21.31%, respectively, compared to [31], and approximately 185.72% and 30.95%, respectively, compared to [32].

- Regarding the carry output, while the proposed method excels in improving carry output for SCOD compared to [31] (with a remarkable improvement of approximately 147.17%), it slightly lags behind [32], showing a decrease of approximately 53.26%. This lag can be attributed to the design of the underlying architecture of HA. However, for DCOD and ECDD, the suggested method outperforms both references significantly, achieving improvements of approximately 200% and 57.89%, respectively, compared to [31], and approximately 33.33% and 13.40%, respectively, compared to [32].

This comparison further underscores the robustness and potential efficacy of the proposed HA design in larger circuits.

### 3.2. Design and implementation of the VM

The QCA based VM design strategically utilizes co-planar wire-crossing for effective wire intersection, as illustrated in figure 11.

The proposed structure utilizes four MGs and two half-adder circuits to implement the low-complexity 2-bit VM circuit. The number of cells in each clock zone are strategically used in order to ensure proper data propagation and transmission consistency. In particular, this design covers an area of 0.17 $\mu m^2$ utilizing only 179 QCA cells. The 4-bit VM can be constructed using four modules of the proposed 2-bit VM, three ripple carry adders, and one OR gate, as illustrated in figure 12(a). The corresponding QCA implementation, shown in figure 12(b), achieves an efficient design with a minimum of 1762 cells, occupying an area of 1.76 $\mu m^2$ and operating with a latency of 4.25 clock cycles (17 clock zones).

Figure 13 shows the input waveforms of P($P_1$ $P_0$) and Q ($Q_1$ $Q_0$) and the output waveforms of the product Z marked as $Z_0$, $Z_1$, $Z_2$ and $Z_3$. The simulation is performed using the QCA Designer-E(v2.2) software in the coherence vector (with energy) simulation engine with default parameters as listed in table 9.

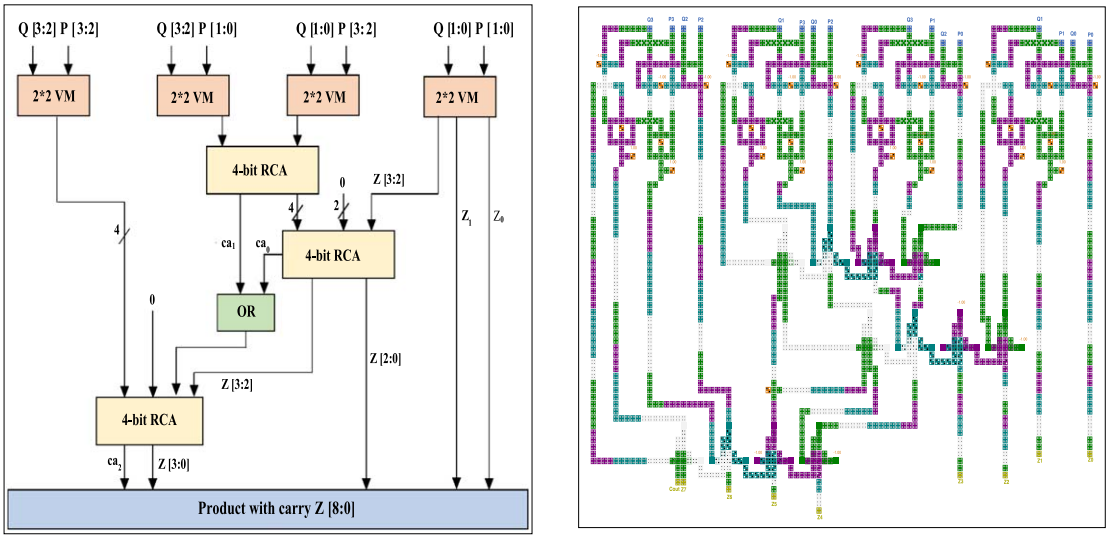



**Figure 12.** The proposed 4-Bit VM.

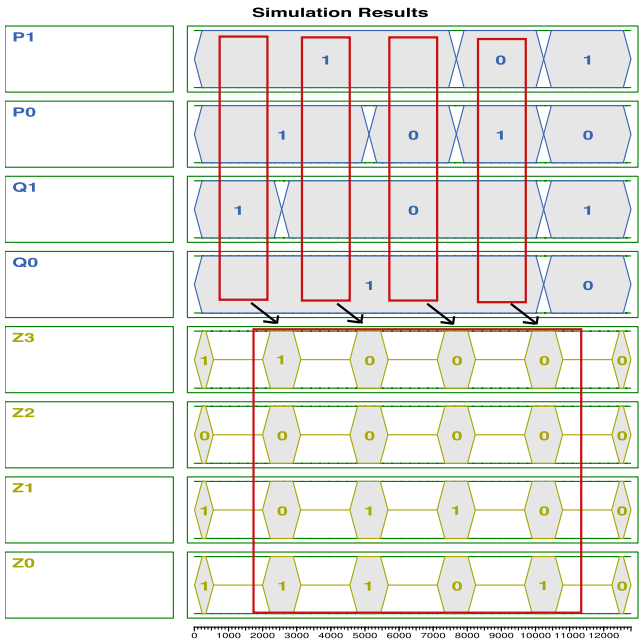



**Figure 13.** Simulated outputs of proposed 2-bit VM.

## 4. Holistic examination and comparative assessment of the proposed circuit

### 4.1. Circuit complexity analysis

In the realm of QCA design, circuit complexity is typically assessed by evaluating the total number of majority gates, inverters, and crossovers (CV) employed [33]. Among these metrics, the number of MGs holds particular significance, as it directly influences the circuit's irreversible power dissipation. Additionally, the number of crossovers in the design serves as an indicator of fabrication difficulty, with multi-layer crossovers incurring a higher cost than co-planar crossovers. Therefore, minimizing these metrics becomes crucial to optimizing circuit performance.

In this regard, the proposed VM design is a single-layer architecture, exclusively employing co-planar crossovers with a complexity of only 14, thereby minimizing both fabrication complexity and power dissipation.

**Table 9.** Default design parameters in QCA Designer-E(v2.2) for coherence vector with energy simulation engine.

| Parameter | Values | Unit |
|---|---|---|
| Quantum Cell Size | 18×18 | $nm^2$ |
| Quantum Dot | 5 | nm |
| Layer Separation | 11.5 | nm |
| Radius of effect | 80 | nm |
| Temperature | 1 | *K* |
| Clock Amplitude factor | 2 | |
| Relaxation Time | 1.00e−15 | *s* |
| Time Setup | 1.00e-16 | *s* |
| Relative Permittivity | 12.9 | |
| Total Simulation Time | 5.00e−11 | *s* |
| Clock Low | 3.80e-23 | *J* |
| Clock High | 9.80e+22 | *J* |
| Clock Shift | 0.00e+00 | |
| Clock Signal Period | 4.00e−12 | *s* |
| $\gamma_{slope}$ | 1.00e−12 | *s* |
| Input Signal Period | 4.00e−12 | *s* |

### 4.2. Area analysis

The overall area of a circuit is a prime consideration in assessing the efficiency and performance of the circuit. A smaller circuit area generally means lower power consumption and faster signal propagation.

*Cell Area:* This metric is computed by multiplying the cell count of the circuit by the default area of a cell ($18 \times 18\ nm^2$). In this context, the design efficiency of the presented VM circuit is established by cell area of 0.05 $\mu m^2$.*Total Area:* In assessing the architecture, the overall rectangular region required for the implementation of the design is the total area of the circuit. An effective design minimizes the total area. The proposed VM circuit achieves a compact area of 0.176 $\mu m^2$.*Area Utilization Factor:* The area utilization factor of a circuit is a measure of the efficiency with which available space is utilized. It is calculated as the ratio of cell area to total area, multiplied by 100. The proposed VM design demonstrates a commendable area utilization of 29.41%, indicating a favorable design approach, as a higher percentage is generally desirable.

### 4.3. Energy calculation

Energy dissipation is one of the pivotal parameters in the evaluation of the efficiency of an electronic design. In the context of QCA circuits, two prominent methodologies are commonly employed to estimate power consumption. These methods involve using two well-established power estimation tools: QCAPro [34] and the QCADesigner-E tool [35] known for their effectiveness in estimating power consumption.

According to the upper bound energy dissipation model [36], the cumulative power dissipation within the circuit [37, 38] can be expressed as follows:

$$P_t = \frac{dE}{dt} = \frac{\hbar}{2}\left[\frac{\mathrm{d}\,\vec{\Gamma}}{\mathrm{d}\,t}.\vec{\lambda}\right] + \frac{\hbar}{2}\left[\vec{\Gamma}.\frac{\mathrm{d}\,\vec{\lambda}}{\mathrm{d}\,t}\right] = P_1 + P_2 \tag{7}$$

In this context, $\hbar$ is the reduced Planck constant, $\vec{\lambda}$ signifies the coherence vector and $\vec{\Gamma}$ denotes the three-dimensional energy vectors that characterize the state of the QCA cell. $P_1$ represents the flow of power in and out of the clock as well as the intercell power flow, whereas $P_2$ denotes the dissipated power.

The non-adiabatic upper bound power dissipation model presented in [36], derived from the quasi-adiabatic model by Timler *et al* [38], is integrated into the QCAPro software. This model is utilized in our analysis to calculate the leakage, switching, and total energy dissipation of the proposed circuits at $T = 2K$ for different tunneling energies. figure 14 illustrates the thermal layout depicting the average energy dissipated in each cell of a 2-bit VM. The darker spots indicate cells that, on average, dissipate a larger amount of energy. The layout was obtained through simulations covering all possible input switching combinations for $E_k = 0.5$. The energy dissipation per cell scaled in units of $10^{-3}$ eV.

Table 10 provides the comparative average, switching, and leakage analyses of energy dissipation for the suggested designs with the recent reported designs, assuming all QCA cells have constant power consumption. In these computations, three distinct tunneling energies (0.5 $E_k$, 1.0 $E_k$, and 1.5 $E_k$) are utilized, with the standard operational temperature for QCA energy analysis set at 2*K*. The percentage per eV unit (denoted as % per eV) in the table represents the corresponding decrease in energy dissipation of the suggested circuits when compared to

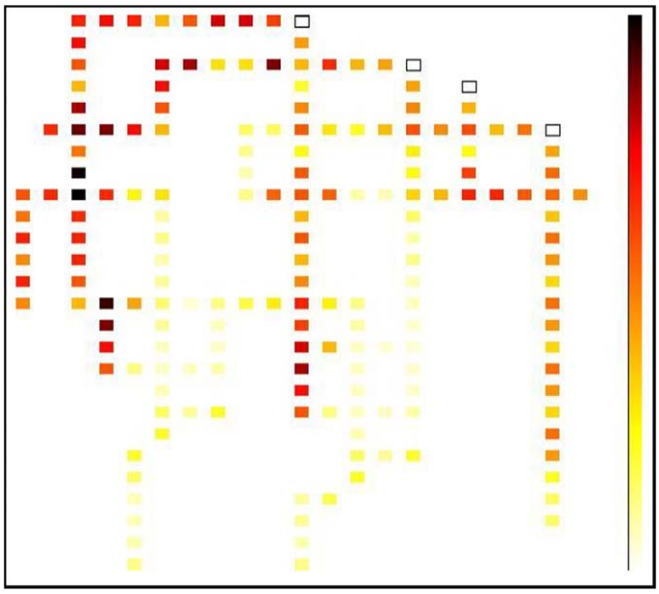


**Figure 14.** Thermal distribution of average energy dissipation in each cell of a 2-bit VM at $T = 2K$ with tunneling energy $E_k = 0.5$.

the reference designs. The consistent improvement of the proposed designs across varying tunneling energy levels further reinforces the potential of the circuits in achieving efficient and sustainable computational processes.The evaluation of the proposed design metrics is also done by using the QCADesigner-E (QDE) tool [39]. The QDE tool is an energy estimation engine where each of the QCA cells in each clock cycle is considered an energy 'bath' [35]. In this context, E_bath_total signifies the overall energy transferred to the bath by all quantum cells ($ET_{bath1}$, $ET_{bath2}$, $ET_{bath3}$, ...) during each clock cycle. In this context, E_bath_total signifies the overall energy transferred to the bath by all quantum cells ($ET_{bath1}$, $ET_{bath2}$, $ET_{bath3}$, ...) during each clock cycle. Table 11 presents a comparative energy dissipation analysis for the proposed designs with the recent designs using QDE tool where Sum_Ebath, denoted as *TED*, encompasses the total energy dissipation, where *TED* represents the energy transfer alongwith the corresponding errors. Avg_Ebath, designated as *AED* indicates the average energy dissipated in each clock cycle, and the error for every clock cycle. For the whole simulation process, the dimensions of the QCA cell are 18 × 18 nm.

The percentage per eV unit (denoted as % per eV) in the table represents the reduction in energy dissipation of the suggested circuits when compared to the reference design. Notably, both the proposed HA and VM designs exhibit significant reductions in total energy dissipation.

### 4.4. Assessment of temperature on output polarization

The Average Output Polarization (AOP) for an output cell is determined by calculating the difference between the maximum and minimum polarization values, measured in electron volts (eV), and then dividing this value by 2 with the results expressed in joules (J) [40]. In our specific example, at a temperature (T) of 2 K, the maximum polarization reaches 9.54e-01, while the minimum polarization is -9.54e-01 for the output cell. Hence, the AOP for the output cell can be computed as shown in equation (8):

$$AOP = \frac{P_{\max} - P_{\min}}{2} = 3.509 \tag{8}$$

To further assess the impact of temperature on the AOP [41] of the 2-bit VM, a comprehensive investigation is conducted using the QDE tool and the output polarization is measured at varying temperatures employing the Euler method of coherence vector simulation engine. The results are visually depicted in figure 15 and the following observations are drawn:

- The AOP of the circuits remains relatively stable until 6K. Beyond this point, the output polarization gradually declines. Hence, as the temperature increases, the AOP decreases. This indicates that the computational fidelity of the circuit decreases with increasing temperature.
- The VM circuit exhibits stable and reliable computation over the low temperature range. This suggests that the circuits is robust to thermal fluctuations.

**Table 10.** Comparative energy dissipation analysis using QCAPro.

(a) Leakage energy Dissipation

| Circuits | $0.5E_k$ | | $1.0E_k$ | | $1.5E_k$ | |
|---|---|---|---|---|---|---|
| | eV | % per eV | eV | % per eV | eV | % per eV |
| ***HA Designs*** | | | | | | |
| [31] | 0.0077 | 30.34 | 0.0206 | 24.83 | 0.0350 | 25.34 |
| | 0.0080 | 33.37 | 0.0226 | 31.51 | 0.0391 | 33.3 |
| Proposed | 0.0054 | — | 0.0155 | — | 0.0261 | — |
| ***VM Designs*** | | | | | | |
| [10] | 0.0689 | 15.67 | 0.2008 | 37.65 | 0.3515 | 41.96 |
| Proposed | 0.0581 | — | 0.1251 | — | 0.204 | — |

(b) Switching energy dissipation

| Circuits | $0.5E_k$ | | $1.0E_k$ | | $1.5E_k$ | |
|---|---|---|---|---|---|---|
| | eV | % per eV | eV | % per eV | eV | % per eV |
| ***HA Designs*** | | | | | | |
| [31] | 0.0214 | 1.68 | 0.0194 | 0.78 | 0.0167 | 2.82 |
| [32] | 0.0213 | 1.22 | 0.0193 | 0.62 | 0.0166 | 2.64 |
| Proposed | 0.0210 | — | 0.0192 | — | 0.0162 | — |
| ***VM Designs*** | | | | | | |
| [10] | 0.1551 | 3.99 | 0.1332 | 5.33 | 0.1125 | 6.93 |
| Proposed | 0.149 | — | 0.1261 | — | 0.104 | — |

(c) Total energy dissipation

| Circuits | $0.5E_k$ | | $1.0E_k$ | | $1.5E_k$ | |
|---|---|---|---|---|---|---|
| | eV | % per eV | eV | % per eV | eV | % per eV |
| ***HA Designs*** | | | | | | |
| [31] | 0.0291 | 9.26 | 0.0400 | 13.18 | 0.0516 | 18.07 |
| [32] | 0.0293 | 10.03 | 0.0420 | 17.28 | 0.0558 | 24.15 |
| Proposed | 0.0264 | — | 0.0347 | — | 0.0423 | — |
| ***VM Designs*** | | | | | | |
| [10] | 0.2241 | 7.54 | 0.334 | 24.79 | 0.464 | 33.47 |
| Proposed | 0.2071 | — | 0.2512 | — | 0.3086 | — |

### 4.5. Cost calculation

In assessing the design efficiency, another widely used metric is the generalized cost function for semiconductor QCA. The importance of the cost metric becomes evident when evaluating the efficacy of QCA designs. $Cost_{AD}$ stands out as the most extensively used parameter, determined by multiplying the area (in $\mu m^2$) with the square of the latency (measured in CC), as defined in equation (9) [33].

$$Cost_{AD} = Area*Latency^2 \tag{9}$$

The equation illustrates a clear connection among the cost function, occupied area, and the associated delay (measured in clock cycles) of the circuit. Thus, optimizing the area and minimizing latency emerge as pivotal elements in minimizing the quantum cost and improving the efficiency of QCA circuit designs. The calculated $Cost_{AD}$ of the suggested VM is 0.52.

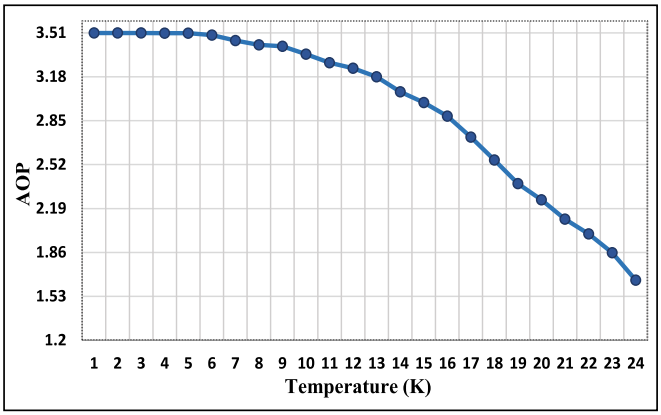



**Figure 15.** Variation in average output polarization (AOP) of 2-bit VM with temperature.

**Table 11.** Comparative energy dissipation analysis of the proposed designs with existing designs using QCADesigner-E.

| Ref. | Avg_Ebath (AED) | % per eV | Sum_Ebath (TED) | % per eV |
|---|---|---|---|---|
| *HA designs* | | | | |
| [31] | 1.03e-003eV (Error: -1.04e-004 eV) | 14.56 | 1.14e-002 eV (Error:-1.15e-003 eV) | 14.56 |
| [32] | 1.64 e-003eV (Error: -1.77 e-004 eV) | 46.3 | 1.80e-002 eV (Error:-1.95 e-003 eV) | 46.3 |
| Proposed | 0.88 e-003 eV (Error: -0.85 e-004 eV) | — | 9.77e-003 (Error: -9.43e-003 eV) | — |
| *2-bit VM designs* | | | | |
| [10] | 6.20e-003eV (Error:-5.47e-004 eV) | 11.77 | 6.83e-002 eV (Error:-6.02e-002 eV) | 11.77 |
| [15] | 6.28e-003eV (Error: -) | 12.89 | 6.90e-002 eV (Error: -) | 12.75 |
| Proposed | 5.47e-003 eV (Error: -4.68e-004 eV) | — | 6.02e-002 eV (Error: -5.15e-003 eV) | — |

The figure of Merit (FoM) [33], is another critical parameter to analyse QCA-specific cost and encompasses key factors such as the count of majority gates, inverter gates, circuit delay, and wire crossings as evaluation parameters. The mathematical expression of this generalized QCA evaluation cost function is represented in equation (10) [33]:

$$FoM = (MG^{\alpha} + INV + CV^{\beta}) \times D^{\delta},\ \alpha,\ \beta,\ \delta \geqslant 1 \tag{10}$$

where, MG, INV, CV and D indicate the number of majority gates, inverters, wire crossing (co-planar crossing $C_{cp}$ or multilayer crossing $C_{ml}$), and latency, respectively. The exponential values $\alpha$, $\beta$, $\delta$ are the subsequent weight metrics. By considering the fabrication complexity, the cost of multilayer crossing $C_{ml}$ is considered three times greater than the cost of co-planar crossing $C_{cp}$. It is worth mentioning that the FoM of the suggested 2-bit VM is 251.1.Moreover, the energy-delay cost function represents a distinct trade-off in performance evaluation. It is calculated as (equation 11) [20],

$$Cost_{ED} = E^{x} * L^{y} \tag{11}$$

where L is the delay, E is the dissipated energy, x and y are experimental weights with standard values of 2 [20, 33]. Therefore, with the energy value generated by QDE, the energy-delay cost for the proposed circuit is 0.011.

### 4.6. Comparative analysis and discussion

A comprehensive comparison of different QCA-based multiplier designs, including array [42–44], matrix[45], Baugh-Wooley [46], Booth [47] and proposed Vedic multiplier is presented. Table 12 provides a succinct summary of the performance metrics evaluated across these diverse multiplier architectures reinforcing the efficacy of the proposed VM in digital applications.

**Table 12.** Comparison of different 2-bit QCA multiplier architectures.

| Reference | CC | Latency | Area | Layer |
|---|---|---|---|---|
| Matrix [45] | 7102 | 20 | 15.69 | Multilayer |
| Serial [48] | 306 | 8 | 0.48 | Coplanar |
| Array [43] | 1598 | 7 | 1.76 | Coplanar |
| Array [42] | 493 | 5 | 0.72 | Coplanar |
| Array [44] | 439 | 1.75 | 0.49 | Coplanar |
| Baugh-Wooley [46] | 688 | 3 | 0.91 | Multilayer |
| Booth [47] | 1257 | 7 | - | Coplanar |
| **Proposed VM** | **179** | **1.75** | **0.17** | **Coplanar** |

**Table 13.** Comparative study of novel 2-Bit VM design against previous approaches.

| Parameters | 2-bit VM Designs | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | [49] | [7] | [8] | [9] | [10] | [12] | [15] | Proposed |
| Layer | Multi | Multi | Single | Single | Single | Single | Single | **Single** |
| Cell Count | 115 | 291 | 1683 | 726 | 194 | 135 | 266 | **179** |
| Latency | 1 | 1.75 | 15 | 8.25 | 2.25 | 1.75 | 3.25 | **1.75** |
| Cell Area | 0.037 | 0.0942 | 0.5452 | 0.2352 | 0.062 | 0.043 | 0.086 | **0.05** |
| Total Area | 0.13 | 0.4 | 8.42 | 2.66 | 0.17 | 0.2 | 0.53 | **0.17** |
| No. of MGs | 8 | 10 | 10 | 13 | 7 | 4 | 10 | **8** |
| No. of INVs | 0 | 2 | 3 | 9 | 0 | 0 | 3 | **2** |
| No. of CVs | 15 | 9 | 3 | 7 | 7 | 2 | 4 | **4** |
| Complexity | 23 | 21 | 16 | 29 | 14 | 6 | 17 | **14** |
| $Cost_{AD}$ | 0.13 | 1.225 | 1894.5 | 181.046 | 0.8606 | 0.6125 | 5.59 | **0.52** |
| FoM | 289 | 560.43 | 25200 | 15450 | 496 | 61.25 | 1256.94 | **251.1** |
| $Cost_{ED}$ | - | 0.255 | 12.42 | 11.19 | 0.059 | 0.0137 | 0.729 | **0.0111** |
| Area Utilisation Factor | 28.46% | 23.55% | 6.48% | 8.84% | 36.47% | 21.50% | 16.23% | **29.41%** |

Data Not Available.

The comparative analysis in table 13 highlights the significant advancements of the suggested design against previous approaches, including the most-recent VM by Daggula and Bevara [15], demonstrating careful consideration of trade-offs between speed, area, and complexity in QCA implementation.

Our design achieves a latency of 1.75 clock cycles, outperforming the recent design [15] (3.25 clocks) and matching the best performance of Khan *et al* [12] and Premananda *et al* [7], while maintaining a reduced cell count of 179 in a single layer. While multi-layer designs like Huang *et al* [49] achieve lower cell counts, our single-layer approach offers a more cost-effective fabrication process [33]. This approach optimizes speed without excessive area consumption or fabrication complexity, as evidenced by the total area of 0.17 $\mu m^2$, which is considerably smaller than the 0.53 $\mu m^2$ reported in [15]. The design's complexity (14) is moderate and matches that of Safoev *et al* [10], while being lower than the complexity of 17 in [15]. Notably, our implementation achieve significant improvements in $Cost_{AD}$, FoM, and $Cost_{ED}$ over all compared designs, indicating superior performance-to-resource ratios. These metrics demonstrate that our design effectively navigates the trade-offs inherent in QCA implementation, resulting in a cost and area-efficient multiplier. Additionally, figure 16 shows the percentage improvement made by the suggested VM design in contrast to the recent single layer architectures in various design metrics.

Furthermore, table 14 provides a comparative analysis of 4-bit VM designs against recent state-of-the-art implementations. The results indicate that the proposed design achieves reductions of 40.07% in cell count, 76.12% in area, 77.63% in latency, and 98.81% in area-to-delay cost compared to the most recent design by [15].

In addition to these quantitative evaluations, it is important to consider practical challenges that arise when scaling QCA circuits. The traditional cost models ($Cost_{AD}$, FoM, and $Cost_{ED}$) as introduced by Liu *et al* [33] and used for evaluating the proposed circuits, may not fully account for interconnect complexity, signal degradation, or long-term maintenance costs, particularly in larger circuits [50]. As designs scale, increased wire lengths and interference can reduce signal integrity, requiring repeaters or buffering mechanisms. Likewise, optimized partitioning, placement, and advanced routing become essential to minimize interconnect overhead. Additionally, maintenance costs such as those related to testing and fault diagnosis can be reduced through the implementation of fault-tolerant design architectures. While traditional models offer valuable insights, they

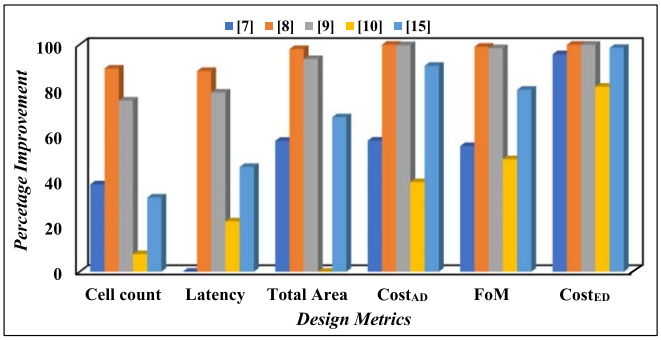



**Figure 16.** Proposed VM versus previous works: percentage improvement in design metrics.

**Table 14.** Comparative study of 4-bit VM against recent approaches.

| Circuit | CC | Area | Latency | $Cost_{AD}$ | Crossover |
|---|---|---|---|---|---|
| [6] | 1955 | 2.55 | 4.25 | 46.05938 | Multilayer |
| [9] | 15673 | 74.68 | 66 | 325306.1 | Coplanar |
| [10] | 1726 | 1.74 | 5.25 | 47.96 | Multilayer |
| [15] | 2940 | 7.33 | 19 | 2646.13 | Coplanar |
| **Proposed** | **1762** | **1.76** | **4.25** | **31.61** | **Multilayer** |

could be extended to reflect real-world constraints, including material expenses and design tool limitations. We recognize the importance of these practical considerations and intend to explore them as part of our future research.

## 5. Conclusion and future works

The realm of nanotechnology has witnessed remarkable advancements in nano-electronics circuits, and QCA technology has emerged as a promising alternative to traditional transistor-based technology. QCA offers significant advantages in terms of lower power consumption and higher speed, making it an attractive choice for future circuit design. In this study, an optimized QCA-based half-adder circuit is used for the subsequent creation of a cell-interaction-based co-planar 2-bit Vedic multiplier within the QCA framework, resulting in substantial improvements in circuit metrics such as cell area, quantum cost, and latency. The half adder is further assessed for cell omission and cell deposition defects. It is observed that the sum and carry outputs show remarkable improvements against single- and double-cell omission and extra cell deposition defect. This comparison underscores the potential efficacy of the HA design in VM circuit. Comprehensive simulations and in-depth performance analysis for various design metrics is performed using the QCADesigner-E v2.2 and QCAPro tool for the suggested circuits. Notably, the proposed design achieves improvements of 22.22% and 81.35% in latency and cost, respectively, with a reduced number of crossovers as compared to the recent single-layer design. Further, the impact of temperature on the average output polarization is presented for the 2-bit VM design. In general, the proposed two-bit Vedic multiplier based on cell interaction demonstrates the potential of QCA technology for efficient and high-performance circuit design. Future research can explore integrating this high-performance, low-power multiplier unit into arithmetic and logic units (ALUs) and multiply-accumulate (MAC) units for signal and image processing systems in mobile devices, where energy efficiency and real-time processing are crucial.

## Statements and declarations

## Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

## Conflict of interests

The authors have no competing interests to declare that are relevant to the content of this article.

## Data availability statement

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

## ORCID iDs

Hemanshi Chugh https://orcid.org/0000-0002-9324-4383
Sonal Singh https://orcid.org/0000-0003-0509-8302

## References

[1] Rairigh D 2005 Limits of cmos technology scaling and technologies beyond-cmos *Citeseer* **Institute of Electrical and Electronics Engineers, Inc, Piscataway** 1–5
[2] Kassa S, Ahmadpour S-S, Lamba V, Kumar Misra N, Jafari Navimipour N and Kotecha K 2024 A cost-and energy-efficient sram design based on a new 5 ip majority gate in qca nanotechnology *Materials Science and Engineering: B* **302** 117249
[3] Navimipour N J, Ahmadpour S-S and Yalcin S 2024 A nano-scale arithmetic and logic unit using a reversible logic and quantum-dots *The Journal of Supercomputing* **80** 395–412
[4] Ahmadpour S-S, Navimipour N J, Mosleh M, Bahar A N and Yalcin S 2023 A nano-scale n-bit ripple carry adder using an optimized xor gate and quantum-dots technology with diminished cells and power dissipation *Nano Communication Networks* **36** 100442
[5] Kassa S, Misra N K, Ahmadpour S S, Lamba V and Vadthiya N 2023 A novel design of coplanar 8-bit ripple carry adder using field-coupled quantum-dot cellular automata nanotechnology *The European Physical Journal Plus* **138** 731
[6] Chudasama A and Sasamal T N 2016 Implementation of 4× 4 vedic multiplier using carry save adder in quantum-dot cellular automata *In International Conference on Communication and Signal Processing (ICCSP)* vol 2016 (IEEE) pp 1260–4
[7] Pooja V G, Premananda B S and Ramesh G S 2018 Design of compact vedic multiplier for high performance circuits *3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT)* (IEEE) pp 1168–72
[8] Gassoumi I, Touil L, Ouni B and Mtibaa A 2019 Design of 4-bit 4-tap fir filter based on quantum-dot cellular automata (qca) technology with a realistic clocking scheme *In Advances in Quantum Communication and Information* (IntechOpen) pp 79–94
[9] Gassoumi I, Touil L and Ouni B 2019 Design of efficient quantum dot cellular automata (qca) multiply accumulate (mac) unit with power dissipation analysis *IET Circuits, Devices & Systems* **13** 534–43
[10] Safoev N and Jeon J-C 2020 Design and evaluation of cell interaction based vedic multiplier using quantum-dot cellular automata *Electronics* **9** 1036
[11] Bahar A N and Wahid K A 2019 Design of qca-serial parallel multiplier (qspm) with energy dissipation analysis *IEEE Trans. Circuits Syst. Express Briefs* **67** 1939–43
[12] Khan A, Bahar A N and Arya R 2021 Efficient design of vedic square calculator using quantum dot cellular automata (qca) *IEEE Trans. Circuits Syst. Express Briefs* **69** 1587–91
[13] Bhoi B K, Misra N K and Pradhan M 2019 A novel and efficient design for squaring units by quantum-dot cellular automata *Microelectronics, Electromagnetics and Telecommunications: Proceedings of the Fourth ICMEET* (Springer) pp 23–31
[14] Noorallahzadeh M, Mosleh M, Ahmadpour S-S, Pal J and Sen B 2023 A new design of parity preserving reversible vedic multiplier targeting emerging quantum circuits *Int. J. Numer. Modelling, Electron. Netw. Devices Fields* **36** e3089
[15] Daggula R and Bevara V 2024 An ultra-low power qca based vedic multiplier for digital radar application *e-Prime-Advances in Electrical Engineering, Electronics and Energy* **9** 100695
[16] Ahmadpour S-S, Noorallahzadeh M, Al-Khafaji H M R, Darbandi M, Jafari Navimipour N, Javadi B, Ain N U, Hosseinzadeh M and Yalcin S 2024 A new energy-efficient design for quantum-based multiplier for nano-scale devices in internet of things *Comput. Electr. Eng.* **117** 109263
[17] Xu Q, Gao J, Wang S, Wang Y, Liu D and Wang J 2021 Quantum dots in cell imaging and their safety issues *Journal of Materials Chemistry B* **9** 5765–79
[18] Das P *et al* 2019 Heteroatom doped blue luminescent carbon dots as a nano-probe for targeted cell labeling and anticancer drug delivery vehicle *Mater. Chem. Phys.* **237** 121860
[19] Chudasama A, Nath Sasamal T and Yadav J 2018 An efficient design of vedic multiplier using ripple carry adder in quantum-dot cellular automata *Comput. Electr. Eng.* **65** 527–42
[20] Bahar A N and Wahid K A 2020 Design of an efficient n× n butterfly switching network in quantum-dot cellular automata (qca) *IEEE Trans. Nanotechnol.* **19** 147–55
[21] Khan A and Arya R 2022 Design and energy dissipation analysis of simple qca multiplexer for nanocomputing *The Journal of Supercomputing* **78** 8430–44
[22] Riki S and Serajeh Hassani F 2023 Designing a one-bit coplanar qca alu using a novel robust area-efficient three-input majority gate design *The Journal of Supercomputing* **79** 17897–918
[23] Patidar M and Gupta N 2021 Efficient design and implementation of a robust coplanar crossover and multilayer hybrid full adder–subtractor using qca technology *The Journal of Supercomputing* **77** 7893–915
[24] Patidar M, Singh U, Shukla S K, Prajapati G K and Gupta N 2023 An ultra-area-efficient alu design in qca technology using synchronized clock zone scheme *The Journal of Supercomputing* **79** 8265–94
[25] Negahdar K, Mosleh M, Ahmadpour S-S, Jafari Navimipour N and Shahrbanoonezhad A 2024 Toward implementing robust quantum logic circuits using effectual fault-tolerant majority voter gate *Materials Science and Engineering: B* **301** 117161
[26] Momenzadeh M, Huang J, Baradaran Tahoori M and Lombardi F 2005 Characterization, test, and logic synthesis of and-or-inverter (aoi) gate design for qca implementation *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **24** 1881–93

[27] Chugh H and Singh S 2023 Efficient co-planar adder designs in quantum dot cellular automata: energy and cost optimization with crossover elimination *Integration* **94** 102103
[28] Kim K, Wu K and Karri R 2006 The robust qca adder designs using composable qca building blocks *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **26** 176–83
[29] Majeed A and Alkaldy E 2022 High-performance adder using a new xor gate in qca technology *The Journal of Supercomputing* **78** 11564–79
[30] Sen B, Dutta M, Mukherjee R, Nath R K, Sinha A P and Sikdar B K 2016 Towards the design of hybrid qca tiles targeting high fault tolerance *J. Comput. Electron.* **15** 429–45
[31] Majeed A H, Zainal M S B, Alkaldy E and Nor D M 2020 Full adder circuit design with novel lower complexity xor gate in qca technology *Transactions on Electrical and Electronic Materials* **21** 198–207
[32] Safoev N and Jeon J-C 2020 Design of high-performance qca incrementer/decrementer circuit based on adder/subtractor methodology *Microprocess. Microsyst.* **72** 102927
[33] Liu W, Lu L, O'Neill Máire and Swartzlander E E 2014 A first step toward cost functions for quantum-dot cellular automata designs *IEEE Trans. Nanotechnol.* **13** 476487
[34] Srivastava S, Asthana A, Bhanja S and Sarkar S 2011 Qcapro-an error-power estimation tool for qca circuit design *IEEE International Symposium of Circuits and Systems (ISCAS)* vol 2011 (IEEE) pp 2377–80
[35] Sill Torres F, Wille R, Niemann P and Drechsler R 2018 An energy-aware model for the logic synthesis of quantum-dot cellular automata *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **37** 3031–41
[36] Srivastava S, Sarkar S and Bhanja S 2008 Estimation of upper bound of power dissipation in qca circuits *IEEE Trans. Nanotechnol.* **8** 116–27
[37] Tougaw P D and Lent C S 1996 Dynamic behavior of quantum cellular automata *J. Appl. Phys.* **80** 4722–36
[38] Timler J and Lent C S 2002 Power gain and dissipation in quantum-dot cellular automata *J. Appl. Phys.* **91** 823–31
[39] Walus K, Dysart T J, Jullien G A and Arief Budiman R 2004 Qcadesigner: A rapid design and simulation tool for quantum-dot cellular automata *IEEE Trans. Nanotechnol.* **3** 26–31
[40] Abdullah-Al-Shafi M, Bahar A N, Bhuiyan M M R, Shamim S M and Ahmed K 2018 Average output polarization dataset for signifying the temperature influence for qca designed reversible logic circuits *Data in Brief* **19** 42–8
[41] Hassan M K, Nahid N M, Bahar A N, Bhuiyan M M R, Abdullah-Al-Shafi M and Ahmed K 2017 Dataset demonstrating the temperature effect on average output polarization for qca based reversible logic gates *Data in Brief* **13** 713–6
[42] Yang B and Afrooz S 2019 A new coplanar design of multiplier based on nanoscale quantum-dot cellular automata *Int. J. Theor. Phys.* **58** 3364–74
[43] Hanninen I and Takala J 2007 Pipelined array multiplier based on quantumdot cellular automata *18th European Conference on Circuit Theory and Design* vol 2007 (IEEE) pp 938–41
[44] Yan A, Li X, Liu R, Huang Z, Girard P and Wen X 2023 Designs of array multipliers with an optimized delay in quantum-dot cellular automata *Electronics* **12** 3189
[45] Lu L, Liu W, O'Neill Máire and Swartzlander E E 2010 Qca systolic matrix multiplier *IEEE Computer Society Annual Symposium on VLSI* (IEEE) pp 149–54
[46] Bhoi B K, Misra N K, Pradhan M and Rout R 2019 Synthesis methods of baugh-wooley multiplier and non-restoring divider to enhance primitive's results of qca circuits *Smart Intelligent Computing and Applications: Proceedings of the Second International Conference on SCI* (Springer) pp 237–45
[47] Kamrani H and Heikalabad S R 2021 Design and implementation of multiplication algorithm in quantum-dot cellular automata with energy dissipation analysis *The Journal of Supercomputing* **77** 5779–805
[48] Walus K, Jullien G A and Dimitrov V S 2003 Computer arithmetic structures for quantum cellular automata *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers* vol 2 (IEEE) pp 1435–9
[49] Huang J and Lale S 2022 A novel nano-scale architecture of vedic multiplier using majority logic in quantum-dot cellular automata technology *Electron. Lett.* **58** 660–2
[50] Jana S, Sultana R, Metia R, Chakrabarti K and Sarkar A 2024 Innovative technique for generation of x-ray based on molecular quantum harmonic oscillator and subsequent application in determination of carcinoma of lungs and breast cells using gold nanowire network *African Journal of Biomedical Research* **27** 7062–76

Environmental Earth Sciences (2025) 84:308
https://doi.org/10.1007/s12665-025-12309-x

ORIGINAL ARTICLE

# Assessment of arsenic contamination in drinking water sources and its effects on human health in rural regions: a study in the Malistan district, Ghazni province of Afghanistan

**Ali Reza Noori**[1,2] · **S. K. Singh**[1] · **Assadullah Rezai**[2]

Received: 23 December 2024 / Accepted: 4 May 2025
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2025

**Abstract**
Arsenic in drinking water, even in trace levels, can cause cancer, skin conditions, heart disease, and developmental abnormalities in children. The Malistan District in Ghazni Province of Afghanistan heavily relies on groundwater. This study aimed to evaluate the concentration of arsenic, iron, and nitrate in drinking water sources and identify any potential health hazards. Seventy-three water samples, including wells, springs, and surface water, were collected and analyzed for pH, total dissolved solids (TDS), electrical conductivity (EC), temperature, $NO_3$, Fe, and arsenic levels. Descriptive statistical analysis categorized the data by water source type, comparing results to World Health Organization (WHO) and Afghanistan National Standard Authority (ANSA) guidelines. Arsenic contamination was identified in six wells and four springs, ranging from 5 to 20 µg/L, with iron concentrations up to 4.5 mg/L. Nitrate concentrations, up to 40 mg/L, were observed in most villages. EC and TDS showed a high correlation ($R = +0.998$), while a moderate correlation existed between arsenic and iron ($R = +0.6205$). Rabat village reported 23 cancer deaths from 2000 to 2023, with stomach cancer accounting for about 48% in the 60–80 age group. Health risk assessments revealed hazard quotient (HQ) values exceeding 1 for all samples, indicating potential noncancerous effects, and carcinogenic risk (CR) values greater than 0.0001, suggesting a potential risk of various cancers for children and adults. Urgent measures are needed to address arsenic contamination and associated health risks in the study area. The study suggests several key recommendations, including utilizing alternative water sources, employing arsenic treatment technology, implementing public awareness campaigns, instituting consistent water quality monitoring, advancing healthcare initiatives, and fostering community involvement.

**Keywords** Water quality · Drinking water contamination · Health risk assessment · Arsenic · Malistan district

## Introduction

Groundwater serves as a critical source of drinking water for millions of people worldwide, playing a vital role in sustaining human life and supporting ecosystems. Since groundwater frequently naturally filters through layers of rock and soil, it is less vulnerable to pollutants and other impurities than surface water. The credibility of this resource is underscored by its reliability, accessibility, and consistent quality, which is particularly valuable in regions grappling with pollution or water scarcity issues, as highlighted by Noori and Singh (2021a). Furthermore, groundwater serves as a vital buffer during droughts and fluctuations in surface water availability, offering essential support during emergencies, as emphasized by Noori and Singh (2021b). As population growth and water demands escalate, responsible protection and management

✉ Ali Reza Noori
a.noori@kpu.edu.af

S. K. Singh
sksinghdce@gmail.com

Assadullah Rezai
a.rezai@kpu.edu.af

[1] Department of Environmental Engineering, Delhi Technological University, Delhi, India

[2] Department of Water Supply and Environmental Engineering, Faculty of Environmental and Water Engineering, Kabul Polytechnic University, Kabul, Afghanistan

Published online: 26 May 2025

of groundwater reservoirs become imperative to ensure future generations'access to clean drinking water. Numerous investigations have been conducted on various facets of groundwater management (Noori and Singh 2021c, 2023a, b, 2024; Singh and Noori 2022).

When the concentration of arsenic in water sources is higher than the permissible limit and poses a severe threat to the health of people and ecosystems, it is called arsenic pollution. Arsenic contamination can originate from both natural and anthropogenic sources. Naturally, groundwater containing abundant arsenic minerals can absorb arsenic from geological formations. Industrial processes that release arsenic into the environment include mining, smelting, and pesticide use. These are examples of anthropogenic sources. Arsenic poisoning of drinking water has a long history that crosses continents. Due to mining, industrial runoff, and agricultural activities, arsenic entered water systems in the late 19 th and early twentieth centuries as industrialization developed (Ravenscroft et al. 2009). Groundwater-dependent areas had an incredibly high level of pollution because naturally occurring arsenic deposits have the potential to seep into aquifers gradually (Smith et al. 2000b). Global attention was drawn to the problem by significant occurrences like the poisoning epidemics in the 1950 s in Toyama Prefecture, Japan.

Contamination of groundwater by arsenic is concerning, particularly in localities where people use groundwater as a domestic and drinking water source. There is consistent evidence linking prolonged exposure to elevated levels of arsenic in drinking water to a higher risk of cancer. Researches indicate that those who have been exposed to arsenic are at a higher risk of developing skin, lung, bladder, kidney, and liver cancers while maintaining human health and human well-being is an essential concern for human society, which is carried out in many studies in this direction (Ayeni et al. 2022; Ayejoto et al. 2023, 2024; Alao et al. 2023; Ayejoto and Egbueri 2024). For instance, studies in Taiwan discovered a correlation between drinking water arsenic levels and the prevalence of liver, lung, and bladder cancer (Chen et al. 1988). According to a similar study conducted in Bangladesh, skin cancer growth and arsenic exposure are significantly correlated (Smith et al. 2000a). Several additional studies have demonstrated a higher correlation between prolonged exposure to arsenic and the emergence of cancers of the skin, lungs, bladder, and kidneys, among other cancer types (Mandal and Suzuki 2002; Bhowmick et al. 2018). As a carcinogen, arsenic damages DNA and interferes with biological functions, which eventually results in the development of malignant cells (Abedin et al. 2002; Jomova et al. 2011). Many studies used the arsenic health risk assessment approach to evaluate the health impact of arsenic-contaminated water sources (Varol 2021; Hoang et al. 2022; Riaz et al. 2022).

Arsenic-related groundwater contamination is a major environmental problem that causes serious health concerns worldwide. Arsenic contamination of groundwater occurs naturally in various regions of the world and is made worse by human activities such as mining, industrial operations, and agricultural practices. The interaction of geological and human elements increases the intricacy of resolving this pollution problem. Geological reasons, such as sediments high in arsenic in aquifers, cause most pollution. Different regions of the world have varying levels of groundwater arsenic pollution. High concentrations of arsenic have been found in many areas, including Bangladesh, West Bengal, India, parts of China, Vietnam, and some regions of South and Central America. Millions of people in these regions rely primarily on groundwater for their drinking water, which puts them at risk for health problems related to arsenic exposure. The effects of groundwater pollution by arsenic have been brought to light by some recent studies (Smith et al. 2006; Van Geen et al. 2007; Brammer and Ravenscroft 2009; Smedley and Kinniburgh 2013; Shaji et al. 2021).

Arsenic and its contamination of Afghanistan's water resources are also significant challenges. Studies on water quality in the nation have addressed this problem sporadically, as its presence in drinking water sources can lead to several illnesses, most notably malignancies. For instance, a survey by Saffi and Kohistani (2016) Showed that water resources in some parts of Panjshir province are contaminated with arsenic. Based on this investigation, Panjshir's arsenic concentration varied from 0.00 mg/L to 0.4 mg/L. Arsenic is found in the thermal springs in the province of Bamyan, according to a study conducted by Jawadi et al. (2021). The discovery indicates that these springs have an arsenic content of up to 100 µg/L. The existing literature indicates that arsenic contamination is present in the following provinces: Herat, Kabul, Panjshir, Maidan Wardak, Logar, Ghazni, Helmand, and Bamyan (Hayat and Baba 2017; Zahidi et al. 2020; Mahaqi et al. 2021; Rasouli and Vaseashta 2023).

With serious consequences for the overall well-being of the populace, cancer is a significant public health concern in Afghanistan. The country has shown a steady increase in cancer incidence in recent years, according to World Health Organization (WHO) estimates. Colorectal, stomach, lung, and breast cancers are the most prevalent cancer types in the nation. According to a study by Joya et al. (2020), esophageal and breast cancers were the most common malignancies found in Afghanistan. In the age range of 50 to 70 years, stomach and esophageal cancers were the most common malignancies among men. On the other hand, women were more likely than males to have esophageal and breast cancers in the age range of 25 to 65. The difficulties in treating cancer in the country stem from a variety of causes, such as poor screening

and diagnostic facilities, restricted access to healthcare services, and a lack of knowledge about preventative measures. Furthermore, the availability of drinking water resources and quality may significantly impact cancer rates in Afghanistan. Urgent attention, increased investment, and the implementation of comprehensive cancer control programs are necessary to improve the prevention, diagnosis, and treatment of cancer in Afghanistan.

Cancer rates are rising in the study region. There might be a well-established correlation between certain types of cancer and drinking water quality. This is especially true for heavy metals like arsenic, a known carcinogen. The main goal of this study is to examine the concentration of arsenic in the drinking water of the study area. The specific objectives are as follows: 1) Determining the concentration of arsenic, iron, and nitrate in the drinking water sources of the study area. 2) Determining the number of cancer deaths in Rabat village around geothermal springs. 3) Assess the potential impact of elevated arsenic levels on human health within the study area. This study holds significant importance as it addresses a critical public health concern. The study is the first attempt in the study area to address drinking water quality. Understanding the potential association between arsenic exposure and cancer rates can inform public health interventions, lead water quality improvement measures, and contribute to developing targeted strategies to mitigate health risks in affected communities.

## Materials and methods

### Malistan district

Malistan district is located in the western part of Ghazni province. Its area is about 1,750 $km^2$, with an elevation ranging from 2525 to 4434 m above sea level (Fig. 1). Geographically, the district is located between 32°54"44"N and 33°38′46"N latitude and 66°51"24"E and 67°37"31"E longitude (Fig. 1). Its distance from the center of the province is about 120 km. Malistan is a cold and mountainous area whose winter is cold and snowy, and summer is mild and pleasant. The district has 288 villages. Malistan is one of the most populated districts of Ghazni Province. In 2024, its population is estimated to be 99,355 people, whose a growth is 2.7% based on the National and Information Authority (NSIA 2024).

Its high elevation and mountainous terrain characterize the district's climate. Due to its continental climate, the region has pleasant summers and frigid, snowy winters. Winter months from December to February can see sharp drops in temperature, with snowfall being predicted. There is a transitional phase in the springtime, with sporadic rainfall in March and April. Malistan experiences moderate, pleasant summers ranging from comfortable to warm.

The hydrology of the study area is shaped by its hilly topography and seasonal precipitation. Rivers, streams, and springs are partly formed by winter snowfall and spring rainfall. In terms of geography, this region stretches from east to west and features a sizable river. It has a significant center valley and smaller surrounding valleys. The river has a branch that flows westward from Malistan's easternmost point (Dala village). Apart from merging at"Shine Deh village"and heading into the Jaghori district, another branch arose from the"Kushnag"and"Miradina"regions west of Malistan to reach other rivers.

Additionally, agriculture—which is bolstered by smaller water sources like springs and streams—is the primary source of revenue for the local population. However, hydrology may be impacted by things like human activity and climate change (UNHCR Field Office Ghazni 2003). Thus, effective water resource management and conservation are essential for the district's sustainable growth. Most of the people in this region are engaged in agriculture, and their common crops include wheat, barley, and potatoes (Wikipedia contributors 2025).

Malistan district, located in the central highlands of Afghanistan, is characterized by a complex geological landscape. The area's geology is characterized by complex geological features, mainly mountainous regions, and composed of sedimentary, metamorphic, and igneous rocks. The region comprises limestone, sandstone, and shale, which characterize its stratigraphic features. Tectonic activities in this region have led to the formation of many folds and faults. The geological processes have formed a mountainous landscape with peaks and deep valleys, affecting the region's natural resources and mineral potential. The surficial geological map of the district is illustrated in (Fig. 2).

### Study area

This research's study area is located inside the Malistan district (Fig. 1). This study covers the villages of Qulamzid and Kadlag. It partially covers the villages of Bughra (upper, middle, and lower), Rabat, Ghighanto, Pashakhana, and Bewa. The villages of Bughra, Qulamzid, and Kadlak are in the vicinity and include wider valleys to the west of the main road of Malistan and Jaghori. About 30 households live in Qulamzid and Kadlak villages, and about 200 live in Bughra villages. The village of Gheganto is a relatively narrow and long valley from east to west. It has many hills and valleys and is surrounded by relatively higher mountains. The villages of Rabat, Pashakhana, and Bewa are situated near the main roads of Malistan and Jaghori and have lower altitudes than the other mentioned villages.

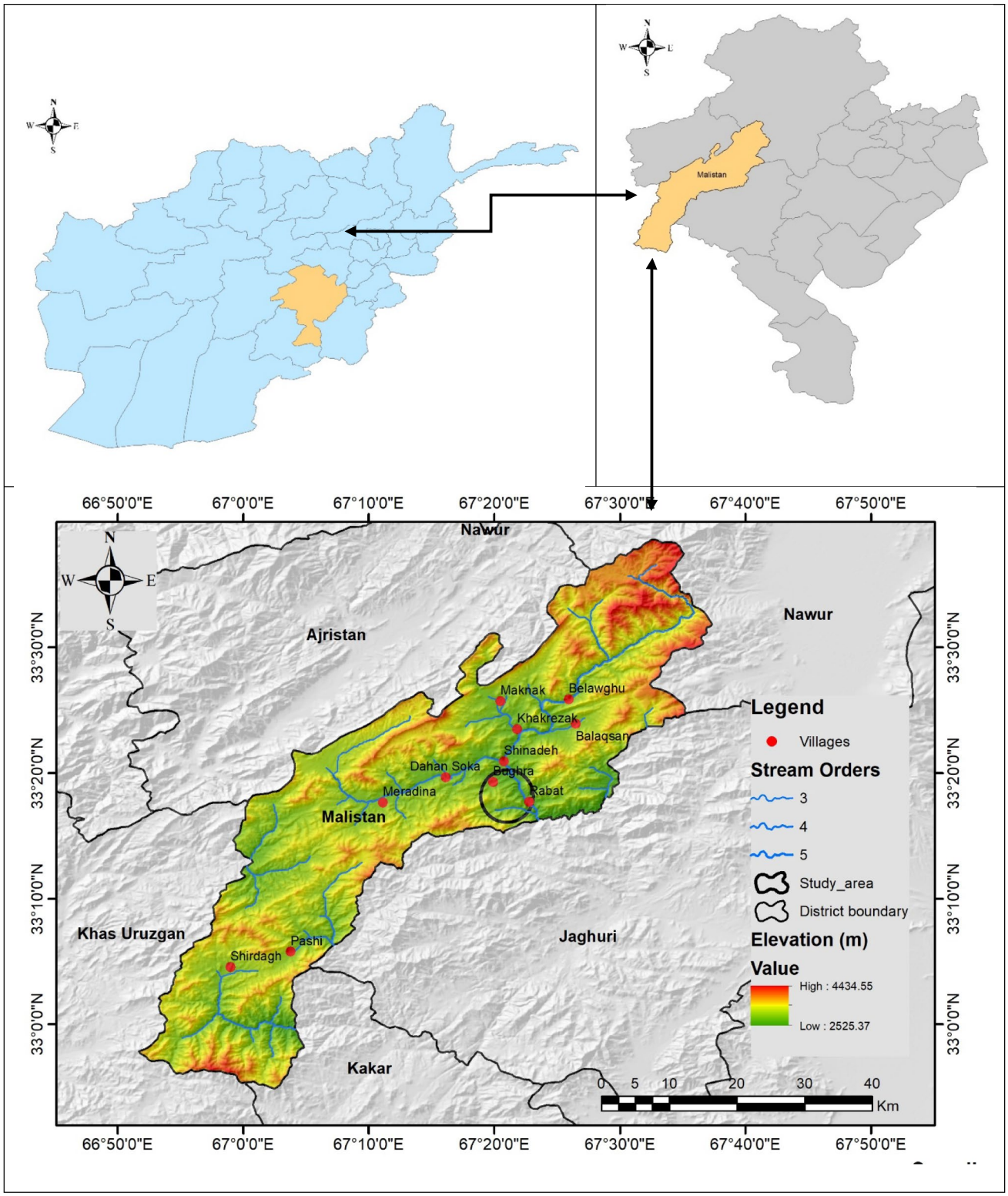



**Fig. 1** Location map of the Malistan district

The research region has three primary water sources: wells (mainly shallow wells), springs, and surface water originating from upper-level springs. In the study region, wells—especially those of the shallow variety—are an essential domestic water source. Shallow wells are typically dug into the earth to reach groundwater reserves. These

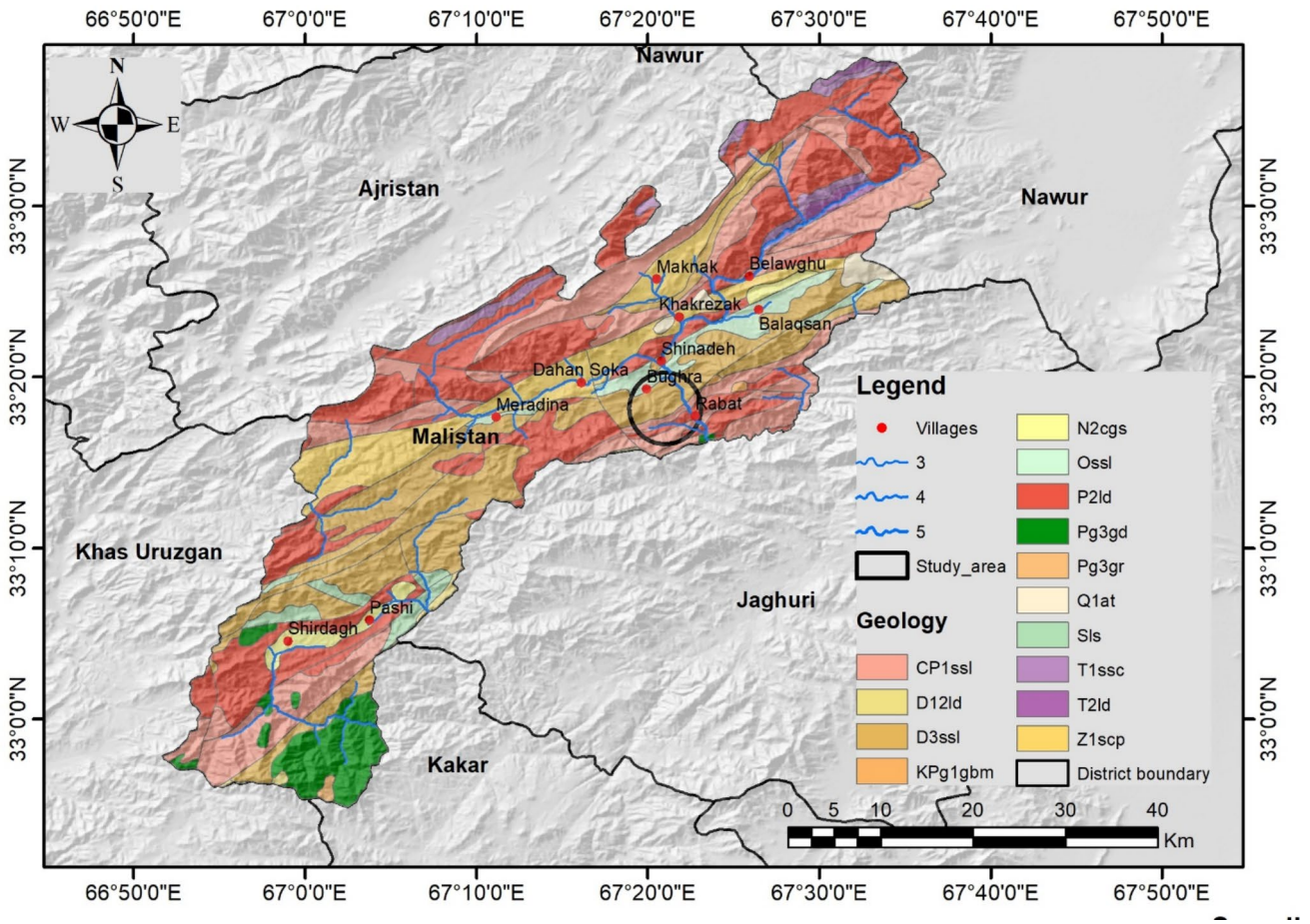



**Fig. 2** Geological setting of the Malistan district

wells tap into shallow aquifers relatively close to the surface. Several factors, including human activities, nearby pollution sources, and geogenic causes, can impact the water quality of shallow wells.

The second reliable source that supplies most of the domestic water in the study area is springs. These springs provide a reliable water source scattered within the research area. Because of natural filters through rock and soil layers, spring water is usually considered suitable for domestic uses. Besides domestic usage, spring water in most villages of the study area is used as an irrigation and agricultural water source.

Lastly, groundwater—which manifests as springs in higher altitudes—is also the source of surface water in the research region. Streams and other bodies of water from snowmelt and rainfall runoff make surface water sources in the study area. Surface water in the study area is mainly utilized for irrigation purposes. However, in some parts of the study area, it was used for domestic purposes, including drinking and cooking, without further treatment. However, it is essential to consider the quality of surface water, as it can be influenced by natural pollution and human activities.

Eventually, the research region depends on three primary water sources for drinking: surface water, wells (mostly shallow wells), and springs next to households. These sources are vital to supplying the community's water needs, and understanding their attributes, accessibility, and quality is necessary for sustainable water management and planning in the region. Location map of the study area and water sampling points are illustrated in (Fig. 3).

## Water sample collection

The water sampling involved in situ testing and collection in High-Density Polyethylene (HDPE) bottles. Some samples were directly tested on-site in the field without being transferred to any other container. This method allowed for immediate analysis and evaluation of the groundwater quality at the sampling location.

In general, the wells of the study area possess two mechanisms for water extraction: a hand pump and a water pump equipped with a solar system. The water was pumped out for five minutes prior to sampling or direct testing, after which it was acquired for on-site water quality testing or preserved

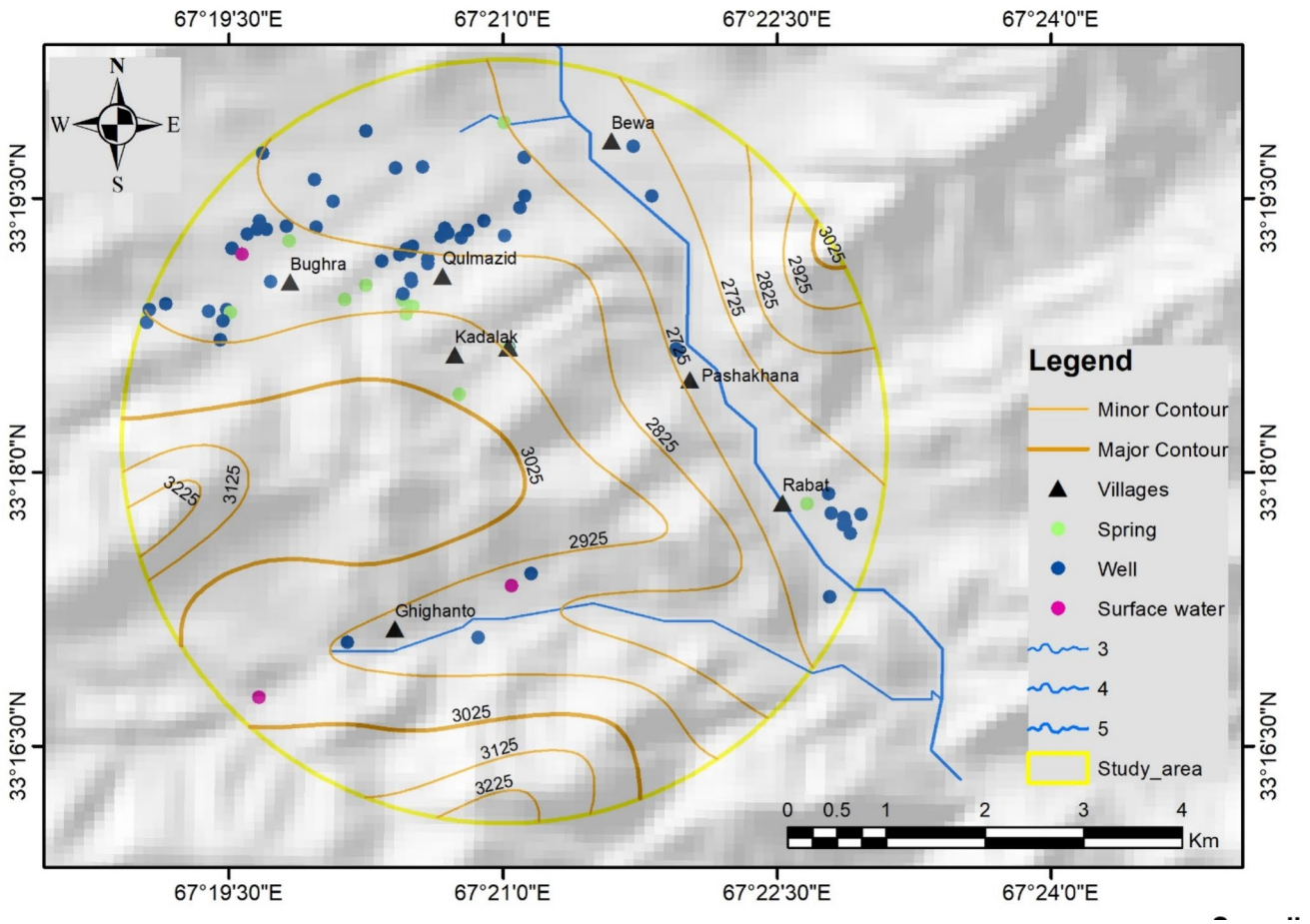



**Fig. 3** Location map of study area and sampling points

in HDPE bottles. In the springs, sampling is done directly from the mouth of the spring, where the local people usually install a channel to collect water.

Other samples were gathered using HDPE bottles. HDPE bottles are commonly used for storing water samples as they are chemically inert and do not react with the water or contaminate it. The cleanliness of these bottles is checked before sampling, and then they are washed with sample water three to four times. They are then carefully filled with source water samples at the sampling locations and immediately closed with their airtight cap. After collection, the HDPE bottles were securely sealed to prevent any leakage or contamination during transportation. For additional examination, these bottles were moved to the laboratory afterward.

## Collection of cancer data

The data collection and analysis method utilized a questionnaire to gather information on cancer cases across various villages systematically. The questionnaire was designed to capture essential details, including the type of cancer, age, gender, and the specific village of the respondents. In the village of Rabat, a local individual collected the data, which was then confirmed by the residents, ensuring the accuracy and reliability of the information. In the villages of Qulmazid and Kadalak, the questionnaire revealed two confirmed cancer-related deaths, offering valuable insight into the prevalence of cancer in these areas. Additionally, the questionnaire highlighted the existence of various types of cancer, particularly internal cancers, in parts of Bughra village, although specific numbers of cancer cases or deaths were not collected. While the data from Ghighanto village, where cancer-related deaths were reported to be significantly high, was not available. Since the study couldn't cover the areas of these last two villages from a water quality analysis perspective, the cancer data collection also did not cover them. Overall, the data collection involves obtaining information on cancer cases from multiple sources, including local individuals, residents, and reported cases from specific villages, while acknowledging the limitations of missing data for certain villages.

## Water quality analysis

Seventy-three water samples have been gathered, consisting of 59 from water wells, 11 from spring sources, and three from surface water sources (streams). The samples'pH, TDS, EC, and temperature were evaluated using the"Multifunction Water Tester."The instrument was calibrated using reagents with specific concentrations provided in its box. The $NO_3$ levels in the samples were determined using a portable"HI 3874 Nitrate Test Kit"from HANNA Instruments. The calorimetric technique used in this kit has a measuring range of 0 to 50 mg/L. Since the location of some sources (wells) were very close to each other. Most of the nitrate and iron tests showed the same results, so testing of all samples for iron and nitrate tests was omitted, and only a certain number of them were used, which covers the geospatial extent of the study area. 44 out of the 73 water samples underwent the nitrate test. A portable testing kit from HANNA instruments dubbed the"HI 3834 Iron Test Kit", was employed to determine the water samples'iron content. The testing kits were calibrated using samples with known concentrations to ensure the accuracy. The kit's calorimetric technique ranges from 0 to 5 mg/L. Forty of the seventy-three water samples were tested for iron concentration.

A portable testing kit known as the"S0005521 Arsenic test kit; extended range"was employed to ascertain the amount of arsenic contained in the samples. This kit uses the Gutzeit technique to visually detect soluble inorganic arsenic ($As^{+3}$ and $As^{+5}$) by creating a colored strip. With the color scale indicating distinct measurement ranges ("0–5–10–20–30–40–50–60–80–100–150–200–250–300–400–500 µg/L $As^{+3/+5}$"), the kit provides a detection range of up to 500 µg/L. The arsenic test was performed on all of the gathered water samples.

The accuracy assessment of the arsenic test kit was crucially informed by six water samples previously analyzed at the water lab of GreenTech Construction and Engineering Company in Kabul. Following concerns arising from cancer-related deaths near a specific water well (sample 2), these samples were examined for arsenic content. Out of six samples, only one exhibited an arsenic concentration of 10 µg/L, while the remaining five showed no detectable arsenic. This revelation, alongside multiple cases of various cancers in the area, prompted an immediate study by the authors. The results from the Green Tech water lab and the test kits showed perfect alignment, confirming the reliability of the arsenic test kits. Statistical analysis further substantiated this alignment with a mean squared error (MSE) of zero and a correlation coefficient of determination ($R^2 = 1$), highlighting the test kits'accuracy in detecting arsenic contamination in water samples (Table 1). These findings increase the working accuracy of these portable kits used in field studies and confirm their reliability. All study maps were provided using ArcGIS, and the study's data was analyzed using MS Excel.

**Table 1** Mean square error of the results obtained from the Green Tech water lab and testing kit used for the current study

| Sample No | GreenTech (µg/L) | Test Kit (µg/L) | Error | Squared Error |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 2 | 10 | 10 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |
| 22 | 0 | 0 | 0 | 0 |
| 47 | 0 | 0 | 0 | 0 |
| **MSE (Mean Square Error)** | | | | **0** |
| **R (Coefficient of Correlation)** | | | | **1** |
| **$R^2$ (Coefficient of Determination)** | | | | **1** |

## Descriptive statistical analysis

The data in this study have been categorized according to the type of source used for the water samples (wells, springs, and surface water sources). This categorization makes a more in-depth examination of the water quality from each source possible. A descriptive statistical approach is then employed to summarise and interpret the data. Several statistical measures are used to comprehend the properties and distribution of the qualitative parameters, including minimum, maximum, mean, and standard deviation. These parameters are then compared with the water quality guidelines provided by the WHO and the Afghanistan National Standard Authority (ANSA). This comparison enables the identification of deviations or non-compliance with established standards, highlighting potential issues or areas of concern regarding water quality.

## Arsenic health risk assessment

### Exposure assessment

There are several ways that arsenic may enter the human body, including by inhalation, skin contact, and the food chain. However, in measuring exposure to arsenic, the effect of intake through other means (such as inhalation) is typically considered negligible compared to oral intake. As a result, exposure to arsenic is primarily assessed and measured through two main pathways: oral (ingestion) and dermal (absorption through the skin). These two routes of exposure are considered the most significant and relevant when evaluating the potential health risks associated with arsenic exposure (Varol 2021). USEPA (2005) Recommends a lifelong average daily intake as an exposure measure to

estimate a person's daily exposure. The daily exposure for the ingestion route is represented by the following equation (Eq. 1):

$$ADI = (C \times IR \times EF \times ED)/(ABW \times ATL) \quad (1)$$

Based on the Equation, ADI indicates the average daily intake in (mg/kg/day), C represents the concentration of pollutants in drinking water in (mg/L), IR indicates ingestion rate per unit time (L/day), EF means the exposure frequency in (days/year), ED illustrates the exposure duration in (years), ABW presents average body weight (kg), and ATL defines average lifetime (days). A deterministic approach is utilized for exposure assessment to determine individual exposures to trace elements using Eq. (1). Following this approach, performing risk characterization is necessary to conduct a comprehensive risk assessment. The outcomes of exposure and toxicity evaluations are combined to calculate cancer risk and hazard indices. These approximations offer numerical measures to comprehend the possible hazards connected to the trace components.

### Human health risk assessment

This study assessed the relationship between trace element concentrations in water sources and the risk levels for chronic and cancerous diseases. The following equation can be used to calculate the Hazard Quotient (HQ):

$$HQ = \frac{ADI}{RfD} \quad (2)$$

The terms"ADI"and"RfD"stand for average daily intake and reference dose of arsenic, respectively. The USEPA has established 0.0003 mg/kg*day as the reference dosage for arsenic (Hoang et al. 2022). The HQ shows the non-cancer risk of one or more compounds and exposure routes. Assume that a substance's exposure level surpasses the reference dose value (RfD), resulting in a higher than 1 HQ value. In that case, it suggests that the substance may have potentially harmful effects that are non-carcinogenic. A higher HQ value indicates a greater likelihood of adverse health effects, including carcinogenic effects. The following formula (Eq. 3) is used to calculate the cancer risk (CR) associated with a trace element on human health:

$$CR = CSF \times ADI \quad (3)$$

ADI stands for average daily intake in milligrams per kilogram, CSF is the cancer slope factor in (mg/kg.day), and CR stands for carcinogen risk.

The CSF for the arsenic contaminant was specified by USEPA (2005) as 1.5 mg/kg/day. If the calculated carcinogenic risk (CR) is larger than 0.0001, it indicates a potential carcinogenic risk associated with exposure to the contaminant (USEPA 2005). A CR value exceeding 0.0001 suggests that the level of exposure is above the threshold for carcinogenic risk, raising concerns about the potential development of cancer.

## Results and discussion

### Descriptive statistics

Table 2 summarises the findings from the descriptive statistical analysis for the various kinds of water sources. The table presents the conclusions drawn from the study, which includes comparing the multiple parameters with the drinking water quality guidelines provided by the World Health Organization (WHO 2011) and the local regulatory body (ANSA 2013). The comparison helps assess the water sources'compliance with the established standards. The guideline values for three parameters (TDS, As, Fe) in the well's water have been surpassed. Among the 59 water samples collected from wells, two samples exceeded the guideline values for TDS and arsenic concentrations, while ten samples had iron concentrations that exceeded the specified guideline. Out of the nine tested springs'water, only arsenic was detected in two springs, surpassing the permissible limits. In all the surface waters that have been tested, no parameters have been found to exceed the allowable limits.

### Temperature

The temperature of water samples during the test ranges from 9.1 to 31 °C (Table 2). However, the arsenic test kit manual indicated that it would be good to perform the test with a water sample temperature ranging from 22 to 29 degrees Celsius. However, the test was done in differing water temperature ranges. The water temperature during the test is the temperature of the ambient environment. To comprehend how the kit should work correctly at different temperatures, a sample that tested positive for arsenic was retested at various temperatures (based on the given range), and the results were the same.

### pH

Water interacts with $CO_2$ in the ground to generate carbonic acid. Therefore, measuring water samples'pH is essential when determining how corrosive they are. The equilibrium between $CO_2$, $CO_3$, and $HCO_3$ significantly influences the pH of natural water. In the present study, the measured pH levels ranged from 6.95 to 8.43, aligning with the recommended range established by the WHO as indicated in Table 2. These results imply that the pH is not acidic in

**Table 2** The outputs of descriptive statistical analysis

| Parameter | Unit | Number of samples | Min | Max | Mean | SD | WHO Limit | > WHO | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | N | % |
| Source = Well | | | | | | | | | |
| Temperature | °C | 59 | 9.1 | 31 | 17.5 | 3.9 | NA | NA | NA |
| pH | | 59 | 6.95 | 8.43 | 7.6 | 0.3 | 6.5—8.5 | Nill | Nill |
| EC* | µS/cm | 59 | 305 | 1150 | 579.6 | 192.4 | 1500 | Nill | Nill |
| TDS | mg/L | 59 | 115 | 569 | 289.4 | 97.4 | 500 | 2 | 3.4 |
| As | µg/L | 57 | 0 | 20 | 1.4 | 4.1 | 10 | 2 | 3.5 |
| Fe | mg/L | 32 | 0 | 4.5 | 0.4 | 1.1 | 0.3 | 10 | 31.3 |
| NO3 | mg/L | 34 | 0 | 40 | 5.5 | 9.7 | 45 | Nill | Nill |
| Source = Spring | | | | | | | | | |
| Temperature | °C | 8 | 13.8 | 24.8 | 17.8 | 3.4 | NA | NA | NA |
| pH | | 8 | 7.04 | 8.2 | 7.6 | 0.4 | 6.5—8.5 | Nill | Nill |
| EC* | µS/cm | 8 | 210 | 841 | 596.0 | 200.1 | 1500 | Nill | Nill |
| TDS | mg/L | 8 | 105 | 420 | 298.8 | 99.4 | 500 | Nill | Nill |
| As | µg/L | 8 | 0 | 15 | 5.0 | 6.6 | 10 | 2 | 25.0 |
| Fe | mg/L | 3 | 0 | 0 | 0.0 | 0.0 | 0.3 | Nill | Nill |
| NO3 | mg/L | 4 | 0 | 0 | 0.0 | 0.0 | 45 | Nill | Nill |
| Source = Surface water | | | | | | | | | |
| Temperature | °C | 3 | 14.9 | 19 | 16.4 | 1.8 | NA | NA | NA |
| pH | | 3 | 8.1 | 8.2 | 8.2 | 0.0 | 6.5—8.5 | Nill | Nill |
| EC* | µS/cm | 3 | 334 | 485 | 386.7 | 69.6 | 1500 | Nill | Nill |
| TDS | mg/L | 3 | 166 | 245 | 194.0 | 36.1 | 500 | Nill | Nill |
| As | µg/L | 3 | 0 | 0 | 0.0 | 0.0 | 10 | Nill | Nill |
| Fe | mg/L | 1 | 0 | 0 | 0.0 | 0.0 | 0.3 | Nill | Nill |
| NO3 | mg/L | 2 | 0 | 10 | 5.0 | 5.0 | 45 | Nill | Nill |

*The permissible limit of EC is obtained from ANSA

the study region. The pH fluctuations that naturally occur in groundwater before it reaches the aquifer system are depicted in Eqs. (4) and (5).

1. The combination of rainfall and atmospheric gases leads to the formation of carbonic acid.

$$H_2O + CO_2 = H_2CO_3 \quad (4)$$

2. Acidity results from hydrogen ions being released when carbonic acid splits into bicarbonate.

$$H_2CO_3 = \left(HCO_3\right)^- + H^+ \quad (5)$$

## Electrical conductivity (EC)

The conductivity of water determines its ability to transmit electric current. According to the ANSA guideline, a maximum conductance level of 1500 µS/cm is permitted. The detectable range of electrical conductivity (EC) is between 210 and 1135 µS/cm. Groundwater quality is categorized based on EC concentration, as illustrated in Table 3 (Singh and Noori 2022). In the research area, 1.4% of the samples exhibited an excellent water type. Approximately 84% of the samples are good water, while 15% are permissible. Based on EC analysis, none of the water samples in the research region had EC values higher than 2000 µS/cm, suggesting that none were doubtful or inappropriate.

**Table 3** Classification of drinking water based on EC

| EC | Water type | Number of samples | Percentage of the samples (%) |
|---|---|---|---|
| < 250 | Excellent | 1 | 1.369 |
| 250—750 | Good | 61 | 83.561 |
| 750—2000 | Permissible | 11 | 15.000 |
| 2000—3000 | Doubtful | 0 | 0.000 |
| > 3000 | Unsuitable | 0 | 0.000 |
| | Total | 73 | 100 |

**Table 4** Classification of drinking water based on TDS concentrations

| TDS | Classification | No. of Samples | Percentage of samples (%) |
|---|---|---|---|
| < 500 | Desirable for drinking | 71 | 97.26 |
| 500–1000 | Permissible for drinking | 2 | 2.74 |
| 1000–3000 | Useful for irrigation | 0 | 0 |
| > 3000 | Unfit for drinking and irrigation | 0 | 0 |
| | Total | 73 | 100 |

**Table 5** Drinking water type based on TDS

| TDS | Classification | No. of Samples | Percentage of samples (%) |
|---|---|---|---|
| < 1000 | Fresh water type | 73 | 100 |
| 1000–10000 | Brackish water type | 0 | 0 |
| 10,000–100000 | Saline water type | 0 | 0 |
| > 100,000 | Brine water type | 0 | 0 |
| | Total | 73 | 100 |

## Total dissolved solids (TDS)

"Total dissolved solids"refers to the trace amounts of organic molecules and minerals in diluted situations. The main factors affecting TDS management are the deterioration, porosity, and penetration of natural rock formations. Human activities like fertilizer runoff and sewage discharge also contribute to TDS levels. According to the WHO guideline, the maximum recommended TDS value is 500 mg/L. This study's highest and lowest recorded TDS values were 567 mg/L and 105 mg/L, respectively (Table 2).

About 97% of the data indicate TDS levels below 500 mg/L, which is desirable for consumption. About 3% of the samples fall within the 500–1000 mg/L range, ideal for drinking. None of the samples had TDS values exceeding 1000 mg/L, categorized as suitable for irrigation but unfit for drinking and irrigation (Table 4). Based on the classifications in Table 5, all of the water samples (100%) fall under the freshwater type (Freeze and Cherry 1979).

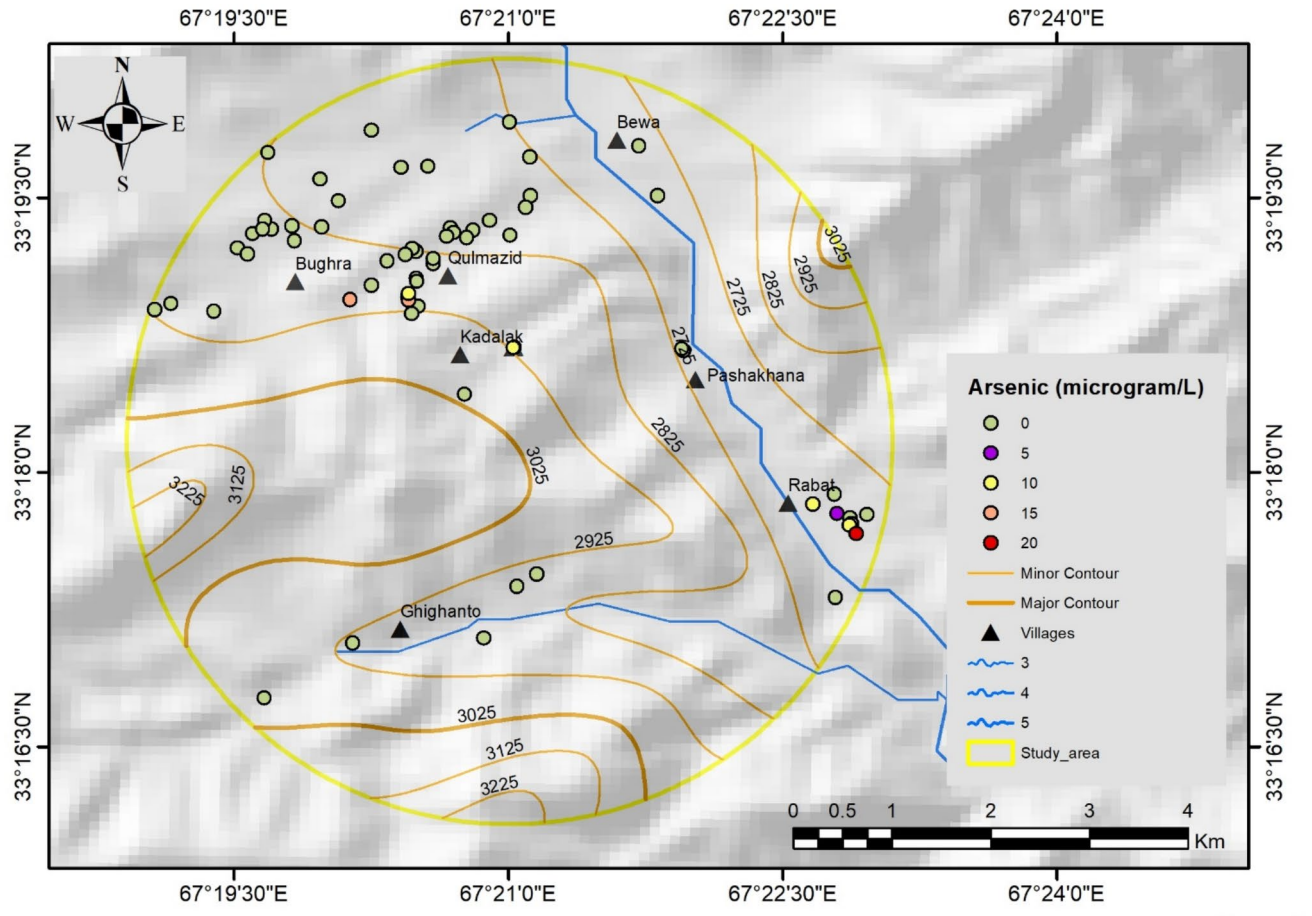



**Fig. 4** Arsenic concentration of the study area

## Arsenic (As)

According to the WHO guideline, the permissible concentration of arsenic in drinking water is limited to 10 µg/L. The concentration of arsenic in the current study area ranged from 0 to 20 µg/L (Table 2 and Fig. 4). Arsenic was identified in the villages of Qulmazid, Kadalak, Bughra, and Rabat in the current investigation. It was discovered that Qulmazid village's four water sources (shallow wells and springs) were contaminated with arsenic. Two shallow wells and one spring in its vicinity belong to a family that also has a death case due to stomach cancer (a man over 70 years old). Another contaminated spring is slightly far from the residential area and has a slight discharge. Its water was previously used for agricultural purposes, mixing with other spring sources located lower than it. Still, a family is situated next to this spring; its water is piped for domestic use. Arsenic was detected in a spring next to a shallow well in the Kadalak village. Households living next to this spring drank water from this contaminated and another neighboring spring (currently dried). There is also a death case due to neck and nose cancer (a man over 60 years old). Additionally, it was discovered that arsenic was present in a shallow water well in Bughra village (upper Bughra). Many sources contaminated with arsenic have been placed in Rabat village. There is a geothermal spring that contains arsenic pollution. The people who live around this geothermal spring have suffered from different types of cancer for a long time. Also, some of the tested shallow wells around this geothermal spring have arsenic concentrations.

## Iron (Fe)

The maximum iron concentration is about 4 mg/L in the area (Fig. 5). Iron in Bughra village has been distinguished in nine sources, including eight wells and one surface water source. Except for two wells, all remaining wells have an iron concentration of less than one mg/L. Iron was found in three wells among the resources checked in Qulmazid village. These three wells have a concentration of less than one mg/L. In Kadalak village, iron, which contains arsenic, has been observed in the spring.

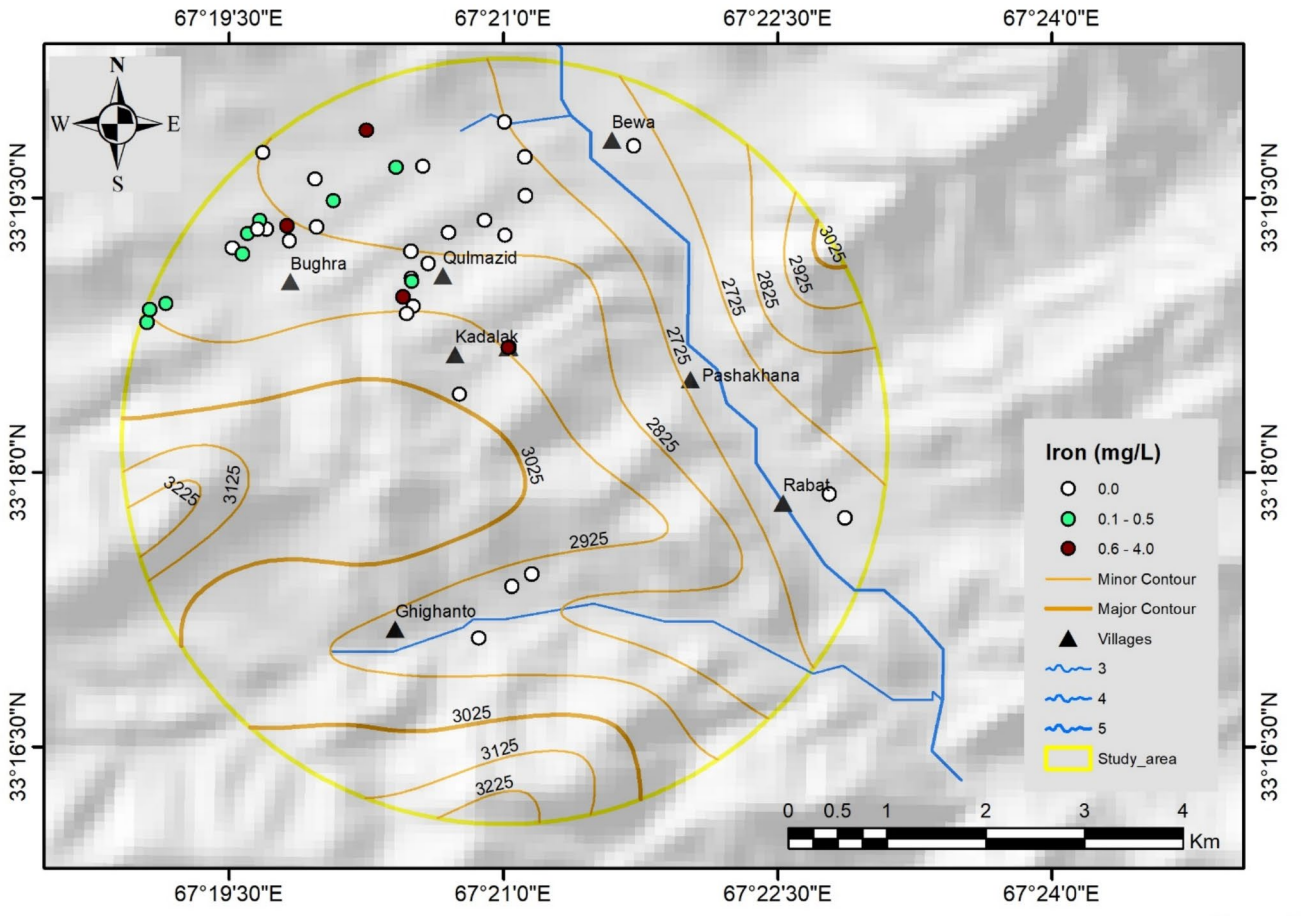



**Fig. 5** Iron concentration in the study area

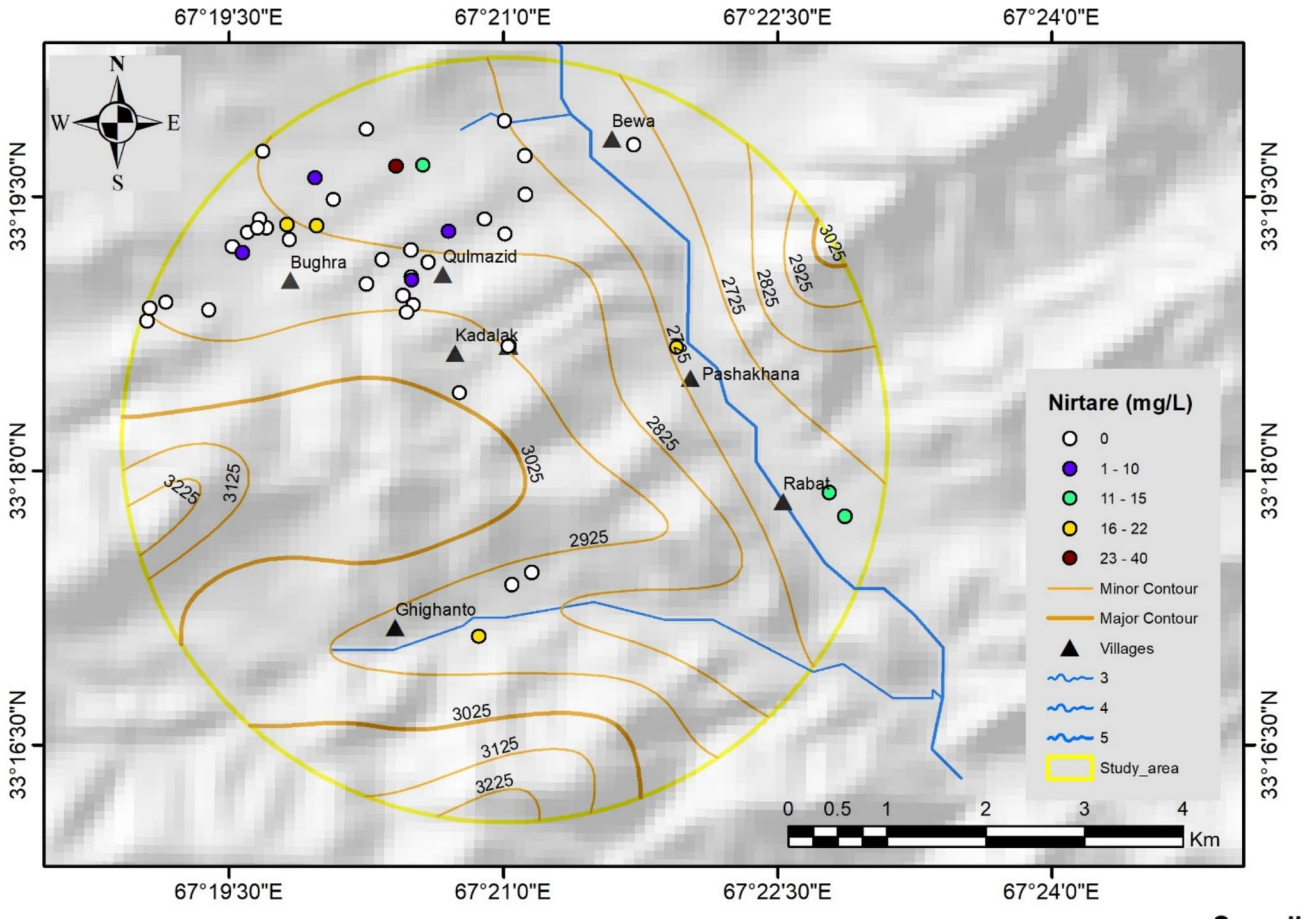



**Fig. 6** Nitrate concentration of the study area

## Nitrate ($NO_3$)

Nitrate in the study area ranges from zero to 40 mg/L (Fig. 6). Out of 34 water samples tested for nitrate concentration, only 12 samples have positive results. In Bughra village, six water samples have nitrate concentrations (5 wells and one surface water sample). Two samples in Qulmazid (wells), one in Pashakhana (well), two in Rabat (wells), and one in Ghighanto (well) have been confirmed. WHO determines the permissible limit of $NO_3$ at 45 mg/L. The samples with $NO_3$ concentration are below the guideline limit for drinking water. However, as can be seen, most of the wells and sources located in the middle or near the agricultural fields are all the sampling points where nitrate is present.

Frequent pollutants (iron and nitrate) in water sources can seriously endanger human health. Higher quantities of iron can give the clothes and plumbing fixtures a metallic flavor and discolor them. The main worry with iron, however, is that it may encourage bacterial growth in water distribution systems, which might result in the creation of biofilms and a decline in water quality. Conversely, leaching from fertilizers, sewage disposal, and agricultural runoff are the main ways nitrate gets into water sources. High nitrate consumption can affect the blood's capacity to transport oxygen, leading to methemoglobinemia, known as"blue baby syndrome,"especially in newborns and early children. Prolonged exposure to nitrate has also been linked to a higher chance of developing several types of cancers.

## Correlation matrix

The coefficient of correlation evaluation approach ascertains the relationship between numerical values. This approach involves calculating the correlation coefficient, which compares the concentrations of different water quality metrics. It helps in understanding the functioning of geochemical processes in groundwater. Correlation analysis assesses the magnitude and direction of the linear association between independent and dependent variables. The Pearson correlation coefficient is commonly used to determine the extent of the linear connection between these variables. (Rehman et al. 2018).

The correlation analysis in this study utilized Pearson correlation coefficients (R) to examine the relationships

**Table 6** Correlation matrix of the observed data

| *Correlation* | *Test T* | *pH* | *EC* | *TDS* | *Arsenic* | *Iron* | *Nitrate* |
|---|---|---|---|---|---|---|---|
| Test T | 1 | | | | | | |
| pH | 0.00663 | 1 | | | | | |
| EC | 0.05300 | -0.4796 | 1 | | | | |
| TDS | 0.05138 | -0.4901 | 0.99846 | 1 | | | |
| Arsenic | -0.28995 | -0.1371 | 0.12123 | 0.12055 | 1 | | |
| Iron | -0.12100 | 0.05161 | 0.03238 | -0.0140 | 0.62056 | 1 | |
| Nitrate | 0.12167 | 0.23135 | 0.25811 | 0.25475 | -0.1121 | -0.051 | 1 |

among seven water quality measures. The correlation matrix, presented in Table 6, provides the R values after conducting the correlation study on the groundwater samples in the study area. This analysis allows for a quick assessment of water quality based on the observed correlations between different characteristics of the water samples. It is evident from the computed Pearson correlation coefficient analysis that, with a correlation value of $R = +0.99846$, EC with TDS exhibits the strongest correlation. It suggests that their tendencies to fluctuate together are similar. The robust association indicates that alterations in EC may accurately anticipate shifts in TDS and vice versa. EC and TDS frequently indicate water salinity and the concentration of dissolved ions. This solid correlation implies that either EC or TDS alone can operate as a stand-in for the other, streamlining monitoring procedures.

In the research region, arsenic and iron have a modest correlation ($R = +0.62$). The moderate positive correlation between iron and arsenic indicates a tendency for greater arsenic concentrations to correlate with higher iron concentrations and vice versa. This discovery has important implications for research on groundwater pollution and geogenic metal sources. By directing focused mitigation efforts or health treatments, this link assists in identifying places where arsenic and iron may be present in increased amounts. The pH, EC, and TDS variables have a negative correlation (Table 6), indicating that when one of these variables rises, the others tend to fall. Because pH impacts the solubility of minerals and ions, it is essential to understand water chemistry. The presence of a negative association might suggest that fluids with more salinity or mineral content (higher EC and TDS) typically have lower pH values. This could have implications for the waters'appropriateness for specific applications, such as irrigation, or their potential for corrosiveness. Eventually, correlation analysis can be helpful for future studies in data interpretation, predictive models, risk assessment, and methodological consideration.

## Evaluation of arsenic pollution in groundwater

Arsenic is the twentieth most abundant element in the earth's crust and is extensively distributed. At 1–2 mg/kg, it may be found in the continental crust (Varol 2021). In addition to human activity, natural processes like decomposition and volcanic eruptions can release arsenic into the environment. It can potentially travel significant distances, either as suspended particles in water or as aerosols. When considering human health risks, the primary focus is on inorganic arsenic. It is mainly present as ($As5^{+}$) or ($As3^{+}$), with the prevalence of each form based on the surroundings. When arsenic is present in soil, it has the potential to contaminate various food products, including vegetables, fruits, and grains. This contamination can occur after drinking water or applying phosphate fertilizers. Additionally, the disposal of liquid and solid wastes from arsenic-containing pesticides and herbicides can contaminate water bodies and soil with arsenic.

Significant dangers to human health can result from acute and chronic health impacts caused by arsenic pollution in drinking water. When excessive amounts of arsenic are consumed acutely, symptoms such as intense stomach pain, vomiting, diarrhea, and circulatory collapse can occur right away. If treatment is delayed, these symptoms could even be fatal. Prolonged and continuous exposure to reduced amounts of arsenic has been linked to many serious health issues, such as skin, lung, and bladder malignancies, skin blemishes, cardiovascular problems, type 2 diabetes, and neurological abnormalities. Chronic effects from years of exposure can significantly reduce the quality of life and increase death rates in populations affected.

In the research area, the concentrations of arsenic in groundwater varied within the range of 5 to 20 µg/L (Fig. 4). According to the guidelines set by the World Health Organization (WHO 2011), the maximum acceptable level of arsenic in drinking water is 10 µg/L. Nevertheless, the investigation showed that arsenic concentrations in several groundwater samples in the research region were higher than this acceptable level. The current study assesses adults'and

children's ADI, HQ, and CR. According to Eq. 1, the C value represents the arsenic concentration in the tested water. The intake ingestion rate (IR) for adults is assumed to be 2 L/day, while for children, it is 1 L/day. In a single year, there are 365 days in which the exposure frequency (EF) occurs. For adults, the exposure duration (ED) is considered 50 years, while for children, it assumes 10 years. The average body weight (ABW) for adults is supposed to be 55 kg, whereas for children, it is 20 kg. Adults'average lifespan (ATL) is projected to be 18,250 days (50 years), and children's ATL is to be 5475 days (15 years).

## Health risk assessment

### Cancer deaths

The data collected from completed questionnaires were analyzed using Microsoft Excel to better understand the cancer cases in the studied villages. According to the data, there have been 23 cases of cancer deaths from 2000 to 2023 in Rabat village alone. Based on (Fig. 7), stomach cancer has the highest percentage (about 48%) in the study area. The next stage is skin cancer, which is about 13%. Bone and ventricle cancer both have almost the same percentage of about 9 percent. Other cancers such as bladder, blood, lung, rectum, and breast are also present in the mentioned village, each of which has a percentage of about 4%. These are usually cancers that are most probably caused by the presence of arsenic in drinking water (Martel et al. 2008; Lynch et al. 2011; Chen et al. 2011; Rosner et al. 2013).

The age and gender statistics are also presented in the (Fig. 8). In terms of age, the highest number of casualties is in people over 60 years old. According to the figure, the number of deaths due to different types of cancer is equal between men aged 60–70 and 70–85 (5 cases in each age group). However, in the case of women, this situation is a little different. The casualties between the age of 60–70 are slightly more than the age of 70–85. According to the figures available in the mentioned village, cancer deaths start at the age of 20 and continue until the age of 82. Still, the number of casualties among men is slightly higher than among women. It is worth mentioning that the water samples tested in Qulmazid and Kadalak villages that contain arsenic also have cases of deaths caused by cancer. In both cases, deaths due to cancer in two men are more than 60 years old.

### Arsenic health risk assessment

Contamination of drinking water with arsenic considerably impacts the prevalence of diseases associated with arsenic exposure. There are two primary categories of risk assessment: carcinogenic and non-carcinogenic. Variations

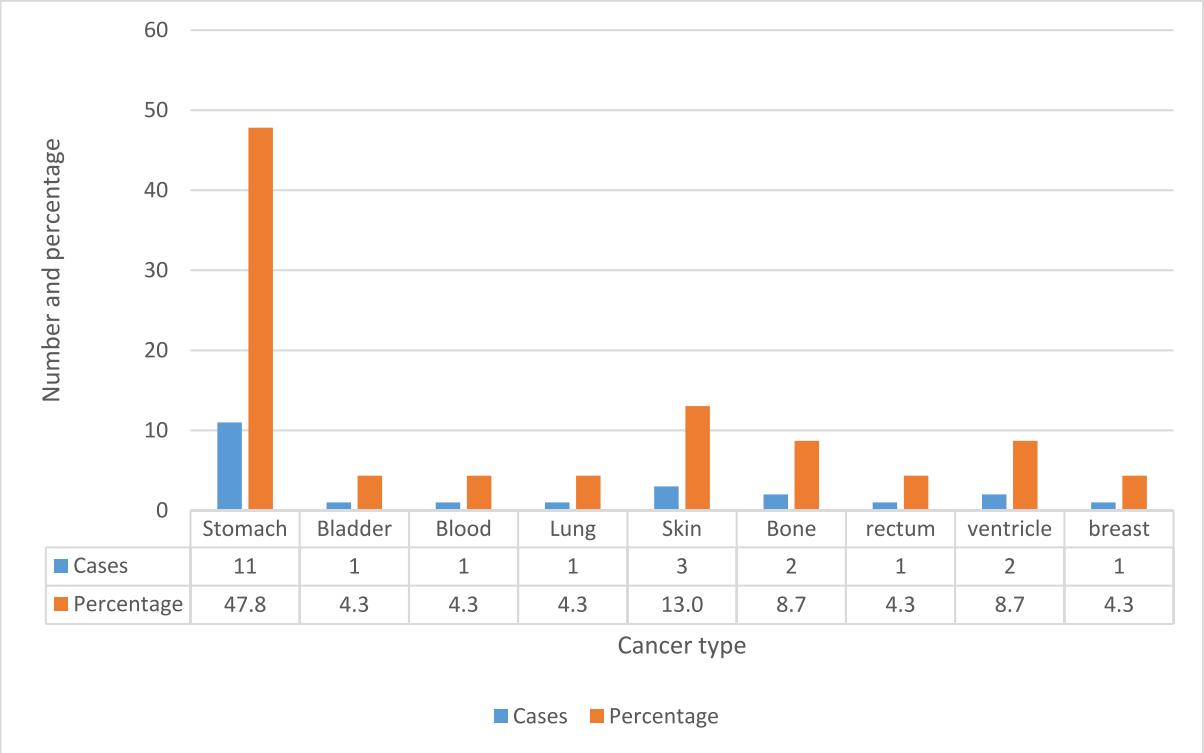



**Fig. 7** Fatalities due to various types of cancer in Rabat village from 2000 to 2023

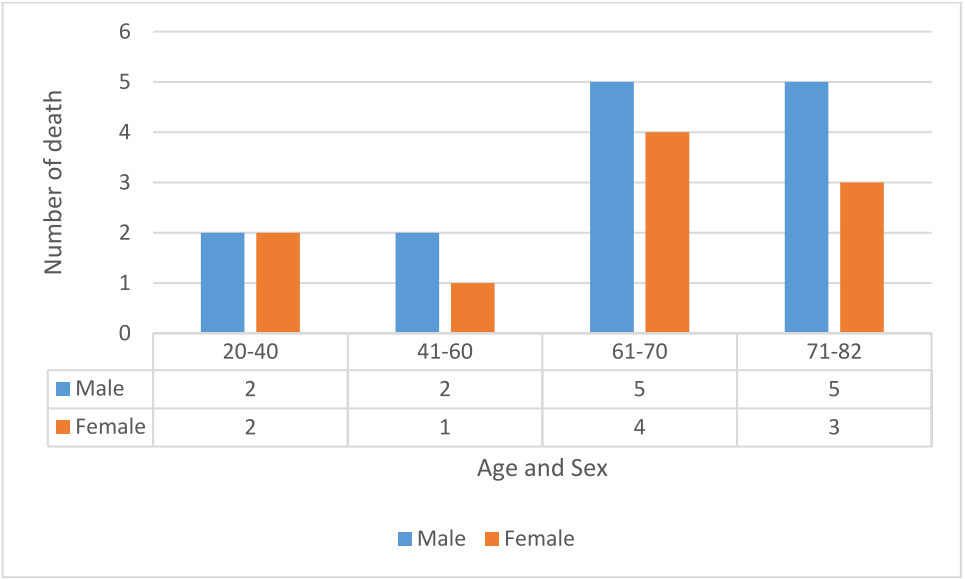



**Fig. 8** Fatalities due to various types of cancer in Rabat village from 2000 to 2023, considering age and gender

in drinking water consumption patterns, arsenic levels in different regions, and personal traits like body weight and height all affect how variable arsenic exposure is. As a result, variables such as daily intake, concentration, water drinking rate, body weight, and potential gender variances must be considered when assessing exposure and its carcinogenic consequences.

The research area's local population depends on groundwater from wells and springs for various uses, including drinking, irrigation, and household needs. This study conducted a health risk assessment to examine potential health risks associated with drinking water exposure to arsenic in Malistan. This assessment focused on chronic and carcinogenic effects, including ADI, HQ, and CR parameters. Table 7 presents the findings of individually calculating the health risk values for adults and children.

According to the data in Table 7, the ADI values for arsenic ranged from 0.000182 to 0.000727 mg/kg for adults and from 0.000167 to 0.000667 mg/kg for children. The non-cancerous hazard quotient values ranged from 0.606061 to 2.4242 for adults and 0.5555 to 2.2222 for children. The carcinogenic risk values ranged from 0.000273 to 0.001091 for adults and from 0.00025 to 0.001 for children.

Carcinogenic risk is the likelihood that a person may get cancer at some point in their life as a result of being exposed to a chemical that has harmful effects. Experts and regulators have established tolerable risk values for poisonous substances, including arsenic, ranging from $10^{-6}$ to $10^{-4}$ (Varol 2021). As shown in Table 7, all HQ values, except

**Table 7** Adult and pediatric hazard quotients (HQ), non-carcinogenic risk (NCR), and carcinogenic risk (CR)

| Sample | Village name | Source type | Lat | Long | C (µg/L) | Adult | | | Child | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | ADI | HQ | CR | ADI | HQ | CR |
| 2 | Qulmazid | Well | 33.316007 | 67.340822 | 15 | 0.000545 | 1.818182 | 0.000818 | 0.0005 | 1.666667 | 0.00075 |
| 6 | Kadalak | Spring | 33.311341 | 67.35042 | 10 | 0.000364 | 1.212121 | 0.000545 | 0.000333 | 1.111111 | 0.0005 |
| 26 | Rabat | Spring* | 33.29709 | 67.377745 | 10 | 0.000364 | 1.212121 | 0.000545 | 0.000333 | 1.111111 | 0.0005 |
| 48 | Qulmazid | Spring | 33.315756 | 67.335577 | 15 | 0.000545 | 1.818182 | 0.000818 | 0.0005 | 1.666667 | 0.00075 |
| 53 | Rabat | Well | 33.296233 | 67.379969 | 5 | 0.000182 | 0.606061 | 0.000273 | 0.000167 | 0.555556 | 0.00025 |
| 56 | Rabat | Well | 33.295295 | 67.381271 | 10 | 0.000364 | 1.212121 | 0.000545 | 0.000333 | 1.111111 | 0.0005 |
| 57 | Rabat | Well | 33.295158 | 67.381074 | 10 | 0.000364 | 1.212121 | 0.000545 | 0.000333 | 1.111111 | 0.0005 |
| 58 | Rabat | Well | 33.294424 | 67.381662 | 20 | 0.000727 | 2.424242 | 0.001091 | 0.000667 | 2.222222 | 0.001 |
| 62 | Qulmazid | Spring | 33.315685 | 67.34089 | 15 | 0.000545 | 1.818182 | 0.000818 | 0.0005 | 1.666667 | 0.00075 |
| 63 | Qulmazid | Well | 33.316264 | 67.340885 | 10 | 0.000364 | 1.212121 | 0.000545 | 0.000333 | 1.111111 | 0.0005 |
| 69 | Bughra | Well | 33.312058 | 67.324216 | 10 | 0.000364 | 1.212121 | 0.000545 | 0.000333 | 1.111111 | 0.0005 |

*It is a geothermal spring

for one sample, are greater than 1, indicating potential non-cancerous health effects from the mentioned water sources. Similarly, all CR values for children and adults in Table 7 are greater than 0.0001, meaning a possible risk of different types of cancer. The presence of cancer-related deaths in the study area further supports this finding (Figs. 7 and 8). Consequently, it is necessary to conduct additional tests on the water sources in these areas, identify alternative water sources for the local population, and take measures to mitigate pollution to address these issues.

The models used for assessing the health hazards associated with arsenic exposure, such as ADI, HQ, and CR, have various limitations and function under particular assumptions. Although ADI may ignore individual susceptibilities and uncertainty in toxicity data, it assumes a linear or non-linear dose–response relationship based on chronic exposure scenarios and standard population factors. Assuming exposure happens through a single channel, such as ingestion, HQ evaluates health risks by comparing predicted exposure levels to reference values. This streamlines risk assessment but could understate the dangers arising from several paths. Assuming a linear correlation between lifetime arsenic exposure and cancer incidence, CR calculates cancer risks by extrapolating data from high-dose studies. Its precision is, however, limited by limitations in data extrapolation and the exclusion of co-exposure effects. These models are quite useful.

#### Recommendations

Based on its findings, the present study makes the following recommendations to improve the health and hygiene conditions in the region. These recommendations are believed to raise awareness and attract the interest of locals, officials, and scholars.

- **Alternative Water Sources:** Investigate and create substitute arsenic-free water sources to give contaminated areas safer and longer-lasting choices. Most of the studied area has arsenic-free waters near residential houses at a further distance. Therefore, the community should use them for domestic uses.
- **Water Treatment Technology:** To ensure that communities have safe drinking water, water treatment technology should be employed to lower the amounts of arsenic in polluted sources. Families in rural areas may use small membrane filters that can remove arsenic for domestic purposes. Or larger filters that can filter more water should be installed near several houses so that several families can use them.
- **Public Awareness initiatives:** Inform the community about the dangers of arsenic exposure and encourage safe water consumption practices. At the local community level, awareness workshops should be undertaken to inform the local population about water contamination and its possible hazards.
- **Community Engagement:** Ensure local communities actively participate in and comprehend the steps taken to mitigate arsenic poisoning by involving them in decision-making processes linked to water resource management.
- **Healthcare efforts:** With an emphasis on cancer prevention and early diagnosis, implement healthcare efforts for people living in locations where recorded incidences of arsenic-related health disorders include routine health check-ups and medical support.
- **Regular Water Quality Monitoring:** To track changes and quickly address new concerns, ensure water quality is regularly monitored and tested.
- **Collaboration with Health Authorities:** Work with health authorities to create health intervention programs that provide the support and medical care needed by people more vulnerable to arsenic exposure.
- **Research and Innovation:** Encourage more study and creative thinking in water purification technology, emphasizing long-term and reasonably priced ways to lower the levels of arsenic in tainted water.
- **International Partnerships:** For continued assistance, resources, and knowledge in resolving water quality problems, fortify alliances with foreign groups.

There are obstacles as well as opportunities to put these measures into action. Other water sources are available, but getting to them might be financially and logistically tricky in certain remote places with poor infrastructure. Membrane filter technologies are practical for residential usage. However, sustained operation and maintenance of these devices require a steady supply of resources and knowledge. It is feasible for locals to pay for this themselves, but providing them with sustainable energy sources is difficult. Initiatives to raise public awareness are doable, but they need funding to reach various demographics and overcome literacy difficulties. Healthcare initiatives like routine examinations are feasible but could require more financial and infrastructural support, particularly in rural regions. Building capability and securing long-term financing are necessary for routine water quality monitoring. Still, their financial requirements are enormous because they require much money to perform water quality testing. Innovation and research are feasible, but they need to be supported by investments in infrastructure and scientific talent. It is conceivable to fortify international relationships, but doing so calls for negotiating geopolitical challenges and guaranteeing reciprocal advantages. Local collaborations, suitable tactics, and ongoing international support are needed to address these issues and enhance the community's water quality and public health outcomes.

## Conclusion

Arsenic in drinking water can lead to severe health consequences, including skin lesions, cardiovascular diseases, and various types of cancer. Most arsenic is naturally present in groundwater, a drinking water source without treatment in many regions. Groundwater is an essential water supply for drinking, residential use, and agriculture in the Malistan District of Ghazni Province. Arsenic, a toxic mineral commonly found in groundwater, presents a serious public health threat due to its potential to cause cancer and arsenicosis. The study sought to evaluate the health concerns associated with arsenic, iron, and nitrate levels in the drinking water sources within the selected region. Seventy-three water samples from shallow wells, springs, and surface water sources were collected for the investigation. Various portable tools and kits were utilized to test temperature, pH, TDS, EC, $NO_3$, Fe, and As.

The data were categorized according to the kind of water source using descriptive statistical analysis, which allowed for a more in-depth investigation of the water quality of each source. Statistical measures such as min, max, mean, and standard deviation were calculated and compared with WHO and ANSA drinking water quality guidelines to identify potential deviations or non-compliance. The study also conducted an arsenic health risk assessment. The human health risk assessment included evaluating chronic and carcinogenic risk levels, with calculations for Hazard Quotient (HQ) and Cancer Risk (CR) using specified formulas and reference doses.

Of the 59 shallow well water samples, two exceeded the guidelines for TDS and As, and 10 surpassed the iron limit. Only arsenic exceeded the limits of eight spring water samples in two springs. No parameters exceeded limits in tested surface waters. pH in the samples is in the acceptable range. Based on EC, most tested samples fall into the good category. According to TDS, most water is fresh and desirable for drinking.

The arsenic contamination was identified in multiple springs and wells with concentrations ranging from 5 to 20 µg/L. Iron concentrations of up to 4.5 mg/L are also detected in some sources. Nitrate concentration appeared in most villages with a maximum concentration of up to 40 mg/L. The Pearson correlation coefficient analysis shows that EC with TDS has the highest correlation ($R = + 0.99846$), and a moderate correlation exists between arsenic and iron ($R = + 0.62$).

There have been 23 cases of cancer deaths from 2000 to 2023 in Rabat village alone. Stomach cancer has the highest percentage (about 48%) in the study area within the age of 60–80 years old. Except for one sample, all HQ values exceed 1, indicating potential noncancerous health effects from the mentioned water sources. Similarly, all CR values for children and adults are greater than 0.0001, meaning a possible risk of different types of cancer.

The study has some limitations, including the use of cross-sectional data, which makes it difficult to determine a very accurate causal correlation between health outcomes and arsenic exposure and variations in arsenic levels between demographic groups and across time. Imbalanced sample distribution and small sample numbers might compromise the study's generalizability. Prospective research avenues may concentrate on continued investigations to elucidate long-term health consequences and integrate biomonitoring to achieve more precise exposure evaluations. To improve knowledge and directly focused health interventions, it would be beneficial to investigate the effectiveness of mitigation options such as alternate water sources and treatment technologies in various situations and the connections between arsenic and other contaminants or genetic variables.

**Author contribution** All authors contributed to the study's conception and design. Ali Reza Noori has done field studies, sample collection and testing. Ali Reza Noori and Assadullah Rezai have done data analysis and the first draft of the manuscript. S.K. Singh has made the final review and approval. All authors read and approved the final manuscript.

**Funding** The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Data availability** Data of the manuscript will be provided from the corresponding author based on request.

**Declarations**

All authors have read, understood, and have complied as applicable with the statement on"Ethical responsibilities of Authors"as found in the Instructions for Authors.

**Competing interests** The authors declare no competing interests.

## References

Abedin MJ, Cresser MS, Meharg AA et al (2002) Arsenic accumulation and metabolism in rice ( Oryza sativa L.). Environ Sci Technol 36:962–968. https://doi.org/10.1021/es0101678

Alao JO, Fahad A, Abdo HG et al (2023) Effects of dumpsite leachate plumes on surface and groundwater and the possible public health risks. Sci Total Environ 897:165469. https://doi.org/10.1016/j.scitotenv.2023.165469

ANSA (2013) National drinking water quality standard. ANSA, Kabul

Ayejoto DA, Egbueri JC (2024) Human health risk assessment of nitrate and heavy metals in urban groundwater in Southeast Nigeria. Ecol Front 44:60–72. https://doi.org/10.1016/j.chnaes.2023.06.008

Ayejoto DA, Agbasi JC, Nwazelibe VE et al (2023) Understanding the connections between climate change, air pollution, and human health in Africa: Insights from a literature review. J Environ Sci Heal Part C 41:77–120. https://doi.org/10.1080/26896583.2023.2267332

Ayejoto DA, Agbasi JC, Egbueri JC, Echefu KI (2024) Assessment of oral and dermal health risk exposures associated with contaminated water resources: an update in Ojoto area, southeast Nigeria. Int J Environ Anal Chem 104:641–661. https://doi.org/10.1080/03067319.2021.2023515

Ayeni EA, Aldossary AM, Ayejoto DA et al (2022) Neurodegenerative diseases: implications of environmental and climatic influences on neurotransmitters and neuronal hormones activities. Int J Environ Res Public Health 19:12495. https://doi.org/10.3390/ijerph191912495

Bhowmick S, Pramanik S, Singh P et al (2018) Arsenic in groundwater of West Bengal, India: a review of human health risks and assessment of possible intervention options. Sci Total Environ 612:148–169. https://doi.org/10.1016/j.scitotenv.2017.08.216

Brammer H, Ravenscroft P (2009) Arsenic in groundwater: a threat to sustainable agriculture in South and South-east Asia. Environ Int 35:647–654. https://doi.org/10.1016/j.envint.2008.10.004

Chen CJ, Chiou HY, Hsu LI (1988) Arsenic and cancer. Lancet 1:414–415

Chen Y, Graziano JH, Parvez F et al (2011) Arsenic exposure from drinking water and mortality from cardiovascular disease in Bangladesh: prospective cohort study. BMJ 342:d2431–d2431. https://doi.org/10.1136/bmj.d2431

Freeze RA, Cherry JA (1979) Groundwater. Englewood Cliffs, p 604

Hayat E, Baba A (2017) Quality of groundwater resources in Afghanistan. Environ Monit Assess 189. https://doi.org/10.1007/s10661-017-6032-1

Hoang MN, Le Vo P, Bui TV, et al (2022) Health risk assessment of arsenic in drinking groundwater: a case study in a central high land area of Vietnam. IOP Conf Ser Earth Environ Sci 964. https://doi.org/10.1088/1755-1315/964/1/012010

Jawadi HA, Malistani HA, Moheghy MA, Sagin J (2021) Essential trace elements and arsenic in thermal springs, Afghanistan. Water (Switzerland) 13:1–16. https://doi.org/10.3390/w13020134

Jomova K, Jenisova Z, Feszterova M et al (2011) Arsenic: toxicity, oxidative stress and human disease. J Appl Toxicol 31:95–107. https://doi.org/10.1002/jat.1649

Joya M, Stanikzai Z, Akbarzadeh I et al (2020) Prevalence of cancers diagnosed in Jamhuriyat Hospital, Kabul. Afghanistan Heliyon 6:e03603. https://doi.org/10.1016/j.heliyon.2020.e03603

Lynch AM, Sasaki JC, Elespuru R et al (2011) New and emerging technologies for genetic toxicity testing. Environ Mol Mutagen 52:205–223. https://doi.org/10.1002/em.20614

Mahaqi A, Mehiqi M, Rahimzadeh M et al (2021) Dominant geochemical reactions and hazardous metal contamination status in the Kabul's aquifers, Afghanistan. Int J Environ Sci Technol 18:4043–4052. https://doi.org/10.1007/s13762-020-03098-w

Mandal BK, Suzuki KT (2002) Arsenic round the world: a review. Talanta 58:201–235

Martel M-J, Rey E, Malo J-L et al (2008) Determinants of the Incidence of childhood asthma: a two-stage case-control study. Am J Epidemiol 169:195–205. https://doi.org/10.1093/aje/kwn309

Noori AR, Singh SK (2021) Spatial and temporal trend analysis of groundwater levels and regional groundwater drought assessment of Kabul, Afghanistan. Environ Earth Sci 80. https://doi.org/10.1007/s12665-021-10005-0

Noori AR, Singh SK (2021) Assessment and modeling of sewer network development utilizing Arc GIS and SewerGEMS in Kabul city of Afghanistan. J Eng Res 22–31. https://doi.org/10.36909/jer.ICARI.15287

Noori AR, Singh SK (2023) Assessment of seasonal groundwater quality variation employing GIS and statistical approaches in Kabul basin, Afghanistan. Environ Dev Sustain. https://doi.org/10.1007/s10668-022-02876-5

Noori AR, Singh SK (2023) Rainfall Assessment and water harvesting potential in an urban area for artificial groundwater recharge with land. Water Resour Manag. https://doi.org/10.1007/s11269-023-03602-0 Rainfall

Noori AR, Singh SK (2021) Status of groundwater resource potential and its quality at Kabul, Afghanistan : a review. Environ Earth Sci 80:1–13. https://doi.org/10.1007/s12665-021-09954-3

Noori AR, Singh SK (2024) Delineation of optimal locations for artificial groundwater recharge utilizing MIF and GIS in a semi-arid area. Environ Earth Sci 83:33. https://doi.org/10.1007/s12665-023-11338-8

National Statistics and Information Authority (NSIA). (2024). National population estimate for the year 1403. Kabul. Retrieved from https://nsia.gov.af

Rasouli H, Vaseashta A (2023) Groundwater Quality Assessment in Pul-e-Charkhi Region, Kabul Afghanistan. Adv Geol Geotech Eng Res 5(4):1–21. https://doi.org/10.30564/agger.v5i4.5949

Ravenscroft P, Brammer H, Richards K (2009) Arsenic pollution: a global synthesis. Wiley-Blackwell

Rehman F, Cheema T, Lisa M et al (2018) Statistical analysis tools for the assessment of groundwater chemical variations in wadi bani malik area, saudi arabia. Glob Nest J 20:355–362. https://doi.org/10.30955/gnj.002237

Riaz RMY, Murtaza G, Farooqi ZUR et al (2022) Assessment of arsenic contamination in groundwater and associated human health risk. Sustainability 14:12460. https://doi.org/10.3390/su141912460

Rosner B, Glynn RJ, Tamimi RM et al (2013) Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers. Am J Epidemiol 178:296–308. https://doi.org/10.1093/aje/kws457

Saffi MH, Kohistani AJ (2016) Arsenic contamination of groundwater in panjsher province. DACAAR Retrieved from https://www.dacaar.org/groundwater

Shaji E, Santosh M, Sarath KV et al (2021) Arsenic contamination of groundwater: a global synopsis with focus on the Indian Peninsula. Geosci Front 12:101079. https://doi.org/10.1016/j.gsf.2020.08.015

Singh SK, Noori AR (2022) Groundwater quality assessment and modeling utilizing water quality index and GIS in Kabul Basin, Afghanistan. Environ Monit Assess 194. https://doi.org/10.1007/s10661-022-10340-0

Smedley PL, Kinniburgh DG (2013) Arsenic in groundwater and the environment. In: Selinus O (ed), Essentials of medical geology: Impacts of the natural environment on public health. Springer Netherlands, pp 279–310. https://doi.org/10.1007/978-94-007-4375-5_12

Smith AH, Hopenhayn-Rich C, Bates MN et al (2000) Cancer risks from arsenic in drinking water. Environ Health Perspect 108:639–642

Smith AH, Lingas EO, Rahman M (2000) Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. Bull World Health Organ 78:1093–1103

Smith AH, Marshall G, Yuan Y et al (2006) Increased mortality from lung cancer and bronchiectasis in young adults after exposure to arsenic in utero and in early childhood. Environ Health Perspect 114:1293–1296. https://doi.org/10.1289/ehp.8832

UNHCR Field Office Ghazni (2003) Malistan, Afghanistan: country of origin information. UNHCR. Retrieved May 19, 2025, from https://www.ecoi.net/en/file/local/1226640/10552_accord216_malistan.pdf

U.S. Environmental Protection Agency (2005) Guidelines for carcinogen risk assessment. EPA/630/P-03/001F. https://www.epa.gov/sites/default/files/2013-09/documents/cancer_guidelines_final_3-25-05.pdf

Van Geen A, Cheng Z, Jia Q et al (2007) Monitoring 51 community wells in Araihazar, Bangladesh, for up to 5 years: implications for arsenic mitigation. J Environ Sci Heal Part A 42:1729–1740. https://doi.org/10.1080/10934520701564236

Varol S (2021) Potential health risk assessment related to arsenic pollution and hydrogeochemistry of groundwaters in Akşehir and surroundings (Konya/Turkey). J Water Health 19:97–107. https://doi.org/10.2166/WH.2020.107

World Health Organization (2011) Guidelines for drinking-water quality. 4th ed. https://iris.who.int/bitstream/handle/10665/44584/9789241548151_eng.pdf

Wikipedia contributors (2025) Malistan District. https://en.wikipedia.org/wiki/Malistan_District

Zahidi NM, Ansari A, YahyaSargand G, Achak J (2020) Analysis of heavy metals in surface and groundwater at Lashkergah City, Helmand-Afghanistan. Int J Sci Technol Res 9:87–91

**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.

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-025-10233-7

RESEARCH ARTICLE–ELECTRICAL ENGINEERING

# Brain-Inspired Learning Controller for Grid-Connected Electric Vehicle Charging Junction with Improved Power Quality

**Amarendra Pandey**[1] · **Alka Singh**[1] · **Sudhanshu Mittal**[1,2]

Received: 20 August 2024 / Accepted: 14 April 2025
© King Fahd University of Petroleum & Minerals 2025

**Abstract**

In this paper, a brain-inspired learning (BIL) controller for grid-connected electric vehicle (EV) charging junction is developed and analyzed. The system configuration combines a utility grid, nonlinear loads, a voltage source converter (VSC), and a bidirectional EV charger. The system is controlled and analyzed to study the smooth operation of battery charge and discharge of the EV. Further, a VSC is realized as a custom power device (CPD) that improves power quality (PQ) problems caused by nonlinear loads, providing reactive power support and correcting unbalances in the grid current. The paper focuses on designing a biologically inspired intelligent control technique for controlling the VSC. The approach mimics the mammalian brain behavior and inspires the development of brain-inspired learning (BIL) techniques for controlling VSC. The developed controller is further synchronized with a bidirectional EV charger to regulate smooth operation in both grid-to-vehicle (G2V) and vehicle-to-grid (V2G) conditions. The designed BIL technique offers the advantages of quick learning, minimum error, and fast response with medium computational complexity. The response of the designed controller under grid variations is also presented. A suitable experimental setup is developed and investigated for its performance evaluation under steady-state and dynamic load situations. The comparative analysis of the BIL control algorithm with synchronous reference frame (SRF) and least mean square (LMS) techniques during load change is presented. The paper also presents a detailed analysis using Bode plots.

**Keywords** Harmonics · Brain-inspired learning · Nonlinear load · Power quality · Voltage source converter

## 1 Introduction

Over the past decade, Electric cars (EVs) have dominated the vehicle market worldwide. This noteworthy shift has been witnessed owing to multiple noteworthy benefits, such as fewer greenhouse gas emissions and improved vehicular performance at affordable cost. Furthermore, an electric vehicle (EV) might be considered a portable energy storage unit, which plays a significant role in adopting electric mobility. It can be integrated into a microgrid to make better use of the energy charged in the battery [1–5]. In the UK, electric cars (EVs) are predicted to make up 60% of all forthcoming automobile sales by 2030 while developing countries are also rapidly catching up [6].

Given the scale of EVs, both V2G and G2V methods have become increasingly prominent and feasible. Today EVs are being deployed for storing and supplying energy to manage dynamic demand effectively [7]. This is interesting since, due to the nature of power electronic circuits and switching activities, EVs not only show nonlinear characteristics as loads when running in G2V mode but also can operate as intermittent energy sources in V2G mode due of the unpredictable nature of energy consumption patterns [8, 9]. This paper proposes a two-stage onboard charger topology containing a single active bridge with an LLC resonant converter to reduce switching loss and improve the charger's efficiency [10]. Paper [11] proposes an interleaved DC side converter with an AC-DC converter that is utilized to reduce inrush current due to soft switching. Using a sinusoidal source current, the VSC also controls DC link voltage, therefore offering a high power factor. Consequently, power systems could

✉ Sudhanshu Mittal
sudhanshumittal_2k21phdee04@dtu.ac.in

[1] Electrical Engineering Department, Delhi Technological University, Rohini, New Delhi, India

[2] Electrical and Instrumentation Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Published online: 22 May 2025

encounter technical challenges when accommodating large-scale bidirectional EVs. This paper addresses key factors such as power quality and stability as impacted by bidirectional EV integration.

The power quality (PQ) of the power distribution system has also declined due to the increased deployment of loads based on power electronics. Some examples of such loads include diode rectifiers, drives, photocopiers, fax machines, computers, and motor loads with high reactive power demand. Due to this, traditional analysis and control of power systems based on dominant linear loads has undergone a significant change. High proliferation of nonlinear loads at the consumer results in considerable distortion in feeder currents and bus voltage [12, 13]. Voltage and current harmonic distortion are undeniably harmful and cause a reduction in the life of cables and other equipment due to heating, harmonic de-rating, increased power losses, false tripping of protection devices, and radio interference [14–16]. The literature study covers various PQ problems that have been reduced in the distribution system by using a VSC.

Paper [17] proposes a grid-integrated Ultra Fast Charging Station (UFCS) that manages the power flow between the grid and the EV sources using a Linear Programming technique. The design of UFCS includes Photovoltaic (PV) sources and a battery rack system utilized to charge several electric vehicles at a time while also regulating the power flow in the system. However, the designed control is quite complex in nature.

Figure 1 shows the basic block diagram of the system, showing a VSC connected in shunt configuration for a grid-connected single-phase distribution system along with an EV charger. The design of the VSC controller influences the load compensation and accuracy of compensation. In recent years, various control techniques have been proposed for the control of VSC, including conventional and adaptive control algorithms. The traditional algorithms include the SRF theory [18], instantaneous reactive power (IRP) theory [19], second-order generalized integrator (SOGI) [20] and their variants. The SRF controller shows higher mathematical complexity and slow response due to the transformation and processing of signals. The structure and tuning of the IRPT controller require high sampling frequency, and it can be intricate, necessitating a comprehensive comprehension of control theory and power electronics. Some new control methods are mentioned in [21, 22]. In paper [21], the PV array integrated multiport multipurpose charging station is implemented using a reweighted sparse-least-mean mixed-norm (RS-LMMN) adaptive control method. A penalty factor for the sparsity of the input signal matrix is also employed by the RS-LMMN method to increase the convergence speed. However, this again adds to the complexity of the controller. A finite control set model predictive control (FCS-MPC) system regulates the proposed three-port converter (TPC). A hybrid system with two DC and one AC port can have its power control handled by the TPC, which, in comparison with methods, claims to have better power density and simple topology [22]. Thus, recent literature review points to numerous designed controllers, some of which are also adaptive in nature.

An adaptive controller like the leaky least mean fourth control (LLMF) algorithm [23] is highly sensitive to step size and leaky factor selection. In the paper [24], the single-phase grid-connected portable EV charger is designed. The front-end VSC is controlled by applying the frequency fixed second-order generalized integrator (FF-SOGI) algorithm. Paper [25] proposes a PQ-theory-based control technique for grid-connected VSC with EV chargers and renewable energy sources (RES) to enable reactive power to support power quality improvement. The grid-tied off-board charger is controlled using the Joint Logarithmic Hyperbolic Cosine Robust Sparse Adaptive Filter (JLHCAF) method [26]. Adaptive controllers designed in [23–26] are effective but possess additional complexity. Moreover, all the design parameters must be sufficiently tuned to produce the desired results.

Using the TL boost converter, a Brayton-Moser passivity-based controller has been designed in the article [27] for an onboard battery charger to provide well-regulated voltage, current, and almost unity power factor operating when charging an electric car battery. With an extra switch, diode, and capacitor, the TL boost converter operates its switches 180° out of phase. These additional components and interleaved operation substantially lower the switch stress and filter size.

Recent literature review suggests developing several intelligent techniques like artificial neural networks (ANNs) and adaptive neural fuzzy systems, and these algorithms show a high rate of convergence and improved performance [28]. Several adaptive filtering techniques have been designed for reference current generation, such as adaptive neural filtering for single-phase applications, adaptive robust predictive current control, adaptive notch filter, adaptive low pass, and bandpass filters. [29, 30].

A literature review of different controllers [18–30] depicts that the performance of the voltage source converter (VSC)controlled using traditional methods is unsatisfactory in some aspects, viz. slow dynamics, poor convergence, increased complexity, and additional requirement of phase-locked loop circuits. Hence, power engineers focus on devising new and improved controllers for PQ improvement in power distribution systems. Algorithms based on the brain learning process are still in their infancy, yet they hold a new relevance in power systems and problems faced in distribution systems. New controller design for the custom power devices is the focus of research in this paper, and the authors have tried to emulate the mammalian brain for satisfactory

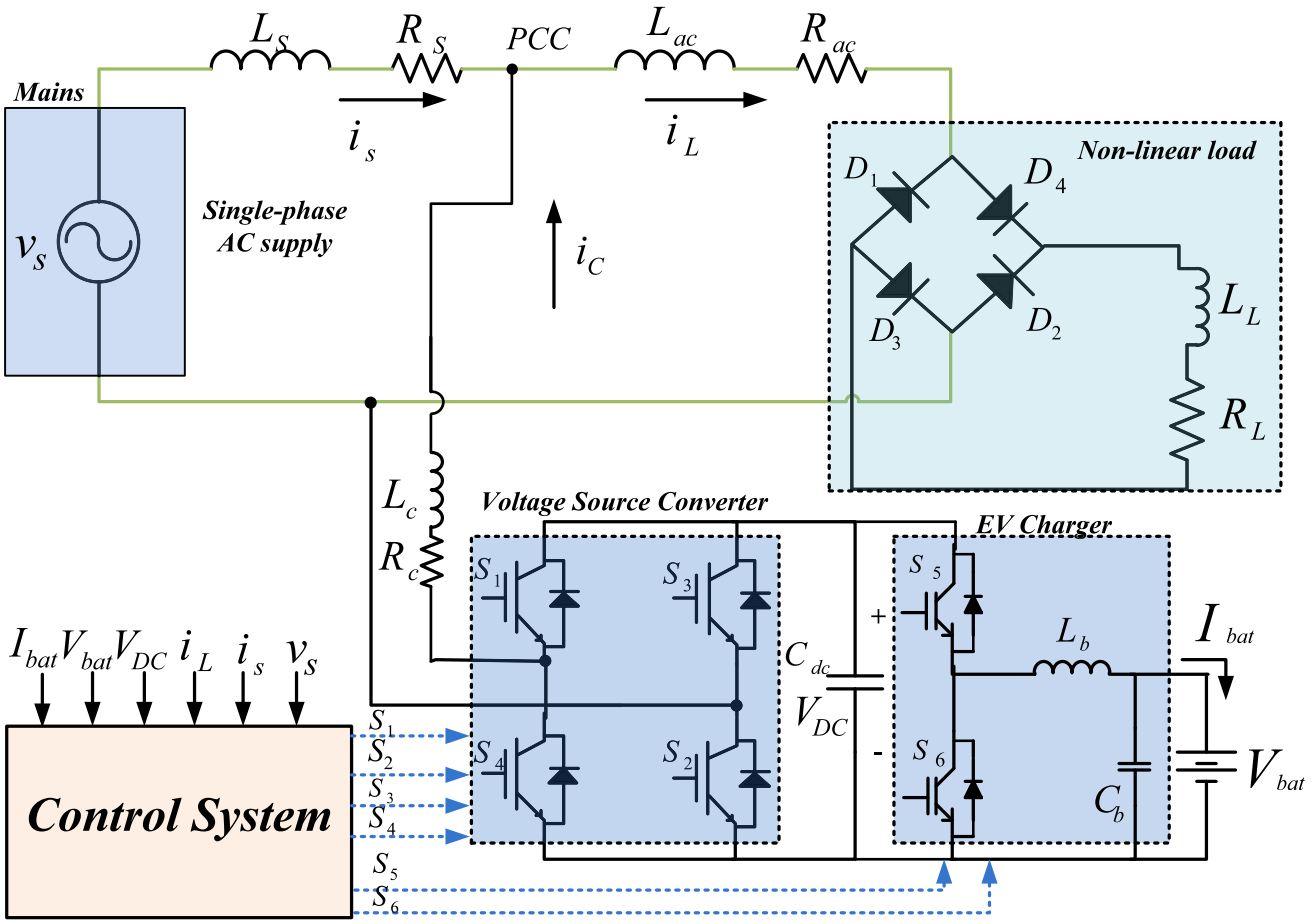



**Fig. 1** Block diagram of the proposed system

operation of EV charging in the designed systems. The concept of designing a new and innovative controller based on the brain's working is not addressed in the literature. Hence, this paper presents a new control technique modeled on the workings of the human brain and its application to a power distribution system. The developed control approach is employed for real-time application to control VSC in a single-phase system and also manages the smooth operation of EV chargers. The paper's key contributions are summarized as follows:

- The developed BIL technique mimics the working of human brain behavior applied for the control of VSC and has not been developed yet.
- A comprehensive analysis of the controller is performed and investigations are carried out for various gain values.
- The developed algorithm involves no transformation equations or PLL block and is fit for real-time applications.
- The suggested controller is self-adaptive and operates well under dynamic load and grid variations as well as during EV charging/discharging operations.
- Suitable comparison with conventional control techniques, viz. SRF theory and adaptive LMS algorithm, are presented.

The developed BIL control possesses attributes such as rapid speed of the teaching–learning process with high efficiency. Moreover, the human brain can reorganize and reorient while learning new concepts. The brain's emotional center amgydala gets stimulated and processes new information effectively for improved performance [31]. Both simulation and hardware results are presented for EV charging and discharging operations with PQ improvement in the subsequent Sections.

## 2 Single-Phase Grid-Connected Voltage Source Converter (VSC) with EV Charger

Figure 1 depicts the standard H-bridge configuration of a single-phase voltage source converter. It has a self-contained DC bus, four IGBT switches, and a single-phase full bridge

VSC. At the point of common coupling (PCC), single-phase source, a nonlinear load, and single-phase full bridge converter are connected via an interface inductor $L_C$. The resistive-inductive (RL) load at the DC terminals of a full-wave bridge rectifier is used to simulate the nonlinear load demand. A single-phase AC power supply feeds the system. The system uses LEM voltage and current sensors to detect voltage and current signals. These signals are fed to the real-time system (RTS) controller, which outputs four gating pulses for VSC and two gating pulses for the EV charger, respectively. The objective of the proposed system is to focus on the enhanced performance of EV charging/discharging operation and improved system PQ.

## 3 Mathematical Modeling

This section discusses the mathematical modeling of the VSC and the proportional integral (PI) controller modeling in detail.

### 3.1 Mathematical modeling of VSC

In this section, the mathematical modeling of single-phase VSC is carried out. The proposed VSC model shown in Fig. 2a incorporates a PI regulator block ($G_{PI}$), current control block and voltage source converter represented as $K_{con}$(s). So, the VSC model is made up of two parts: the internal control loop for current and the exterior control loop for voltage. The voltage on the DC link ($V_{dc}$) of the VSC is regulated to reference DC voltage ($V_{dc}^*$) by the outer control loop. The actual grid current $I_s$ is regulated to reference grid current ($I_s^*$) and appropriate p ulses are developed using the hysteresis current controller (HCC).$K_{con}(s)$ represents the transfer function of VSC, which can be determined with the help of small signal analysis. The VSC-based system is assumed to be under steady-state conditions to determine the transfer function. It shows the connection between the converter's output side (the DC link) and the input side (the AC side) linked to the grid.

Figure 2b shows the configuration of the VSC model with grid-connected nonlinear load to design equations of interfacing inductor, capacitor, and transfer function calculation are discussed in detail.

In VSC configuration, the DC link output ($V_{dc}$) ratio to VSC model input current ($i_c$) constitutes the transfer function. The VSC's average power exchange rate between the input (AC side) and output (DC link side) is computed.

The average energy change rate in VSC across the interfacing inductor, capacitor, and DC link capacitor can be expressed as shown in Eq. 1.

$$P_{cap} = P_{conv} - P_{loss} - P_{ind} \tag{1}$$

Here $P_{cap}$ denotes the rate of energy stored in the capacitor, $P_{conv}$ is the power input to the converter, $P_{loss}$ is active power loss in VSC across resistance ($R_c$) and $P_{ind}$ is power absorbed in an interfacing inductor ($L_c$).

The DC-link capacitance power is

$$P_{cap} = \frac{d}{dt}\left(\frac{1}{2}C_{dc}V_{dc}^2\right) \tag{2}$$

The voltage source converter (VSC) power input ($P_{conv}$) is,

$$P_{conv} = v_s i_c \tag{3}$$

The resistive loss ($R_c$) power is

$$P_{loss} = i_c^2 R_c \tag{4}$$

The interfacing inductor ($L_c$) across power absorbed is

$$P_{ind} = \frac{d}{dt}(\frac{1}{2}L_c i_c^2) \tag{5}$$

Using Eq. 1 to Eq. 5

$$C_{dc}V_{dc}\frac{dV_{dc}}{dt} = \left[v_s i_c - L_c i_c \frac{di_c}{dt} - i_c^2 R_c\right] \tag{6}$$

The power equation is linearized by applying a minor disturbance to the input converter current ($\Delta i_c$). Consider a scenario in which a small voltage $V_{dc}$ perturbs the DC link voltage as well.

Substituting $i_c$ and $V_{dc}$ with appropriate terms including incremental changes $\Delta i_c$ and $\Delta V_{dc}$, respectively, in Eq.

$$i_c = i_{co} + \Delta i_c$$

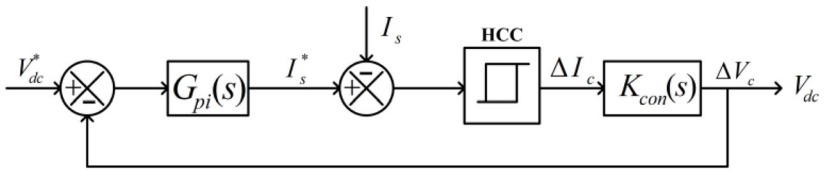
$$V_{dc} = V_{dco} + \Delta V_{dc}$$

where $i_{co}$ and $V_{dco}$ denotes steady-state operating points. By neglecting the higher-order terms Eq. (7) is obtained as

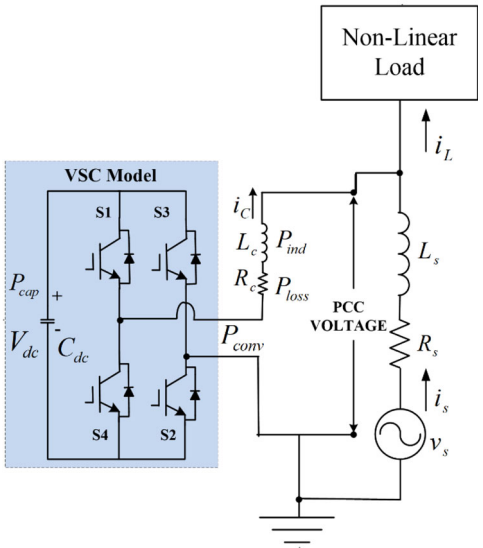
$$C_{dc}V_{dco}\frac{d\Delta V_{dc}}{dt} = \left[V_s i_{co} + V_s \Delta i_c - L_c i_{co}\frac{d\Delta i_c}{dt} - i_{co}^2 R_c - 2 i_{co}\Delta i_c R_c\right] \tag{7}$$

Under steady-state conditions, Eq. 7 can be written as

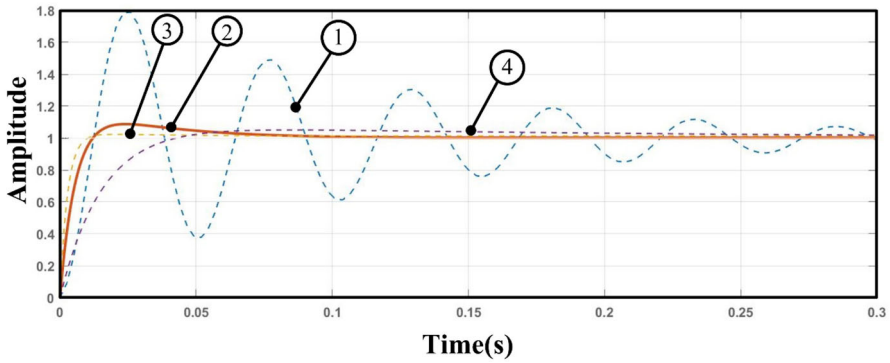
$$0 = [V_s i_{co} - i_{co}^2 R_c] \tag{8}$$



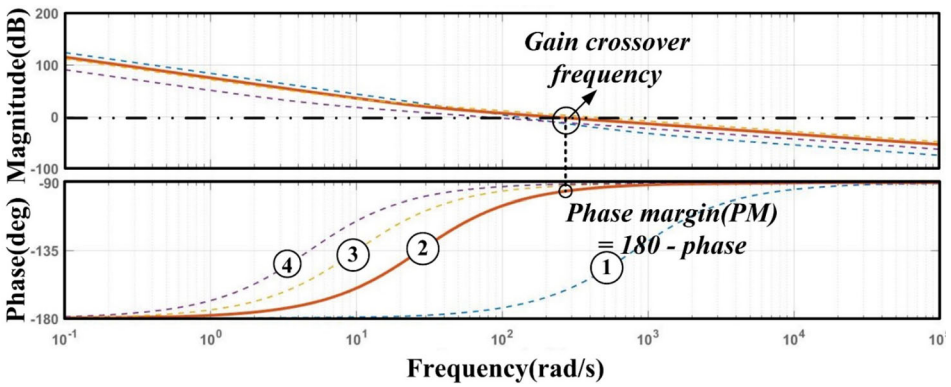

(a) Representation of control loops of the VSC model




(b) Configuration of VSC model with grid-connected nonlinear load




(c) Step response of the system corresponding to different gain parameters




(d) Bode plot of a voltage control loop

**Fig. 2** **a** Representation of control loops of the VSC model, **b** configuration of VSC model with grid-connected nonlinear load, **c** step response of the system corresponding to different gain parameters, **d** Bode plot of a voltage control loop

Now by subtracting Eqs. 7 and Eq. 8, we get

$$C_{dc}V_{dco}\frac{d\Delta V_{dc}}{dt} = \begin{bmatrix} V_s\Delta i_c - L_c i_{co}\frac{d\Delta i_c}{dt} \\ -2i_{co}\Delta i_c R_c \end{bmatrix} \tag{9}$$

Taking Laplace transform of Eq. 9 gives

$$C_{dc}V_{dco}s\Delta V_{dc} = \begin{bmatrix} V_s\Delta i_c - L_c i_{co}s\Delta i_c \\ -2i_{co}\Delta i_c R_c \end{bmatrix} \tag{10}$$

The converter transfer function $K_{con} = \frac{\Delta V_{dc}}{\Delta I_c}$ can be written as

$$K_{con} = \frac{\Delta V_{dc}}{\Delta I_c} = \frac{[V_s - L_c i_{co}s - 2i_{co}R_c]}{C_{dc}V_{dco}s} \tag{11}$$

In the voltage control loop shown in Fig. 2a, the HCC block sends the required gate pulses to the VSC. The output of the PI controller is sent to HCC and used to generate a suitable reference current.

Hence, necessary modifications have been made to better understand the mathematical expressions.

### 3.2 Modeling of PI Controller

The PI regulator, shown by the block $G_{pi}$ (s) in Fig. 2a, is utilized to control the DC link voltage at a reference value. The PI controller is used to regulate the DC link voltage to its reference value such that the steady-state error estimated in $V_{dc}$ reduces to zero. The control scheme is designed so that the grid compensates for the converter losses. The PI controller is depicted as $G_{pi}(s)$ block, and it is represented as

$$G_{pi}(s) = K_p + \frac{K_i}{S} \tag{12}$$

$K_p$ represents the proportional gain, whereas $K_i$ signifies the integral gain of the specified system. Utilizing Fig. 2a, the transfer function of the voltage control loop might be expressed as

$$T(s) = \frac{G_{pi}(s)K_{con}(s)}{1 + G_{pi}(s)k_{con}(s)} \tag{13}$$

Substituting Eq. (11) and (12) and the values of parameters mentioned in Appendix, T(s) expression comes out to be

$$T(s) = \frac{110K_i + 110K_p s}{0.3S^2 + 110K_p S + 110K_i} \tag{14}$$

Appropriate PI gains values must be selected to control the system properly. Accordingly, the system's step response is shown in Fig. 2c, and the Bode plot is shown in Fig. 2d. Table 1 shows four different combinations of $K_p$ and $K_i$ and the corresponding rise time, settling time, overshoot for step response, and phase margin of the system evaluated from the bode plot. From Table 1, $K_p$ and $K_i$ values are selected to be 0.55 and 15, respectively, for which the response of the system is better and stability is ensured. The plots highlight fast settling time and good convergence achieved with moderate overshoot and satisfactory PM of 82.366°.

## 4 Proposed Controller

In this section, the BIL controller design and the complete grid-connected EV charging system have been discussed in detail.

### 4.1 BIL Controller Design

Today, biology-inspired techniques are being developed for several power and control engineering applications. This paper discusses the design of a new BIL controller inspired by the brain's learning. It is based on amygdala-orbito-fronatal model proposed in [32, 33] and is briefly discussed below.

The amygdala and orbito-frontal cortex (OC) are two important parts of the limbic system and play an important role in processing emotional stimulus rapidly and accurately. Some important features of brain-inspired learning include a high rate of learning mechanisms and quick reaction, and it can be widely used for prediction [34] and control application [35]. The interaction between the amygdala and OC makes the BIL controller emulate the functioning of the human brain. Further, the BIL controller possesses advantages over other intelligent controllers, such as neuro-fuzzy, reinforcement learning, expert systems, and machine learning algorithms in terms of simpler structure and low computational requirements. Most intelligent controllers cannot be implemented for real-time control applications because of their complexity; however, a relatively simplified design of BIL works well. This controller inherits features such as adaptive nature, robust performance, and learning under uncertainty. Moreover, it can be used for linear and nonlinear systems [36–38].

The human brain's limbic system includes the sensory cortex, thalamus, amygdala, OC, and other structures. The limbic system divides the brain into separate regions that execute certain tasks [39]. The amygdala is crucial for processing emotional input, and the OC is crucial for learning and responding. A block diagram of the limbic system's components—the amygdala, thalamus, and sensory cortex and their interactions is shown in Fig. 3a. The amygdala and OC are essential for emotional learning. The sensory cortex and thalamus provide the amygdala with emotional information and an external reward signal or emotional stimuli (ES). It

**Table 1** Choice of PI gain parameters

| S.No | $K_p$ | $K_i$ | PM(Phase margin(°) | Ts (Settling time(s) | Tr (Rise time(s) | % Overshoot |
|---|---|---|---|---|---|---|
| 1 | 0.05 | 40 | 8.6572 | 0.4186 | 0.0088 | 78.88 |
| 2 | 0.55 | 15 | 82.366 | 0.0762 | 0.0083 | 8.768 |
| 3 | 1 | 10 | 88.438 | 0.1855 | 0.0056 | 2.338 |
| 4 | 0.19 | 0.9 | 86.119 | 0.2853 | 0.268 | 5.1195 |

converses with the OC, which responds to emotional stimuli based on the signal from the reward system. The OC also receives sensory information from the sensory cortex and assesses the amygdala's reaction to avoid incorrect learning connections. The interactions between these components influence the learning processes of the human brain.

A simpler approach to developing the BIL model considers signals such as sensory input ($SI$), the output of amygdala ($A_i$), the output of OC ($OC_i$) and the reward signal ($Rew$) which are used for updating the weights of amygdale ($A_i$) and OC ($OC_i$). The input to the system is SI, which represents the sensor output. The block diagram representation of BIL control showing the complete signal flow is shown in Fig. 3b. The mathematical expressions for BIL-based intelligent control are depicted by the following equations.

$$A = G_A \times SI \tag{15}$$

$$OC = G_{OC} \times SI \tag{16}$$

$$\frac{dG_A}{dt} = a \times SI \times (ES - A, 0) \tag{17}$$

$$\frac{dG_{OC}}{dt} = b \times SI \times (A - OC - ES) \tag{18}$$

$$O = A - OC \tag{19}$$

In Eqs. (15–19) $G_A$ denote amygdala's gain and $G_{oc}$ represents OC gain, $a$, and $b$ are learning rate constants and O depicts the output of the controller. Using the block diagram developed in Fig. 3b and the equations depicted, the complete transfer function of the given system is evaluated.

The block diagram of the system is shown in Fig. 3b. For finding out the transfer function of the given system, the system is assumed to be under steady-state condition. The controller for PQ mitigation is developed to obtain the fundamental component of load current (O). The SI represents the error signal.

between the sensed load current and the final output and O represents the output of the BIL controller. ES signal is considered as the output of the PID controller as shown in Eq. 21. Learning is affected by two factors *'a* and *'b'* evaluated using Eqs. 22 and 23.

$$SI = e \tag{20}$$

$$ES = k_p e + k_i \int e + k_d \frac{de}{dt} = [k_p + \frac{k_i}{s} + k_d s]e \tag{21}$$

$$a = K_a \times |SI| \tag{22}$$

$$b = K_b \times |SI| \tag{23}$$

$$T(s) = \sum_{K=1}^{n} \frac{P_k \Delta_k}{\Delta} \tag{24}$$

The overall transfer function T(s) is evaluated using Mason's gain formula.

where $P_k$ denotes the k$^{th}$ forward path.

$\Delta_k$= 1-[non-touching loops to kth forward path].

$\Delta$= 1- [Sum of the loop gains of all individual loops] + [Sum of gain products of all possible of two non-touching loops] + …….

Using the developed model as shown in Fig. 3b

$$P_1 = \left(k_p + k_d s + \frac{k_i}{s}\right)\left(\frac{ae^2}{s + ae^2}\right)\left(\frac{-be^2}{s - be^2}\right) \tag{25}$$

$$P_2 = \left(k_p + k_d s + \frac{k_i}{s}\right)\left(\frac{ae^2}{s + ae^2}\right) \tag{26}$$

$$P_3 = \left(k_p + k_d s + \frac{k_i}{s}\right)\left(\frac{be^2}{s}\right) \tag{27}$$

$$L_1 = \left(k_p + k_d s + \frac{k_i}{s}\right)\left(\frac{ae^2}{s + ae^2}\right)\left(\frac{be^2}{s - be^2}\right) \tag{28}$$

$$L_2 = \left(k_p + k_d s + \frac{k_i}{s}\right)\left(\frac{-be^2}{s - be^2}\right) \tag{29}$$

$$L_3 = \left(k_p + k_d s + \frac{k_i}{s}\right)\left(\frac{-ae^2}{s + ae^2}\right) \tag{30}$$

$$T(s) = \frac{P_1\Delta_1 + P_2\Delta_2 + P_3\Delta_3}{1 - [L_1 + L_2 + L_3]} \tag{31}$$

**Fig. 3** **a** Block diagram showing frame work of limbic system [31], **b** detailed block diagram of BIL controller, **c** Bode plot of proposed system for different gains, **d** step response of the proposed system for different sets of gains, **e** grid-connected EV charging system control

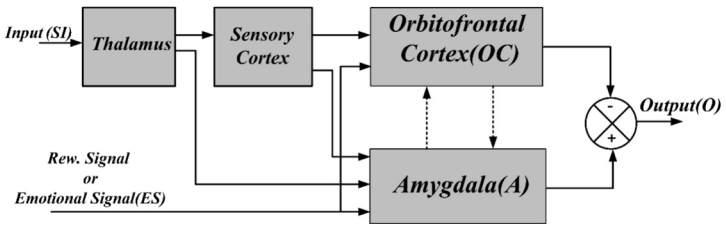



Fig. 3(a) Block diagram showing frame work of limbic system [31]

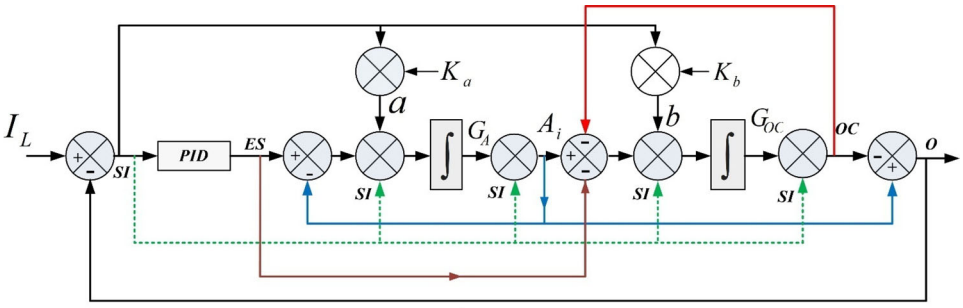



Fig.3(b) Detailed Block diagram of BIL controller

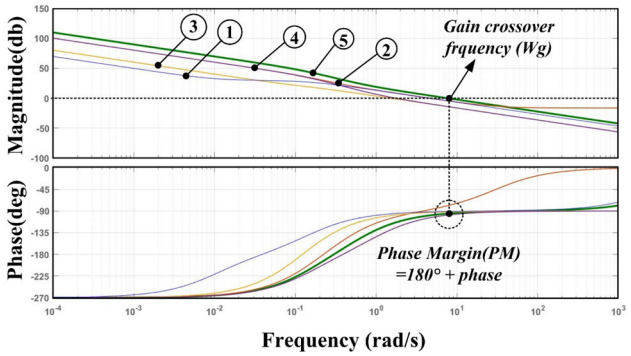



Fig.3(c) Bode plot of proposed system for different gains

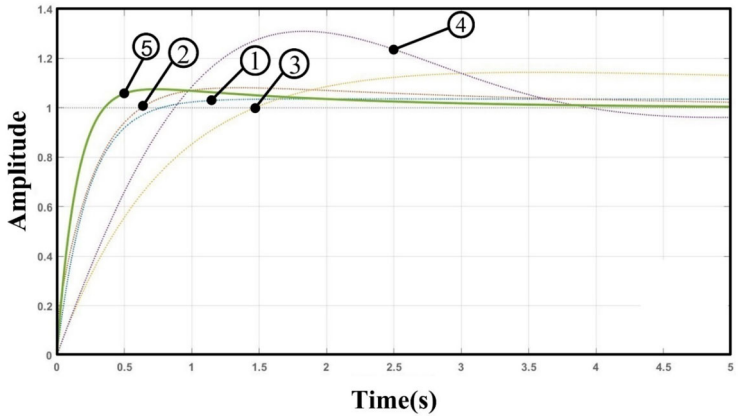
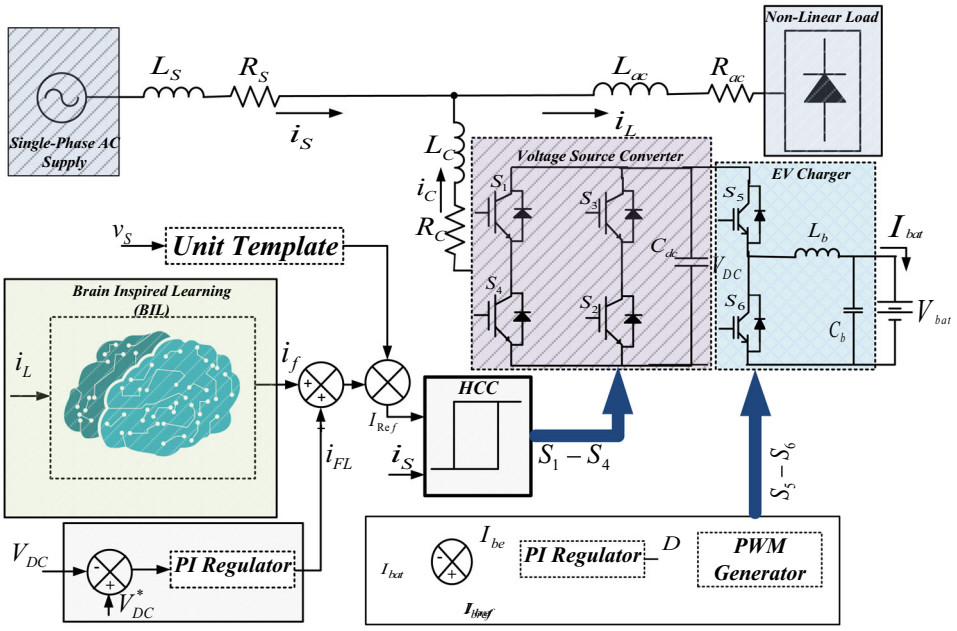



Fig.3(d) Step response of the proposed system for different sets of gains

**Fig. 3** continued




(e) Grid connected EV charging system control

system shows a stable system, and better response in terms of fast settling time, moderate rise time, and low overshoot values.

$$T(s) = \left[ \frac{(k_{p+}k_d s + k_i/s)\left[\left(\frac{ae^2}{s+ae^2}\right)\left(\frac{-be^2}{s-be^2}\right) + \left(\frac{ae^2}{s+ae^2}\right) + \left(\frac{be^2}{s-be^2}\right)\right]}{1 - (k_{p+}k_d s + k_i/s)\left[\left(\frac{ae^2}{s+ae^2}\right)\left(\frac{be^2}{s-be^2}\right) + \left(\frac{-ae^2}{s+ae^2}\right) + \left(\frac{-be^2}{s-be^2}\right)\right]} \right] \tag{32}$$

Substituting the parameter values and simplifying leads to

$$T(s) = \left[ \frac{(k_p + k_d s + \frac{k_i}{s})(\frac{s(a+b)-ab}{(s-a)(s-b)})}{(s+a) + (k_p + k_d s + \frac{k_i}{s})(\frac{s(a+b)-ab}{(s-a)(s-b)})} \right] \tag{33}$$

Now considering the value of $a$ to be much smaller as compared to $b$, the simplified equation can be written as

$$T(s) = \left[ \frac{(k_p + k_d s + \frac{k_i}{s}) \times \frac{b}{s-b}}{s + (k_p + k_d s + \frac{k_i}{s}) \times \frac{b}{s-b}} \right] \tag{34}$$

$$T(s) = \left[ \frac{(k_d s^2 + k_p s + k_i) \times b}{(1 + bk_d)s^2 + (bk_p - b)s + bk_i)} \right] \tag{35}$$

For optimum performance of the system, the bode plot and step response plots of the system are shown in Fig. 3c, d. Table 2 quantifies the responses in terms of PM, $T_s$, $T_r$, and overshoot for a set of selected gains ($K_p$,$K_i$, $K_d$). The Bode plot is shown in Fig. 3c and Table 2 helps to analyze the stability of the system in terms of phase margin. The step response of the system is shown in Fig. 3d. For this proposed system gain set, the last entry has been selected because for this set of values ($K_p$=50, $K_d$=0.01 and $K_i$=30) the proposed

## 4.2 Grid-Connected EV Charging System Control

Figure 3e shows the complete circuit diagram of the grid-connected EV charging system control. The load current ($i_L$) is fed through the BIL controller to extract fundamental current ($i_f$). A PI controller is designed and well-tuned to regulate the overall DC-link voltage to the reference value; further, the output of this controller provides an estimate of the current loss component. Thereafter, both currents fundamental ($i_f$) and loss component ($i_{FL}$) are added together, and this resultant current component is multiplied by the unit template to provide the sinusoidal reference current $i_{Ref}$. Using the grid voltage ($v_s$) and its quadrature phase shifted component, which are both input into the grid voltage ($v_s$) and processed via the peak estimation block, the unit template (u) is formed. This perfectly sinusoidal $i_{Ref}$ is processed by the HCC and compared with the actual supply current ($i_s$) to generate four gating pulses ($S_1$-$S_4$) for IGBT switching operations. These pulses are used to control the VSC. This ensures a perfectly

**Table 2** System performance with gain variation

| S.No | $K_p$ | $K_d$ | $K_i$ | PM (Phase margin(°)) | Ts (Settling time(s)) | Tr (Rise time(s)) | % Overshoot |
|---|---|---|---|---|---|---|---|
| 1 | 30 | 0.01 | 0.3 | 88.0473 | NaN | 0.4438 | 3.6134 |
| 2 | 30 | 1 | 10 | 92.467 | 5.6822 | 0.4311 | 8.1288 |
| 3 | 10 | 0 | 1 | 80.44 | 20.5477 | 1.0513 | 14.358 |
| 4 | 10 | 0 | 10 | 54.961 | 5.973 | 0.6788 | 31.005 |
| 5 | 50 | 0.01 | 30 | 84.387 | 2.860 | 0.0429 | 1.1014 |

sinusoidal grid current ($i_s$) while compensating for the reactive power demand of the load. Now, the DC link capacitor of the VSC is connected to the EV charger (bidirectional buck-boost converter with battery). The battery current ($I_{bat}$) and its reference current ($I_{bref}$) are compared, and its output is passed through a second PI controller. This PI output is fed to the pwm block to produce two pulses ($S_5$ -$S_6$) to control and ensure the smooth operation of the EV system.

## 5 Experimental Results and Discussion

This paper presents the waveforms of a charging and discharging operation for a single-phase VSC with improved PQ problems using a BIL-based intelligent control method. The designed grid-connected EV charger junction aims to facilitate both EV charging/discharging operations and supply reactive power to reduce harmonics in the supply current. The system performance is analyzed with experimental results obtained from the prototype built in the laboratory and depicted in Fig. 4a.

A single-phase 40 V, 50 Hz AC power supply (AE make variable autotransformer) feeds the nonlinear RL load system. The system parameters having LEM voltage (LV-25P) and current sensors (LA-25P) are used in the experimental setup to detect voltage and current signals. The real-time system (RTS) controller is used to implement the designed controller for VSC and the EV charger controller. Thus, the RTS controller output provides the required four gating pulses for VSC and two pulses for the EV charger, respectively. The power quality analyzer captures the steady-state charging/discharging results (HIOKI PQ3100). The four-channel Digital Signal Oscilloscope (DSO-X 2014A) is utilized to collect the dynamic waveform of the charging/discharging operation. The VSC DC link has a fixed reference voltage of 80 V. In this section, the practical results under steady state and dynamics are discussed in detail.

### 5.1 Steady-State Performance

This section covers steady-state performance analysis during EV charging/discharging conditions.

#### 5.1.1 Charging Operation

Figure 4b–d shows the proposed system's steady-state performance during EV charging operation. Figure 4b(a-c) depicts waveforms of source voltage ($v_s$) with source current ($i_s$), load current ($i_L$), and compensator current ($i_c$). Figure 4b–d displays the suggested system's steady-state performance during EV charging operation. Reiterating EV charging activities, Fig. 4b(a) indicates the source voltage ($v_s$) is in phase with the source's current ($i_s$).

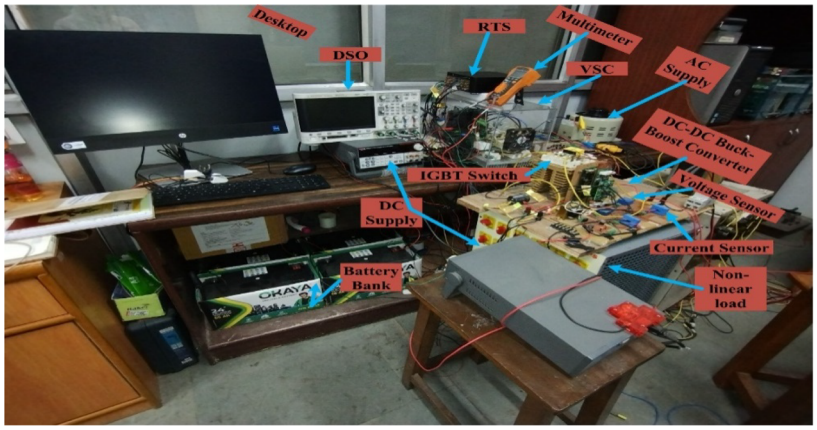
Figure 4c(a–c) provides the source, load, and compensator's power analysis. To the load, the grid provides active 147W and reactive 21var power. The load power requirement is 24var and 50W. During the charging procedure, the compensator injects the necessary reactive power to raise the power factor from 0.94 to 0.99.

The total harmonic distortion (THD) response of different variables is shown in Fig. 4d(a–c). The THD of the '$v_s$' is 1.96%, and the '$i_s$' is 4.56%, as per Fig. 4d(a), which follows the IEEE-519 limitation. The $i_L$ THD is 15.49%, as recorded in Fig. 4d(b). The VSC acting as a compensator shows a THD of 6.04%, as shown in Fig. 4d(c). The figures illustrate that the single-phase VSC performs harmonic reduction and effectively lowers harmonic content in the supply current. Additionally, the power factor is nearly one, and the supply voltage and current operate in phase during charging operations.

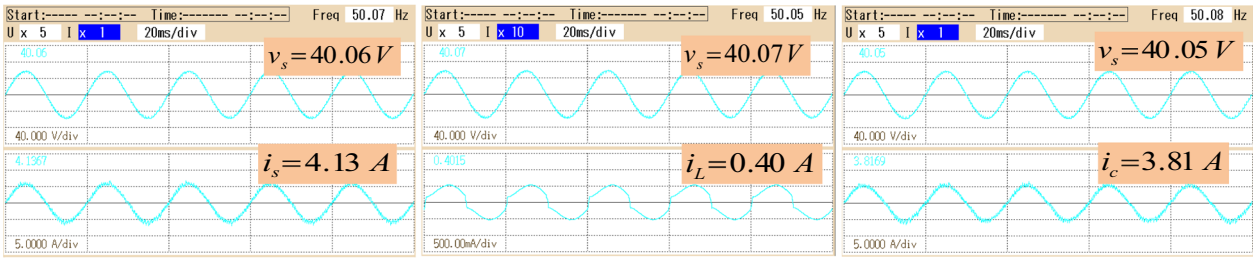
#### 5.1.2 Discharging Operation

The steady-state performance during EV discharging operation mode is shown in Fig. 5a-b. Figure 5a(a–c) shows the source voltage waveforms with plots of source current, load current, and compensator current, respectively. It is observed that the source voltage and current are out of phase. Further, it is proof of the system operating in discharging mode. The compensator injects current to load and surplus current into the grid.

Figure 5b(a–c) shows the power analysis of the grid, load, and compensator during the discharging operation. The EV charger, along with the battery connected at the DC link of the compensator end, supplies 117.5W, and 17 var, as seen

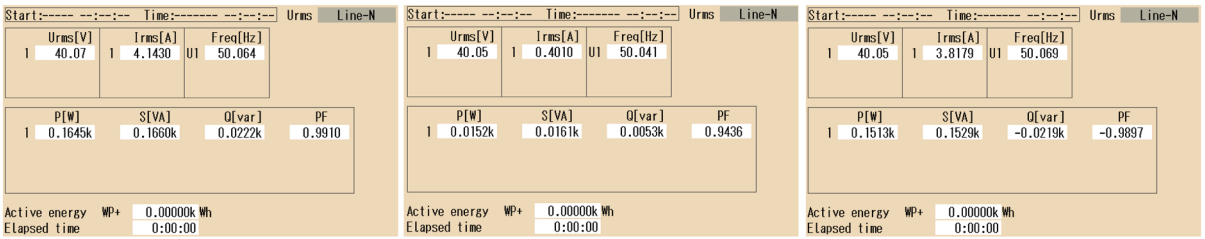



(a) Experimental setup of the complete system



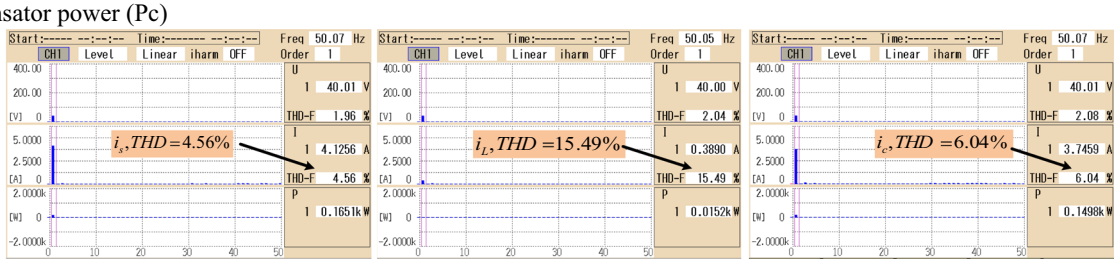

(a) (b) (c)

(b) Steady-state waveforms showing (a) $v_S$ and $i_S$ (b) $v_S$ and $I_L$ (c) $v_S$ and $i_c$ during charging operation.




(a) (b) (c)

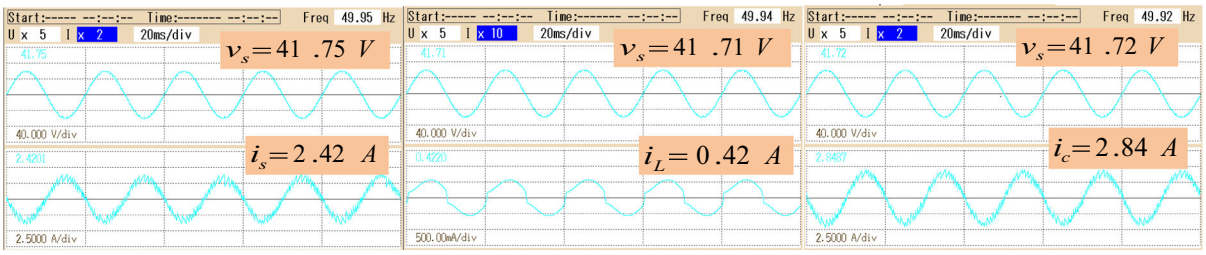
(c) Steady-state performance of power during charging operation (a) Supply power (Ps) (b) Load power ($P_L$)and (c) Compensator power (Pc)




(a) (b) (c)

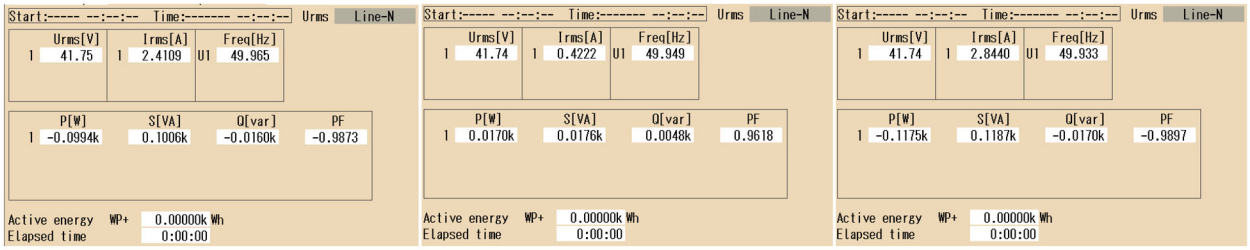
(d) Steady-state performance showing (a). THD of $i_S$(4.56%) (b).$i_L$(15.49%) and (c) THD of $i_c$(6.04%).

**Fig. 4** **a** Experimental setup of the complete system, **b** Steady-state waveforms showing **a** $v_S$ and $i_S$ **b** $v_S$ and $I_L$ **c** $v_S$ and $i_c$ during charging operation, **c** Steady-state performance of power during charging operation **a** Supply power (Ps) **b** Load power ($P_L$)and **c** Compensator power (Pc), **d** Steady-state performance showing **a**. THD of $i_S$(4.56%) **b**.$i_L$(15.49%) and **c** THD of $i_c$(6.04%)



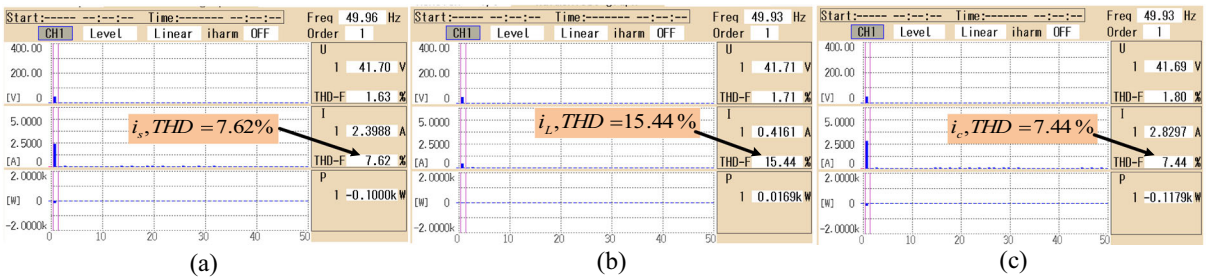

(a) (b) (c)

(a Steady state waveforms showing (a) $v_S$ and $i_S$ (b) $v_S$ and $I_L$ (c) $v_S$ and $i_c$ during discharging operation.




(a) (b) (c)

(b) Steady state performance of power during discharging operation (a) supply power (Ps) (b) load power ($P_L$) and (c) compensator power (Pc)




(a) (b) (c)

(c) Steady state performance showing during discharging operation (a). THD of $i_S$ (7.62%) (b). $i_L$(15.44%) and (c) THD of $i_c$(7.44%)

**Fig. 5** **a** Steady-state waveforms showing **a** $v_S$ and $i_S$ **b** $v_S$ and $I_L$ **c** $v_S$ and $i_c$ during discharging operation, **b** steady-state performance of power during discharging operation **a** supply power (Ps) **b** load power ($P_L$) and **c** compensator power (Pc), **c** steady-state performance showing during discharging operation **a**. THD of $i_S$ (7.62%) **b**. $i_L$(15.44%) and **c** THD of $i_c$(7.44%)

in Fig. 5b(c), to load, and surplus power is injected into the grid. The load power demand is 17W and 4.8var during discharging, as seen in Fig. 5b(b). The extra power is fed back to the grid, which amounts to 99.4W and 16var, as shown in Fig. 5b(a).

The THD of ($i_s$), ($i_L$), and ($i_c$) with ($v_s$) is shown in Fig. 5c(a–c). The THD of ($v_s$) is 1.63%, and the ($i_s$) THD is 7.62%. The THD of ($i_L$) is 15.44%, as recorded in Fig. 5c(b). The EV charger discharges through VSC, and VSC acts as a compensator with THD 7.44%, as seen in Fig. 5c(c).

## 5.2 Dynamic Performance

The dynamic performance of the proposed brain learning system is analyzed during charging/discharging operation under load disturbances. In this section, the dynamic waveforms under charging and discharging modes of operation are separately shown in Figs. 6 and 7.

### 5.2.1 Charging Operation

Figure 6 shows the dynamic performance of the proposed BIL controller under load increment during charging operation. Figure 6a shows the dynamic performances of source voltage ($v_s$), source current ($i_S$), load current ($i_L$) and DC

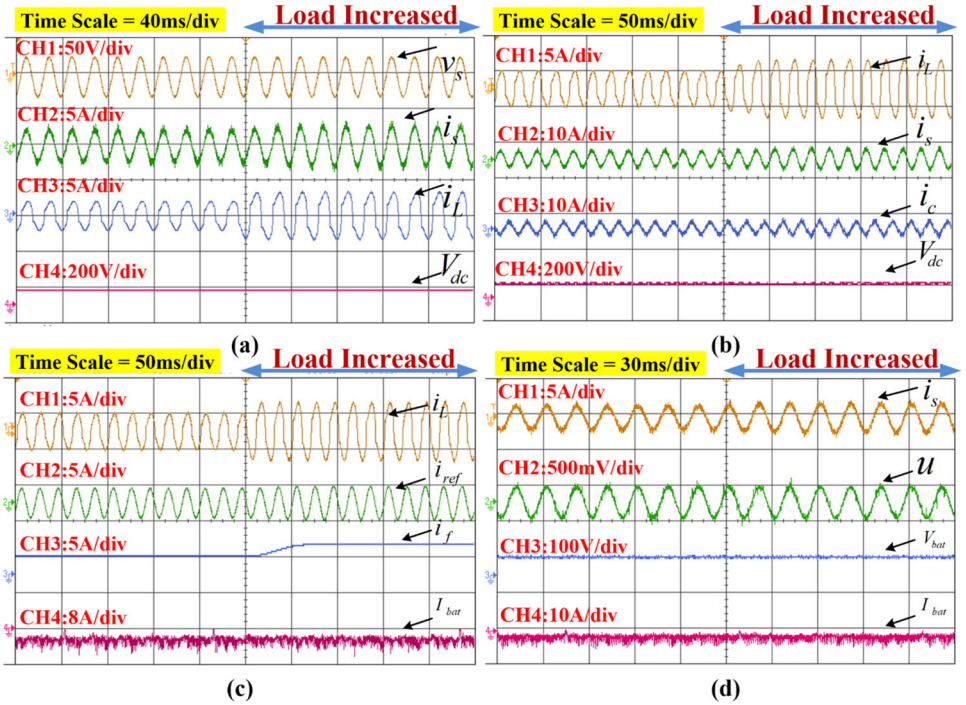



**Fig. 6** Experimental waveform of **a** $(v_S)$,$(i_S)$,$(i_L)$, $(i_c)$, $(V_{dc})$ **b.**$(i_L)$,$(i_S)$,$(i_c)$, $(V_{dc})$ **c.**$(i_L)$,$(i_{ref})$,$(i_f)$, $(I_{bat})$ and **d** $(i_S)$, (u), $(V_{bat})$, and $(I_{bat})$ during charging operation

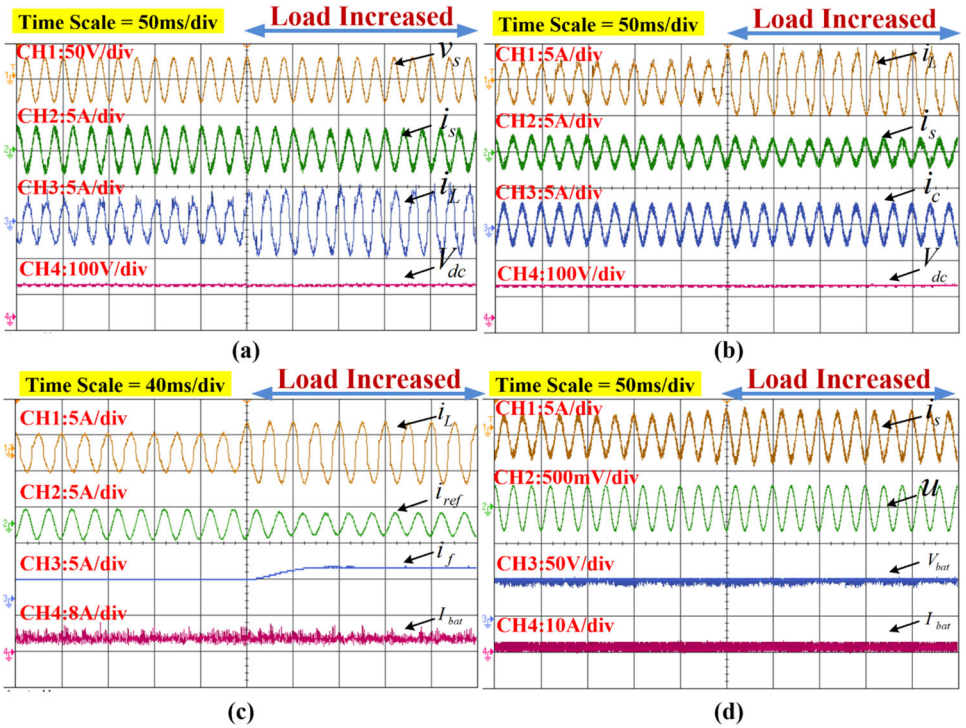



**Fig. 7** Experimental waveform of **a** $(v_S)$,$(i_S)$,$(i_L)$, $(i_c)$, $(V_{dc})$ **b** $(i_L)$,$(i_S)$,$(i_c)$, $(V_{dc})$ **c.**$(i_L)$,$(i_{ref})$,$(i_f)$, $(I_{bat})$ and **d** $(i_S)$, (u), $(V_{bat})$, and $(I_{bat})$ during discharging operation

link voltage ($V_{DC}$). In this figure, the change in source current due to load change can be observed. Further, in phase relationship between source voltage and source current is observed for charging operation. Figure 6b–c shows dynamic performances of $(i_L)$, $(i_S)$, $(i_c)$, $(V_{dc})$, reference current $(i_{ref})$, fundamental current $(i_f)$ and battery current $(I_{bat})$. The change in load definitely increases the source current, reference current, and fundamental current, but the DC link voltage and battery current remain nearly constant, which indicates a stable operation in line with the experimental results. Figure 6d shows the waveforms of $(i_s)$, unit template (u), battery voltage ($V_{bat}$), and battery current ($I_{bat}$). The source current due to load change follows the unit template (u). The battery voltage ($V_{bat}$) continuously charges

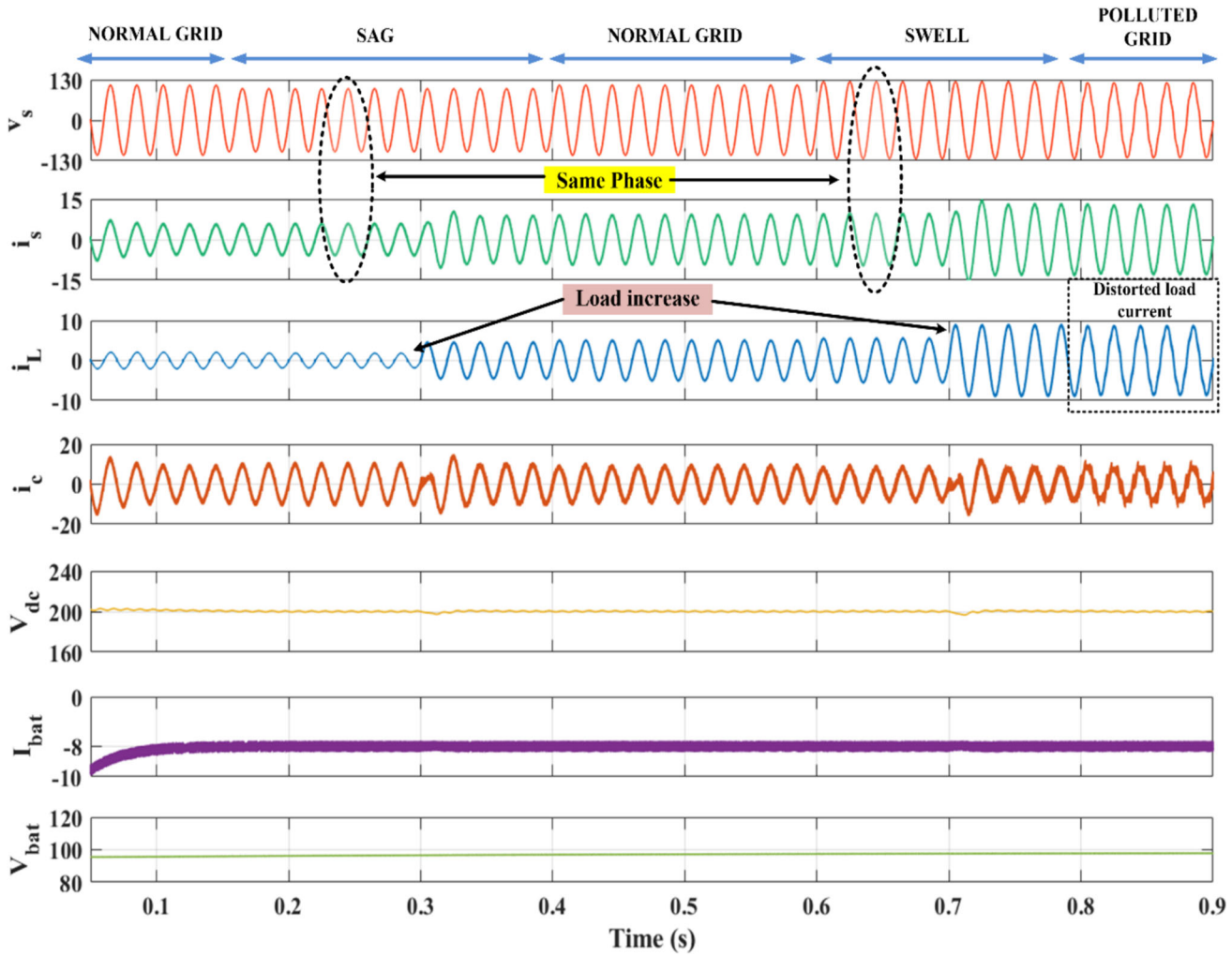



**Fig. 8** Waveforms of ($v_s$), ($i_s$), ($i_L$), ($i_c$), ($V_{dc}$), ($V_{bat}$), and ($I_{bat}$) during G2V at varying grid conditions

from 48V DC. The battery reference current ($I_{bat}$) is set to -3A during charging operation. Figure 6a–d also shows that the grid current is sinusoidal and in phase with the supply voltage. Moreover, the DC link voltage of the VSC is maintained constant even during load increments. The IEEE-519 standard is also valid concerning the harmonic content in grid current.

#### 5.2.2 Discharging Operation

Figure 7a–d shows the dynamic performance of($v_s$), ($i_S$), ($i_L$), ($i_c$), ($V_{dc}$), ($i_{ref}$), ($i_f$), ($V_{bat}$), and ($I_{bat}$) under load increment during the discharging operation mode of the EV. The presented waveforms in Fig. 7 show major differences in terms of the opposite waveform of source voltage and current. During discharging operation mode, the battery current is set to 3A reference, and the actual ($I_{bat}$) follows the reference value as shown in Fig. 7c–d. Further, the DC link voltage settles to a reference value of 80V, and the battery voltage is set to 48V. As dynamic load increases on the system, the reference current increases, as shown by the increased value of if in Fig. 7c. A clear phase opposition in supply voltage and supply current signals is easily noticed during discharging operating conditions, as shown in Fig. 7a.

## 6 Performance Under Varying Grid Conditions

The simulation results under varying grid conditions are shown in Figs. 8 and 9 for charging and discharging operations, respectively.

### 6.1 Charging Operation

Figure 8 shows the waveforms of $v_s$, $i_s$, $i_L$, $i_c$, $V_{dc}$, $I_{bat}$, $V_{bat}$ during charging mode under varying grid conditions. As shown in this figure, voltage sag is simulated from t =

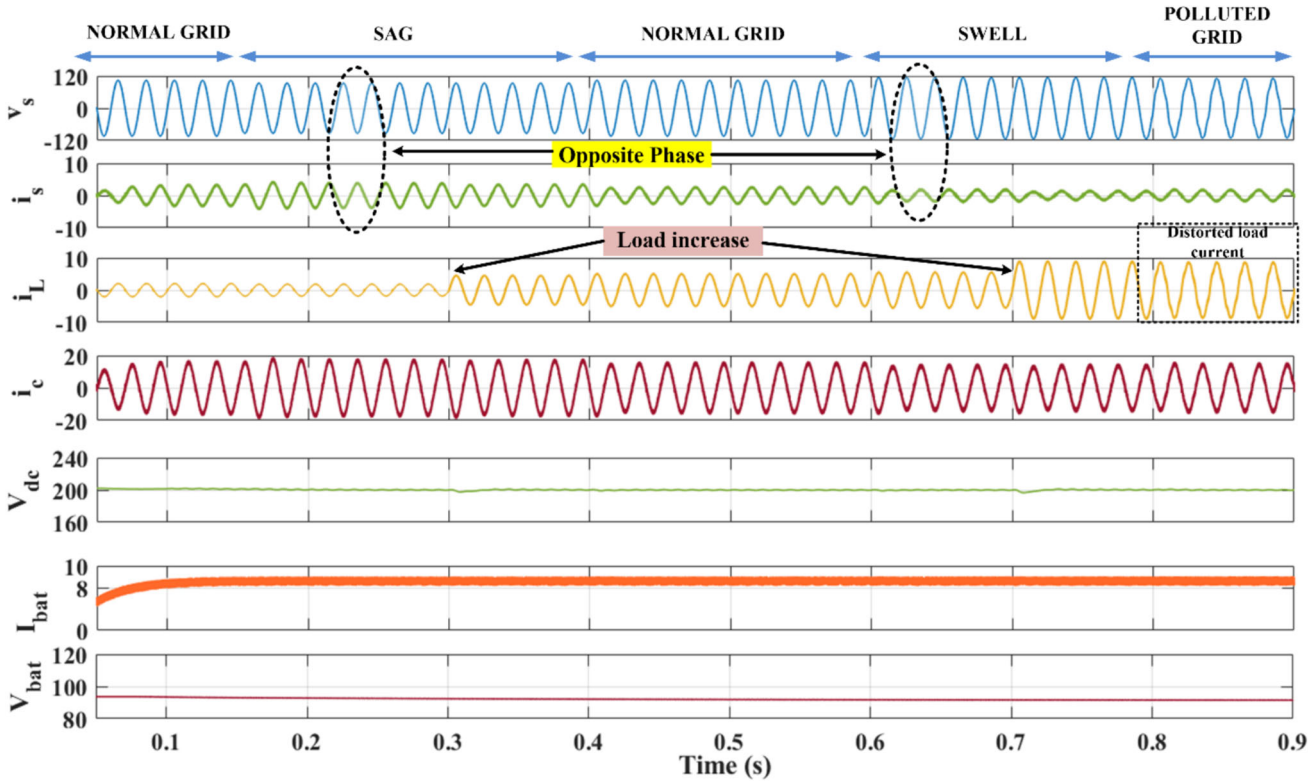



**Fig. 9** Waveforms of ($v_s$), ($i_s$), ($i_L$), ($i_c$), ($V_{dc}$), ($V_{bat}$), and ($I_{bat}$) during V2G at varying grid conditions

0.15 s to t = 0.4 s; swell is simulated from t = 0.6 s to t = 0.8 s, and polluted grid with harmonics is simulated from t = 0.8 s onwards. Further, the load current is increased at t = 0.3 s and t = 0.7 s; the source current also increases correspondingly. The BIL controller works well and computes the fundamental load current to generate reference currents for the VSC. Results show a small dip in DC link voltage during load change, and $V_{dc}$ quickly settles to 200 V, which shows the robustness of the proposed BIL controller.

Distorted grid conditions are also tested in Fig. 8. The 7th harmonics are also injected in the grid voltage at t = 0.8 s, and it is observed that the proposed controller easily removes the harmonics present in the grid current, and the supply current waveforms are found to be sinusoidal. The source voltage and source current are observed to be in the same phase, proving that the system is in charging mode. The battery current is -8A, again denoting the charging operation of the EV.

### 6.2 Discharging Operation

Figure 9 shows the waveforms of $v_s$, $i_s$, $i_L$, $i_c$, $V_{dc}$, $I_{bat}$, and $V_{bat}$ during discharging mode under varying grid conditions. As in Fig. 8 before, voltage sag is simulated from t = 0.15 s to t = 0.4 s, swell from t = 0.6 s to t = 0.8 s, and distorted grid from t = 0.8 s onwards. Further, the effect of load current increase from t = 0.3 s and t = 0.7 s is observed on the system performance. It is analyzed that the '$i_s$' has now decreased. It is due to the compensator injecting current into the load and surplus current being fed into the grid. The EV battery is discharging, feeding the load and/or grid.

The effect of grid variations in the form of voltage sag/swell and harmonics is also shown in Fig. 9. The 7th harmonics are also injected in the grid voltage at t = 0.8 s, and it is observed that the proposed controller easily removes the harmonics present in the grid current. The proposed BIL controller is equipped with a generalized integrated block that can remove the effects of distortions and grid abnormalities. This ensures satisfactory performance in both operating modes. Figure 9 shows the waveform of the supply current is sinusoidal. In Fig. 9, the source voltage and source current are observed to be in the opposite phase, which proves that the system is in discharging mode condition.

## 7 Performance comparison of controllers

Table 3 compares the proposed learning control technique with conventional techniques such as SRF and LMS algorithms. It is observed that the proposed BIL control technique shows better performance in terms of a high learning rate, quick reaction to disturbances and is self-adaptive in nature. The BIL output prediction is fast and correct fundamental

**Table 3** Comparison of performance of BIL with SRF and LMS algorithms

| S.No | Parameters | SRF technique | LMS technique | Proposed BIL |
|---|---|---|---|---|
| 1 | PLL -requirement | Required | No | No |
| 2 | Fundamental current extraction | Oscillations are present | Oscillation is absent | Zero oscillations |
| 3 | Complexity of control | High | Moderate | Moderate |
| 4 | Weight Convergence | Moderate(~ 3 cycles) | Slow (> 3 cycles) | Fast (< 2 cycles) |
| 5 | % THD of $i_s$ | 5% | 4.84% | 4.56% |
| 6 | % THD of $i_L$ | 15.33% | 15.43% | 15.44% |
| 7 | % THD of $i_c$ | 6.83% | 6.48% | 6.04% |
| 8 | Tracking | Moderate | good | Very good |
| 9 | Technique | Non-adaptive | adaptive | Self-adaptive |
| 10 | DC link voltage variation | 4-5 V | 3 V | 2 V |

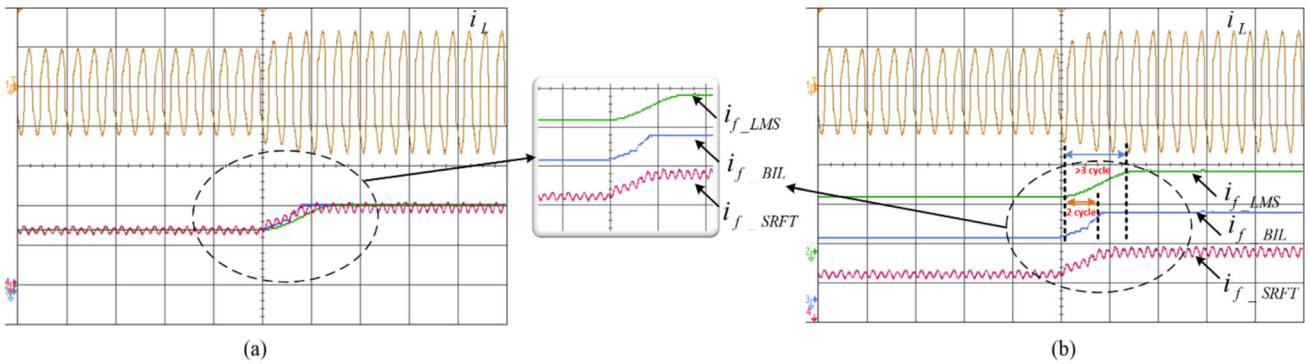



**Fig. 10** Experimental results showing waveform of load current ($i_L$) along with fundamental current extracted through BIL ($i_{f_BIL}$),SRF ($i_{f_SRF}$) and LMS algorithm ($i_{f_LMS}$)

load current ($i_{f_BIL}$) is estimated under wide load variations. Moreover, the proposed control algorithm shows less complexity as compared to SRF and LMS techniques since PLL and other transformation blocks are not required. The extraction of fundamental current ($i_{f_LMS}$) through LMS algorithm response is slow during load change. The convergence of weights takes around 2–3 cycles for the SRF controller but shows continuous oscillations as shown in Fig. 10a–b. This shows that the advanced control technique based on brain learning works well under different situations. After one or two cycles, the steady state is reached.

## 8 Conclusion

In this paper, a brain-inspired learning technique is modeled and proposed for the control of single-phase VSC integrated EV charger. The capacitor of VSC is connected with the EV charging junction, and smooth charging/discharging operation in a phase-connected system is analyzed. The designed controller not only mitigates power quality problems in source current ($i_s$) but also ensures bidirectional G2V and V2G operations. The proposed BIL control technique is innovative in nature; theoretically analyzed, simulated, and validated on an experimental prototype. The controller design is presented in detail, and appropriate gain parameters are selected to obtain good performance in terms of phase margin, overshoot, and settling time. Both the steady state and dynamic responses during the charging/discharging operation are determined to be satisfactory and the experimental results are validated. The comparative analysis of the proposed BIL algorithm with SRF and LMS techniques has shown the superiority of the BIL algorithm in terms of faster system response (< 2 cycles) with reduced total harmonic distortion of 4.56% in grid current. The designed model illustrated in the paper has shown promising results for load compensation as well as smooth operation of EV chargers for aggregated load and EV stations. However, the proposed BIL controller does not take into account the large scale and spread of EV stations, their variability throughout the day, and the complexity of processing such large-scale data.

## Appendix

Experimental Parameters: AC supply voltage = 40V, DC link voltage = 80V, battery bank voltage ($V_{bat}$) = 48V, battery current ($I_{bat}$)= ±3A, diode rectifier with resistive (R) = 125Ω and inductive (L) = 80 mH nonlinear load, interfacing inductor = 5mH, switching frequency=20kHz, sampling time = 50μsec.

## References


1. Vazquez, S.; Lukic, S.M.; Galvan, E.; Franquelo, L.G.; Carrasco, J.M.: Energy storage systems for transport and grid applications. IEEE Trans. Ind. Electron. **57**(12), 3881–3895 (2010)
2. Saber, A.U.; Venayaganiirthy, G.K.: Plug-in vehicles and renewable energy sources for cost and emission reductions. IEEE Trans. Ind. Electron. **58**(4), 1229–1238 (2011)
3. Roberto V. R. et al., "Electric vehicles in power systems with distributed generation: Vehicle to microgrid (V2M) project," in Proc. IEEE EPQU,2011, pp. 1–6.
4. Yoshimi K. et al., "Practical storage and utilization of household photovoltaic energy by electric vehicle battery," In: Proc. IEEE ISGT, 2012,pp. 1–8.
5. Kumar, P.; Bhaskar, D.V.; Behera, R.K.; Muduli, U.R.: Continuous fast terminal sliding surface-based sensorless speed control of pmbldcm drive. IEEE Trans. Industr. Electron. **70**(10), 9786–9798 (2022)
6. Liang, H.; Liu, Y.; Li, F.; Shen, Y.: Dynamic economic/emission dispatch including PEVs for peak shaving and valley fifilling. IEEE Trans. Ind. Electron. **66**(4), 2880–2890 (2019)
7. Wang, Z.; Wang, S.: Grid power peak shaving and valley fifilling using vehicle-to-grid systems. IEEE Trans. Power Del. **28**(3), 1822–1829 (2013)
8. Acharya, S.; Dvorkin, Y.; Pandži´c, H.; Karri, R.: Cybersecurity of smart electric vehicle charging: A power grid perspective. IEEE Access **8**, 214434–214453 (2020). https://doi.org/10.1109/ACCESS.2020.3041074
9. Yilmaz, M.; Krein, P.T.: Review of the impact of vehicle-to-grid technologies on distribution systems and utility interfaces. IEEE Trans. Power Electron. **28**(12), 5673–5689 (2013). https://doi.org/10.1109/TPEL.2012.2227500
10. Alajmi, B.N., Abdelsalam, I., Marei, M.I. and Ahmed, N.A.: On-board battery charger based-on multifunctional two-stage topology for electric vehicles, 2023 IEEE PES/IAS PowerAfrica (pp. 1–5). IEEE. (2023)
11. Alajmi, B.N., Abdelsalam, I., Marei, M.I. and Ahmed, N.A.: (2023), Two stage single-phase EV on-board charger based on interleaved cascaded non-inverting buck-boost converter, 2023 IEEE Conference on Power Electronics and Renewable Energy (CPERE) (pp. 1–6). IEEE.
12. Bajaj, M.; Flah, A.; Alowaidi, M.; Sharma, N.K.; Mishra, S.; Sharma, S.K.: A lyapunov-function based controller for 3-phase shunt active power filter and performance assessment considering different system scenarios. IEEE Access **9**, 66079–66102 (2021). https://doi.org/10.1109/ACCESS.2021.3075274
13. Zong, X.; Gray, P.A.; Lehn, P.W.: 'New metric recommended for IEEE standard 1547 to limit harmonics injected into distorted grids.' IEEE Trans. Power Del. **31**(3), 963–972 (2016)
14. Kececioglu, O.F.; Acikgoz, H.; Yildiz, C.; Gani, A.; Sekkeli, M.: Power quality improvement using hybrid passive filter configuration for wind energy systems. J. Electr. Eng. Technol. **12**(1), 207–216 (2017)
15. Singh, G.K.: 'Power system harmonics research: a survey.' Eur. Trans. Electr. Power **19**(2), 151–172 (2009)
16. Reddy, S.S.; Bijwe, P.R.: Real time economic dispatch considering renewable energy resources. Renew. Energy **83**, 1215–1226 (2015)
17. Rubino, L.; Rubino, G.; Esempio, R.: Linear programming-based power management for a multi-feeder ultra-fast DC charging station. Energies **16**, 1213 (2023). https://doi.org/10.3390/en16031213
18. Bhattacharya S., Divan, D.M., Banerjee B.: "Synchronous frame harmonic isolator using active series filter," in Proc. Eur. Power Electron. Conf., pp. 3030–3035, 1991.
19. Akagi, H., Kanazawa, Y., Nabae, A.: "Generalized theory of the instantaneous reactive power in three-phase circuits," in Proc. IEEE and JIEE IPEC, pp. 821–827, (1983)
20. Mittal S., Singh A., Chittora, P.: "EV Control in G2V and V2G modes using SOGI Controller," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-6, https://doi.org/10.1109/GCAT55367.2022.9972182
21. Verma, A., Singh, B.: "Reweighted Sparse- Least-Mean Mixed-Norm Adaptive Control for Solar PV Integrated EV Charging Station," 2018 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), Delhi, India, pp. 653–658, (2018) https://doi.org/10.1109/ICPEICES.2018.8897351.
22. Neira, S.; Pereda, J.; Rojas, F.: Three-port full-bridge bidirectional converter for hybrid DC/DC/AC systems. IEEE Trans. Power Electron. **35**(12), 13077–13084 (2020). https://doi.org/10.1109/TPEL.2020.2990667
23. Mittal, S.; Singh, A.; Chittora, P.: Design and development of leaky least mean fourth control algorithm for single-phase grid-connected multilevel inverter. Int J Circ Theor Appl. **52**(1), 328–345 (2024). https://doi.org/10.1002/cta.3746
24. Chaurasiya, S., Bhim S.: "A Single-Phase Low Cost, Compact/High Power Density Portable EV Charger for High Voltage EV Battery Packs with Weak/Strong Grid Operation Capability." 2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT). IEEE, (2022).
25. Gorla, N.B.Y., et al. "Improved utilization of grid connected voltage source converters in smart grid through local VAR compensation." IECON 2015–41st Annual Conference of the IEEE Industrial Electronics Society. IEEE, (2015).
26. Kumar, S., Tapesh, U., Om H.G.: "Power quality improvement and signal conditioning of PV array and grid interfaced off-board charger for electric vehicles with V2G and G2V capabilities." Chinese Journal of Electrical Engineering (2023).
27. Shipra, K.; Maurya, R.; Sharma, S.N.: Brayton-moser passivity based controller for electric vehicle battery charger. CPSS Trans. Power Electron. Appl. **6**(1), 40–51 (2021). https://doi.org/10.24295/CPSSTPEA.2021.00004
28. Arya, S.; Singh, B.: Neural network based conductance estimation control algorithm for shunt compensation. IEEE Trans. Ind. Informatics **10**(1), 569–577 (2014)
29. Hashim, H.F.. Omar, R., Rasheed, M.: "Design and analysis of a three phase series active power filter (SAPF) based on hysteresis controller" in 4th IET Clean Energy and Technology Conference (CEAT 2016), https://doi.org/10.1049/cp.2016.1310
30. Saxena, H.; Singh, A.; Rai, J.N.: Adaptive Splinebased PLL for synchronisation and power quality improvement in distribution system. IET Gener., Transmission Distrib. **14**(7), 1311–1319 (2020)

31. Mei, Y.; Tan, G.; Liu, Z.: An improved brain-inspired emotional learning algorithm for fast classification. Algorithms **10**, 70 (2017). https://doi.org/10.3390/a10020070
32. Morén, J., Balkenius, C.: A Computational Model of Emotional Learning in the Amygdala. In: Proceedings of the 6th International Conference on the Simulation of Adaptive Behaviour, Eds.; MIT Press: Cambridge, UK, pp. 115–124, (2000).
33. LeDoux, J.E: "Emotion and the limbic system concept", **2**, 169–199, (1991).
34. Lotfi E.: "Wind power forecasting using emotional neural networks", In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, San Diego, CA, USA, pp. 311–316, (2014).
35. Lin, C.M.; Chung, C.C.: Fuzzy brain emotional learning control system design for nonlinear systems. Int. J. Fuzzy Syst. **17**, 117–128 (2015)
36. Chung J., Kim, H.: "Brain Emotional Learning Based Intelligent Controller Design for DC Motor Speed Control. In: 2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), pp. 000069–000074, (2019), https://doi.org/10.1109/IWOBI47054.2019.9114498.
37. Jafarzadeh, S.; Mirheidari, R.; Jahed, M.; Barkhordar, M.: Designing PID and BELBIC controllers in path tracking problem. Int, J. Comput. Commun. Control **3**, 343–3484 (2008)
38. Mehrabian, A.R.; Lucas, C.: Intelligent adaptive control of nonlinear systems based on emotional learning approach. Int. J. Artif. Intell. Tools **16**, 69–85 (2007)
39. Lotfi, E.; Akbarzadeh, T.: Brain emotional learning-based pattern recognizer. Cybernet. Syst. **44**, 402–421 (2013)

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.

ACS AGRICULTURAL SCIENCE & TECHNOLOGY

pubs.acs.org/acsagscitech Article

# Cellulose Microfibers-Embedded Carboxymethyl Tamarind Kernel Gum Hydrogels as Soil Conditioners and Plant Nutrients Carriers

Tarun Kumar Gayen, Mohammad Amdad Ali,* and Sudhir G. Warkar*

Cite This: https://doi.org/10.1021/acsagscitech.5c00092 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via DELHI TECHNOLOGICAL UNIV on May 16, 2025 at 10:57:11 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

**ABSTRACT:** Synthesis of rice straw cellulose (RSC) microfibers-embedded carboxymethyl tamarind kernel gum (CMTKG) hydrogels was demonstrated, in which the cellulose microfibers serve as a flexible skeleton, contributing to the overall strength and resilience of the hydrogels, while the CMTKG (with a degree of substitution of 0.2) provides carboxyl groups that help retain water molecules within the hydrogel matrix. Freeze−thaw technique was utilized for the gelation process using epichlorohydrin (ECH) as a cross-linker. The synthesized hydrogel [Formulation: 0.1 g of RSC, 0.3 g of CMTKG and 1 mL of ECH] achieved the highest equilibrium swelling ratio of ∼7722% at ∼18 °C in distilled water. Crucially, this hydrogel was upcycled from agricultural waste (rice straw and tamarind seeds). Thus, it is biofriendly and suitable for agricultural applications. The hydrogel-incorporated soil showed a 33% higher maximum water-holding capacity (MWHC) and a 23% higher bulk density (BD), compared with the control. When manganese (Mn) was loaded into the hydrogels, it demonstrated ∼80% Mn release after 184 h in distilled water. Release kinetics showed that the Peppas−Sahlin model was the best-fit model, indicating the release of Mn from the CMTKG-RSC hydrogel matrix with polymer relaxation and Fickian diffusion. The Mn-loaded hydrogel exhibited ∼80% Mn release in soil after 9 days, and release kinetics indicated that the Peppas−Sahlin model was the best-fit model. Furthermore, the Mn-loaded hydrogel significantly enhanced the growth of blackberry plants in pot trials compared to the control group.

**KEYWORDS:** *agricultural waste, biodegradation, controlled fertilizer release, rice straw cellulose, recycle, soil conditioner*

## 1. INTRODUCTION

The global population is expected to double by 2050, significantly increasing the demand for food, while agricultural land remains limited, driving a higher demand for fertilizers.[1,2] Commercially produced fertilizers are expensive and suffer from low nutrient use efficiency.[2] Unused fertilizers contaminate the environment and impede the development of sustainable agriculture.[3] This inefficiency of the synthesized fertilizers has spurred the development of advanced fertilizer systems, such as controlled-release fertilizers, where the slow release of nutrients ensures optimal availability around the roots when needed, reducing nutrient loss to the environment and promoting healthy plant growth.[4] Fertilizer-incorporated polymer composites have been extensively studied for this purpose.[5] However, it is important to note that synthetic polymers are expensive and not environmentally friendly.[6] Taking this into account, biomolecule-based composite fertilizers have been investigated for their biodegradability and renewable nature.

In recent years, hydrophilic polymers have been cross-linked to form a network, known as hydrogels, which helps retain large amounts of water, thereby adding water retention capacity to the fertilizer-loaded polymer network.[7] These fertilizer-loaded polymeric hydrogels can perform dual purposes: slow release of fertilizers and maintaining soil moisture.[8] The water retention capability of hydrogels reduces the need for frequent irrigation and is effective in arid environments.[9] Research efforts have focused on increasing the swelling ratio of biopolymer-based hydrogels by combining different biomolecules with various functional groups. For example, nanocellulose/wood ash-reinforced starch (contains −OH groups)−chitosan (contains $-NH_2$ groups) hydrogel composites have been reported to achieve a swelling ratio as high as 420%,[10] while jackfruit peel cellulose nanocrystal−alginate (contains −COOH groups) hydrogels exhibited a swelling ratio of 300%.[11] The combined presence of various functional groups (−OH, −COOH, and $-SO_4^{2-}$) in the hydrogel matrix has also been studied. For example, composite hydrogels made from carboxymethyl cellulose and gelatin (contains $-NH_2$, −COOH, and −OH groups) demonstrated a swelling ratio of 373%.[12] Hydrogels made from hydroxyethyl cellulose (HEC), carboxymethyl cellulose (CMC), and divinyl sulfone (DVS) have been reported to achieve swelling ratios ranging from 400 to 1300%.[13]

It is to be noted that as the density of functional groups increases, the resulting repulsion between them leads to polymer chain relaxation, creating more space in the hydrogel matrix for water molecules.[14] However, a further increase in the density of functional groups can result in the structural disintegration of the hydrogel, thus limiting its maximum water uptake and stability.[15] This issue can be addressed by

Received: January 28, 2025
Revised: April 24, 2025
Accepted: April 25, 2025

ACS Publications © XXXX American Chemical Society

increasing the amount of cross-linking agent; however, as cross-linking density increases, the pore size of the hydrogel decreases, reducing the available space to accommodate water molecules.[16] To adjust these two competing factors, we have proposed a unique hydrogel design that incorporates a flexible skeleton embedded in a highly absorptive cross-linked superabsorbent polymer (SAP) matrix.[17]

This design has been inspired by plant cells in which cellulose fibers provide structural support and rigidity. The fibrils consist of long chains of glucose molecules linked together, forming a strong network that contributes to the overall strength and resilience of plant tissues. We selected these cellulose microfibrils as flexible skeletons embedded in the hydrogels, which are expected to help maintain the structural integrity of the hydrogel even in an extended swelling state. As the SAP, we selected carboxymethyl tamarind kernel gum (CMTKG), a natural polymer with a high water-holding capacity (up to 648.19%).[14]

We isolated cellulose from rice straw (RS) and synthesized CMTKG hydrogels with a cellulose skeleton and investigated their chemical and mechanical properties. We demonstrated the upcycling of agricultural waste into valuable ingredients. The effects of these prepared hydrogels on soil conditioning, such as water retention, slow release of fertilizers, bulk density, and soil porosity, were also studied. The application of these hydrogels was examined for the cultivation of chickpeas and blackberries. Furthermore, as a proof-of-concept measurement, manganese was loaded into the hydrogel, and its release in water and soil was studied. Manganese was selected as it is an essential micronutrient for plants, playing a key role in photosynthesis, enzyme activation, and metabolic processes.[18]

## 2. EXPERIMENTAL PROCEDURES

**2.1. Materials.** Rice straw was collected from agricultural fields in Delhi, India. Carboxymethyl tamarind kernel gum (CMTKG) (0.2° substitution, Mol. Wt. = $8.5 \times 10^7$ g/mol) was generously provided by Hindustan Gum and Chemicals Ltd., Bhiwani, Haryana, India. Epichlorohydrin (ECH) (98%), sodium hydroxide (98%), urea (99%), toluene (99%), ethanol (99.9%), hydrochloric acid (36%), acetic acid (99.8%), and hydrogen peroxide (30%) were purchased from Thermo Fisher Scientific (Delhi, India). All solutions were prepared by using deionized (DI) water.

**2.2. Instrumentation.** A temperature-controlled hot plate with a magnetic stirrer (Remi Elektrotechnik Limited, New Delhi, India) was used. A hot air oven (PSO-451, Presto Stantest Private Limited, Haryana, India) was used to dry the hydrogels. A pressure plate apparatus (Model: 1600 pressure extractor, Santa Barbara, CA, USA) was employed to study the effects of hydrogels on the water retention capacity of soil at different metric tensions of 0.33 bar (corresponding to the field capacity) and 1, 5, and 15 bar (corresponding to the permanent wilting point).

**2.3. Chemical Characterization.** An FTIR spectrophotometer (PerkinElmer Spectrum, version 10.5.3) was used to collect IR spectra of the synthesized biopolymers and hydrogels. The scans were run in the range of 500−4000 $cm^{-1}$ with a step size of 0.5 $cm^{-1}$. Surface morphologies of the biopolymers and hydrogels were investigated by scanning electron microscopy (SEM) (JEOL JSM-6610LV). SEM samples were coated with a layer (7−10 nm) of Au under an argon atmosphere using a gold sputter coater JEC 300 model in a high-vacuum evaporator. The base pressure in the SEM chamber during analysis was $1 \times 10^{-7}$ Torr. ImageJ software was used to determine the particle-size distribution of the microparticles. Elemental analysis was performed using energy-dispersive X-ray spectroscopy (EDX) (JEOL JSM 6610LV, Japan). Rheological properties were performed using an MCR-101 rheometer (Anton Paar, Japan) with a parallel plate. Freshly prepared hydrogel was employed for running an amplitude sweep test to ascertain the viscoelastic behavior of the hydrogel. Mn content was determined by measuring the absorbance at 500 nm using a UV−Vis spectrophotometer (UV-9000I).

**2.4. Extraction of Cellulose from Rice Straw.** An alkaline procedure with small modifications to achieve optimal conditions was used to isolate cellulose from rice straw (Figure 1).[19,20] Alkaline

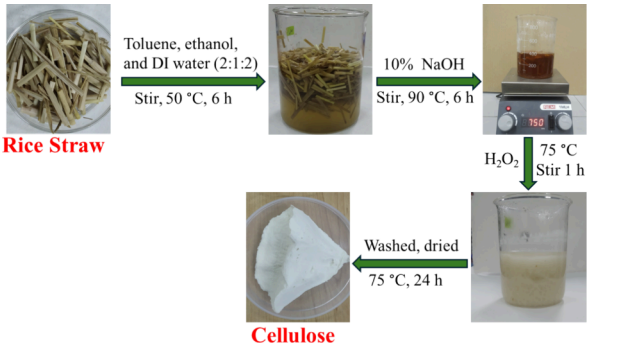



**Figure 1.** Schematic diagram of the isolation process of RS cellulose.

treatment degrades the side chains of esters and glycosides, thus leading to hemicellulose solubilization, cellulose decrystallization, and lignin structural degradation. 20 g of small-cut RS pieces was treated at 50 °C with a solution of toluene, ethanol, and DI water in a 2:1:2 ratio for 6 h, followed by filtration. This step removes nonpolar compounds from the rice straw.[21] The treated RS pieces were dried and then treated with a 10% (w/v) NaOH solution at 90 °C for 6 h. This process breaks down lignin and hemicellulose into soluble compounds.[22] Afterward, the sample was thoroughly washed with excess water and dried in a hot air oven at 50 °C for 24 h. Subsequently, the RS fibers were treated with hydrogen peroxide (15 mL/L) at 75 °C for 1 h. This process removes residual lignin and hemicellulose and bleaches the cellulose fibers.[22] The resulting cellulose fiber was washed thoroughly with water, dried in a hot air oven at 50 °C overnight, and stored in a desiccator for further use. The cellulose content was estimated to be approximately 40% of the rice straw, which is consistent with the values reported in the literature.[23−25] The viscosity-average molecular weight and degree of polymerization of rice straw cellulose were $3.5 \times 10^4$ and 222.3, respectively (experimental details can be found in the Supporting Information (SI)).

**2.5. Synthesis of Rice Straw Cellulose Hydrogel.** The schematic diagram of the synthesis steps of rice straw cellulose (RSC) hydrogels is shown in Figure 2. First, the RSC of different concentrations (Table S1) was added to a solution containing 8.1 g of DI water, 0.7 g of NaOH, and 1.2 g of urea. Afterward, the mixture was stirred for 2 h, and the solution was kept in a refrigerator at −10 °C for 14 h. The frozen solid was thawed and stirred extensively at room temperature to obtain a transparent solution. The cross-linking agent, ECH (1.0 mL), was added to the solution, and the mixture was stirred at room temperature for 1 h. The mixture was placed in a hot air oven at 50 °C for 20 h to form the RSC hydrogel.[26]

**2.6. Synthesis of CMTKG Hydrogel.** The schematic diagram of the synthesis steps of the CMTKG hydrogels is shown in Figure 2. First, CMTKG of different concentrations (Table S1) was added to a solution containing 8.1 g of DI water, 0.7 g of NaOH, and 1.2 g of urea. After stirring for 2 h, the solution was kept in a refrigerator at −10 °C for 14 h. The frozen solid was thawed and stirred extensively at room temperature to obtain a transparent solution. ECH (1.0 mL) was added to the solution and stirred at room temperature for 1 h. The mixture was then placed in a hot air oven at 50 °C for 20 h to form the CMTKG hydrogel.[14] The viscosity-average molecular weight and degree of polymerization of CMTKG were $8.5 \times 10^7$ and 649.3, respectively (experimental details can be found in SI).

**2.7. Synthesis of RSC-CMTKG Hydrogel.** The schematic diagram of the synthesis steps of the RSC-CMTKG hydrogels is shown in Figure 2. First, CMTKG and RSC were added to a solution


https://doi.org/10.1021/acsagscitech.5c00092
ACS Agric. Sci. Technol. XXXX, XXX, XXX−XXX

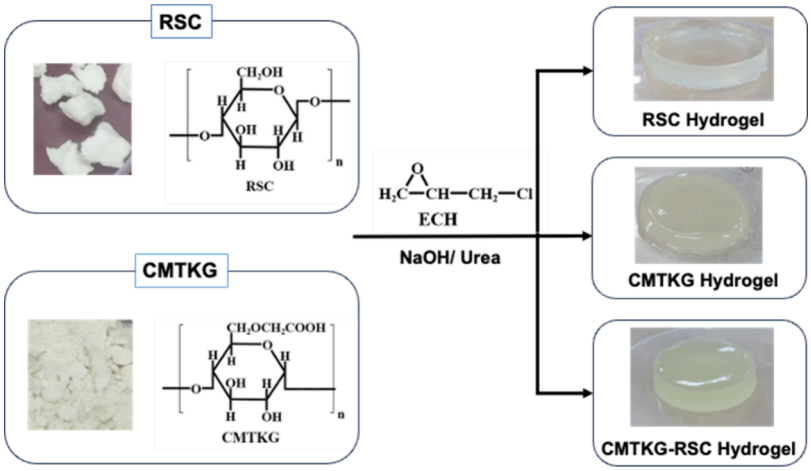



**Figure 2.** Schematic diagram of the synthesis of the hydrogels with different combinations of biopolymers.

containing 8.1 g of DI water, 0.7 g of NaOH, and 1.2 g of urea. Formulations of different concentrations of RSC, CMTKG, and ECH were examined to synthesize the hydrogels (Table S1). After stirring for 2 h, the solution was kept in a refrigerator at −10 °C for 14 h. The frozen solid was thawed and stirred extensively at room temperature to obtain a transparent solution. ECH was added to the solution, and the mixture was stirred at room temperature for 1 h. The mixture was placed in a hot air oven at 50 °C for 20 h to form the CMTKG-RSC hydrogel.[14]

**2.8. Equilibrium Swelling Ratio and Gel Fraction.** The dry hydrogel was weighed ($H_o$) and then soaked in the solution for 96 h. Afterward, it was removed from the solution, and excess surface water was carefully blotted off using filter paper. The swollen hydrogel was then weighed as ($H_E$). This process of weighing was repeated until only a slight decrease in weight was observed. The equilibrium swelling ratio (ESR) was calculated (Table S1) using the formula[26] specified below:

$$\mathrm{ESR}(\%) = \frac{H_E - H_o}{H_o} \times 100\% \quad (1)$$

The dry hydrogel was weighed and then soaked in DI water for 96 h to reach an equilibrium swelling state. Afterward, the hydrogel was removed from the water and air-dried to a constant weight ($H_1$). The gel fraction (GF) was then calculated (Table S1) using the following formula[27]

$$\mathrm{GF}(\%) = \frac{H_1}{H_o} \times 100\% \quad (2)$$

**2.9. Water Retention in Air.** A piece of hydrogel (0.1 g) was soaked in water for 96 h to reach equilibrium swelling ($H_E$). The wet gel was then removed from the solution and allowed to dry naturally at room temperature. Water retention (WR), or the deswelling water ratio of the hydrogel sample, was evaluated using the following equation[28]:

$$\text{Water retention}(\%) = \frac{H_t}{H_E} \times 100\% \quad (3)$$

where $H_t$ is the weight of the sample at time ($t$).

**2.10. Maximum Water-Holding Capacity.** Maximum water-holding capacity (MWHC) is very important for hydrogels in agriculture, as they can modulate moisture content in soil. By retaining large amounts of water, they can help plants survive droughts, leading to healthier plants and yield enhancement.[29] To determine the MWHC of the soil, a fixed quantity of soil was placed in a plastic beaker with small holes at the bottom and layered with filter paper to prevent soil loss. The beakers, along with the control soil sample, were then placed in a water tub to allow maximum water absorption, followed by gravitational drainage to remove excess water. The net weight of water absorbed by the soil was calculated from the weights of the wet soil, dry soil, and filter paper using the following formula[14]:

$$\mathrm{MWHC} = [\text{Wet soil} - \text{Dry soil}] - [\text{Wet filter paper}] \quad (4)$$

**2.11. Available Water Content.** The available water content (AWC) of soil is a critical parameter for the growth of plants.[29] A pressure plate apparatus was used to study the effects of the hydrogel on the water retention capacity of soil at various metric tensions. The soil samples were saturated with water for 12 h, after which the plates were subjected to different metric tensions in a pressure chamber until the water flow from the soil stopped completely. The soil samples were then transferred to moisture boxes, accurately weighed, and subsequently oven-dried at 100 °C for 12 h. After being cooled, the samples were reweighed. The percentage of AWC was calculated using the following formula[14]:

$$\mathrm{AWC}(\%) = \frac{S_{wet} - S_{dry}}{S_{dry}} \times 100 \quad (5)$$

where $S_{wet}$ = weight of the wet soil at specific metric tension. $S_{dry}$ = weight of the dried soil.

**2.12. Determination of Particle Density, Bulk Density, and Porosity of Soil.** A pycnometer was used to determine the particle density of the soil.[14] The soil sample was collected from DTU, Delhi, India. The pH and electrical conductivity of the soil were measured as 8.5 and 0.95 $\mu$S/cm, respectively. The soil was dried in an oven at 100 °C until its weight stabilized. After drying, the samples were mixed with hydrogels of different concentrations (0, 0.1, 0.2, and 0.3%) and filtered through a 2 mm sieve. The oven-dried pycnometer was weighed. The pycnometer was filled with a soil sample and water subsequently, and their weights were measured. By subtracting the weight from the soil-filled pycnometer and DI water-filled pycnometer to the empty pycnometer, we recorded the weight of the soil, and DI water was recorded. The pycnometer was filled with the soil sample, and then DI water was added to remove air within the soil. The soil–water mixture was further heated to remove any air bubbles, and after cooling, the weight of the mixture was measured. To calculate the weight of the soil sample, the weight of the DI water was subtracted from the combined weight of the soil and DI water.

Particle density (PD) was calculated using the formula specified below[14,30]:

$$\mathrm{PD} = \rho_w \times \frac{(W_s - W_a)}{(W_s - W_a) - (W_{sa} - W_w)} \quad (6)$$

where, $W_a$ = weight of the oven-dried pycnometer; $W_s$ = weight of the pycnometer and soil sample; $W_{sa}$ = weight of the pycnometer and soil sample, and water; $W_w$ = weight of the pycnometer and water at room


https://doi.org/10.1021/acsagscitech.5c00092
ACS Agric. Sci. Technol. XXXX, XXX, XXX−XXX

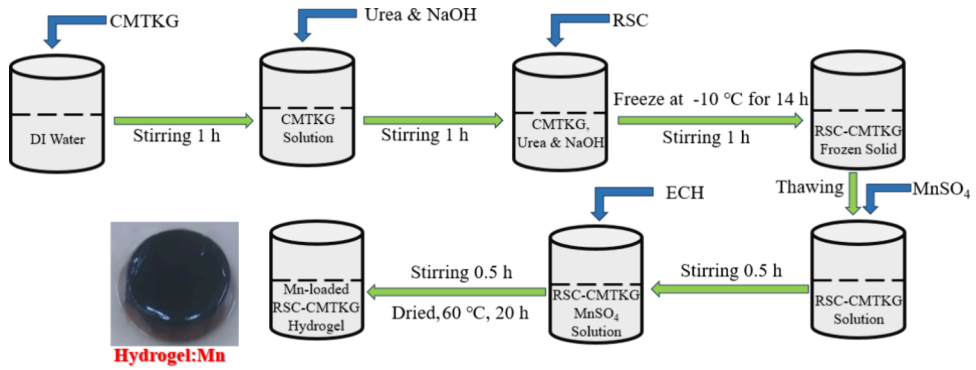



**Figure 3.** Schematic diagram of the synthesis of the Mn-loaded CMTKG-RSC hydrogel.

temperature; $\rho_w$ = density of water at room temperature; $V_a$ = volume of the pycnometer.

The bulk density (BD) was calculated by using the following formula[14,31]:

$$\mathrm{BD} = \frac{(W_s - W_a)}{V_a} \tag{7}$$

The porosity of the soil sample was also calculated using the following formula[14]:

$$\mathrm{Porosity}(\%) = \left[1 - \frac{\mathrm{BD}}{\mathrm{PD}}\right] \times 100 \tag{8}$$

**2.13. Synthesis of Mn-Loaded Hydrogel.** RSC (0.1 g) and CMTKG (0.3 g) were added to a solution of NaOH, urea, and water (7:12:81), and the mixture was stirred for 2 h (Figure 3). The homogeneous mixture was placed in a refrigerator at −10 °C for 14 h. The frozen solid was thawed, $MnSO_4$ solution of different concentrations was added dropwise, and then ECH was added. The resultant solution was placed in a hot air oven at 60 °C for 20 h. Three formulations were applied by varying the concentration of $MnSO_4$. The concentrations of $MnSO_4$ were 0.1, 0.05, and 0.01 M, respectively, and the Mn-loaded CMTKG-RSC hydrogel was leveled as Hydrogel:Mn-1, Hydrogel:Mn-2, and Hydrogel:Mn-3, respectively.[26]

**2.14. Loading Efficiency of Mn.** The following equation is used to calculate the maximum loading efficiency ($L_{max}$) of manganese.[1]

$$L_{max}(\%) = \frac{H_{Mn} - H_o}{H_o} \times 100\% \tag{9}$$

where $H_{Mn}$ is the dry weight of the manganese-loaded CMTKG-RSC hydrogel, and $H_0$ is the dry weight of the CMTKG-RSC hydrogel.

**2.15. Cumulative Mn Release.** To understand plant nutrient uptake from the soil, a release kinetics of nutrients into the soil is essential. A study of nutrient release kinetics helps design nutrient management strategies for nutrient optimization and environmental pollution reduction.[32] Therefore, we investigated Mn release kinetics in water and soil.

To investigate Mn release in water, 1.2 g of Hydrogel:Mn-1, Hydrogel:Mn-2, and Hydrogel:Mn-3 was submerged in 100 mL of DI water separately, and after a certain time interval, 5 mL of manganese-containing water was collected.

To investigate Mn release in soil, 1.4 g of dry Hydrogel:Mn-1 was mixed with 200 g of dried soil, and the soil/Hydrogel:Mn-1 mixture was transferred into a leaching pot. 50 mL of DI water was added slowly to the soil/Hydrogel:Mn-1 mixture. To maintain the soil/Hydrogel:Mn-1 mixture in a consistent environment throughout the experiment, the weight of the leaching pot was kept constant by periodically adding DI water. Leached water (5 mL) was collected by opening the valve, and the Mn content was determined periodically to calculate the cumulative release.

Mn-released water (5 mL) was mixed with 1.6 mL of diluted sulfuric acid, 0.8 mL of phosphoric acid, 0.1 g of potassium periodate, and 8 mL of DI water. The mixture was placed in a water bath and heated at 90 °C for 30 min until a purple-red coloration was formed. After cooling, Mn content was determined by measuring the absorbance at 500 nm using a UV−vis spectrophotometer.

The cumulative release percentage, $R_{cum}$ (%), of Mn was calculated using the equation[33]:

$$R_{cum}(\%) = \frac{M_t}{M_\infty} \times 100\% \tag{10}$$

where $M_t$ is the amount of Mn released at a time ($t$), and $M_\infty$ is the amount of Mn released at equilibrium.

**2.16. Mn Release Kinetics.** The Mn release kinetics were further investigated with the following models[27]:

$$\text{The Korsmeyer} - \text{Peppas model: } \frac{M_t}{M_\infty} = K_{KP} t^n \tag{11}$$

$$\text{The Higuchi model: } \frac{M_t}{M_\infty} = K_H t^{1/2} \tag{12}$$

$$\text{The Peppas} - \text{Sahlin model: } \frac{M_t}{M_\infty} = k_1 t^m + k_2 t^{2m} \tag{13}$$

$$\text{The Zero} - \text{order kinetics model: } \frac{M_t}{M_\infty} = K_o t \tag{14}$$

$$\text{The first} - \text{order kinetics: } \frac{M_t}{M_\infty} = 1 - e^{-K_1 t} \tag{15}$$

where $K_{KP}$ and $n$ are the release constant and diffusion exponent, respectively. $K_H$ is the Higuchi dissolution constant. The Higuchi dissolution constant quantifies the rate at which a substance is released from a matrix system. $k_1$ and $k_2$ are diffusional constants related to the contributions of the Fickian and non-Fickian mechanisms. $m$ is the coefficient of the purely Fickian diffusion exponent. $K_o$ and $K_1$ are the release constants of the zero- and first-order models, respectively.

The Korsmeyer−Peppas model describes Mn release patterns, accounting for both Fickian and non-Fickian diffusion behaviors in the swelling and release system. The release mechanism of Mn can be determined by analyzing the diffusion exponent $n$. When $n < 0.45$, the process follows the Fickian diffusion. For values within the range $0.45 < n < 0.89$, the diffusion process is non-Fickian. If $n > 0.89$, the process is dominated by a simple dissolution mechanism. The Peppas−Sahlin model, on the other hand, integrates both Fickian and non-Fickian contributions. It is important to note that these mathematical models are only applicable for the first 60% of the diffusion process. The zero-order model describes urea release that occurs at a constant rate, independent of concentration, over a specified period. In contrast, the first-order model explains release behavior where the rate is constant within a specific time frame but is dependent on the Mn concentration.


https://doi.org/10.1021/acsagscitech.5c00092
ACS Agric. Sci. Technol. XXXX, XXX, XXX−XXX

**2.17. Statistical Analysis.** ANOVA was used to analyze the parameters of plant growth, and the results were processed statistically. One-way ANOVA was processed at the percentage probability level ($p = 0.05$) to evaluate the significance of differences between any two treatment averages.

## 3. RESULTS AND DISCUSSION

**3.1. Synthesis of the Hydrogels.** Combinations of RSC, CMTKG, and cross-linker concentrations were examined to synthesize hydrogels, and further, their water sorption behavior and gel fraction were investigated. The intrinsic viscosity of RSC and CMTKG was 124.38 and 329.89 $mLg^{-1}$, respectively (experimental details can be found in SI). The isolated RSC was then dissolved in an aqueous solution of NaOH and urea. In this system, NaOH provides the necessary alkalinity to facilitate dissolution by ionizing −OH groups in the cellulose, while urea acts as a cosolvent that helps disrupt the hydrogen bonds between cellulose molecules and resides between them.[20] It should be noted that for the cross-linking reaction to proceed, the polymeric chains need to be in close proximity. Thus, it is expected that when the urea concentration exceeds a certain threshold, it inhibits the gelation reaction.

To determine the optimum ratio of RSC and CMTKG to urea for the gelation reaction, we plotted the gel fraction as a function of (RSC and CMTKG) to urea (Figure 4a). It is observed that the gel fraction increases linearly as the concentration of RSC and CMTKG increases and urea decreases, reaching an optimum state with a gel fraction of 80% when the ratio of (RSC and CMTKG) to urea is 0.5. Beyond this point, further increases in RSC and CMTKG, coupled with a decrease in urea, showed minimal effects on the gel fraction. This suggests that at a ratio of 0.5 for (RSC and CMTKG) to urea, the polymeric chains are in proximity, allowing the cross-linking reaction to occur.

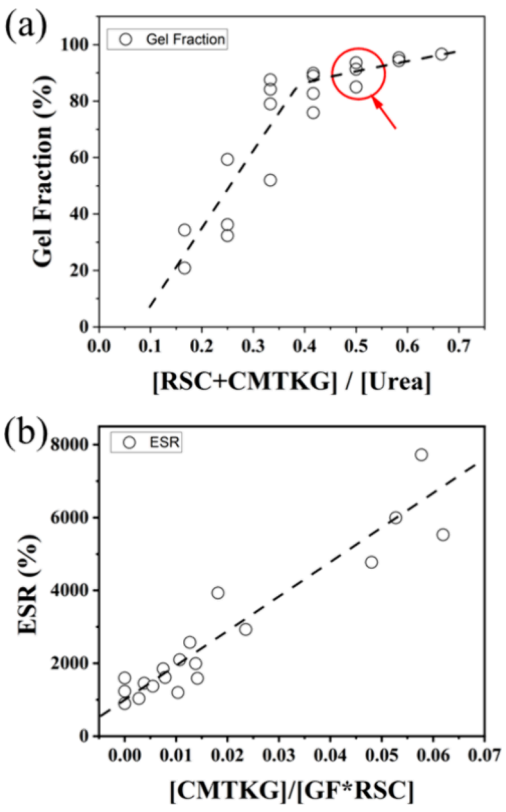



**Figure 4.** (a) Plot of gel fraction versus (RSC+CMTKG)/Urea. (b) Plot of ESR versus CMTKG/(Gel Fraction × RSC). Dashed lines serve as guides to the eye.

Furthermore, since CMTKG provides −COOH groups, it is expected that an increase in the CMTKG content will enhance the water-holding capacity of the hydrogels. In contrast, RSC cellulose tends to form microfibrils under favorable conditions. These microfibrils are highly oriented and crystalline, limiting the available spaces for accommodating water molecules. Additionally, the gel fraction increases, meaning that hydrogels are more integrated, and the spaces within the hydrogel matrix decrease. To understand the effects of the CMTKG-to-RSC ratio and gel fraction on the equilibrium swelling ratio, we plotted ESR as a function of the CMTKG-to-RSC ratio and gel fraction, as shown in Figure 4b. It was observed that ESR increased linearly with higher CMTKG content and decreased gel fraction and RSC, reaching an ESR as high as ∼7800% when the ratio of CMTKG to (gel fraction and RSC) was 0.06. Below this ratio, the hydrogel structures were found to disintegrate due to the low gel fraction (less than 20%).

**3.2. Characterization of the Hydrogels.** FTIR spectra of RSC, CMTKG, RSC hydrogel, CMTKG hydrogel, and CMTKG-RSC hydrogel were recorded to characterize the chemical compounds of the biopolymers and hydrogels (Figure 5a). There are no significant differences observed between the FTIR spectra of the RSC and the RSC hydrogel, indicating that their chemical structures are identical (Figure 5a). Broad peaks between 3600 and 3000 $cm^{-1}$ are attributed to O−H stretching vibrations.[34] C−H and C−$H_2$ stretching vibrations were observed at around 2920 $cm^{-1}$ and 1350 $cm^{-1}$, respectively.[35] The strong peak at 1028 $cm^{-1}$ is attributed to C−O, C−C, and C−O−C stretching vibrations.[26] Peaks at 896 and 666 $cm^{-1}$ represent C−O−C stretching vibration for $\beta$,1,4-glycosidic linkages.[33] These peaks are more pronounced in the RSC polymer compared to the RSC hydrogel, suggesting that the cellulose biopolymer formed microcrystalline structures. Hence, these results indicate the successful extraction of cellulose from rice straw.

The peak at 1748 $cm^{-1}$ was observed in CMTKG, representing −C=O stretching (Figure 5a).[36,14] In the spectrum of the CMTKG-RSC copolymer hydrogel, the indicative peaks for −COOH groups are still present, suggesting the presence of CMTKG in the copolymer hydrogel. In addition, the peaks at 896 and 666 $cm^{-1}$ are also noticeable, suggesting the presence of RSC in the copolymer hydrogels.

To assess the purity of the synthesized polymers and hydrogels, energy-dispersive X-ray spectroscopy (EDX) was performed (Figure S1). In all samples, oxygen and carbon were the primary constituents, and no impurities were detected. The C and O elements were estimated as 55 and 45 at %, respectively, in the RSC polymer, and 59 and 40 at % in the RSC hydrogel. The C and O elements were estimated as 58 and 42 at %, respectively, in the CMTKG polymer, and 53 and 37 at %, respectively, in the CMTKG hydrogel. The slightly higher amount of C in the hydrogel is probably due to the addition of the cross-linking agent. C and O were found to be at 59 and 41 at % in CMTKG-RSC hydrogels, respectively.

SEM images were taken to investigate the surface morphologies of the biopolymers and hydrogels (Figure 5b–f). The SEM image (Figure 5b) shows a fibril structure, and the fibers are in bundles; cross-interlocking points and junctions are observed. This suggests that the alkaline and


https://doi.org/10.1021/acsagscitech.5c00092
ACS Agric. Sci. Technol. XXXX, XXX, XXX−XXX

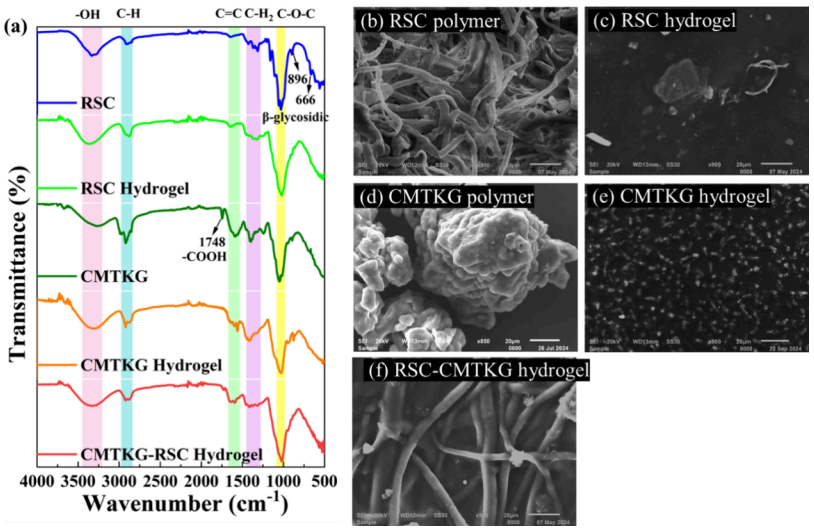



**Figure 5.** FTIR spectra of (a) biopolymers and hydrogels; SEM images of (b) RSC biopolymer, (c) RSC hydrogel, (d) CMTKG biopolymer, (e) CMTKG hydrogel, and (f) RSC-CMTKG hydrogels [hydrogels formulation, Table S1, H-7, H-4, H-11].

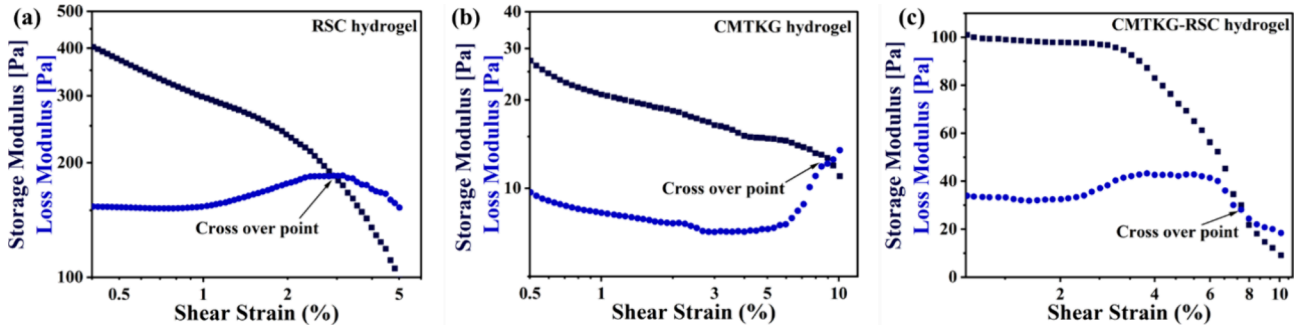



**Figure 6.** Storage modulus and loss modulus as a function of shear strain of: (a) RSC hydrogel (H-8), (b) CMTKG hydrogel (H-4), and (c) CMTKG-RSC hydrogel (H-11).

bleaching processes have separated the plant components into well-separated fibrous strands, and the cellulose chains are arranged in a highly oriented and crystalline fashion. The diameter of the fibers was in the range of 2−5 $\mu$m. This type of fibril texture of cellulose has been well-studied in the literature.[19,37,38] After gelation of the RSC polymer, the surface morphological image of the RS hydrogel is shown in Figure 5c. It is noted that the fibril structure was not observed. This suggests that during the gelation process, cellulose chains were strongly cross-linked, thus preventing the formation of crystalline structures.

Figure 5d depicts a granular structure of the CMTKG polymer.[14] No microcrystalline structure was observed. However, in the CMTKG hydrogel, a continuous spongy-like surface was observed (Figure 5e). The surface topological image of the RSC-CMTKG hydrogel (Figure 5f) reveals the fibril structure of the cellulose molecules. Interestingly, no noticeable microporosity was observed (which was seen in the cellulose polymer), suggesting that during the copolymer gelation process, cellulose molecules reorganize to form a well-defined microfibril structure, and the micropores were filled by highly absorptive CMTKG molecules and then cross-linked by the cross-linker ECH.

An amplitude sweep test was performed to determine the viscoelastic behavior of the RSC hydrogel, CMTKG hydrogel, and RSC-CMTKG hydrogel (Figure 6). The RSC hydrogel showed a viscoelastic region between 0.1 and 3% of strain. The maximum values of the storage modulus ($G'$) and loss modulus ($G''$) of the RSC hydrogel were determined at 403 and 153 Pa, respectively, and the crossover point occurred at around 3% shear strain (Figure 6a), which indicates physical changes from solid to liquid.[39] The highest $G'$ and $G''$ for the CMTKG hydrogel were obtained at 27.27 and 9.69 Pa, respectively. The crossover point of the CMTKG hydrogel was determined at around 9% shear strain (Figure 6b). Furthermore, as RSC was incorporated into CMTKG, the storage modulus ($G'$) and loss modulus ($G''$) increased to 101 and 33.92 Pa, respectively. This suggests that RSC serves as the structural framework in the RSC-CMTKG hydrogel. The crossover point of the CMTKG-RSC hydrogel was determined at around 7.5% shear strain (Figure 6c). In all hydrogels, $G'$ predominates over $G''$ up to the crossover point, indicating that elastic (solid-like) behavior prevails over viscous (liquid-like) behavior, suggesting a typical gel-like behavior. The viscous behavior dominates ($G'' > G'$) after the crossover point of all hydrogels and viscoelastic moduli decrease with strain amplitude, resulting in high deformation with structural breakdown.[39]

**3.3. Effects of pH and Salts on Swelling Behavior.** The swelling capacity of the synthesized optimized hydrogels [Hydrogel: H-11, Table S1] was studied in different media. The parched oven-dried hydrogels were weighed precisely and immersed for 96 h to reach equilibrium state in various pH (=1, 4, 7, 9, and 12) and saline (0.9% NaCl, 0.9% $CaCl_2$, and


https://doi.org/10.1021/acsagscitech.5c00092
ACS Agric. Sci. Technol. XXXX, XXX, XXX−XXX

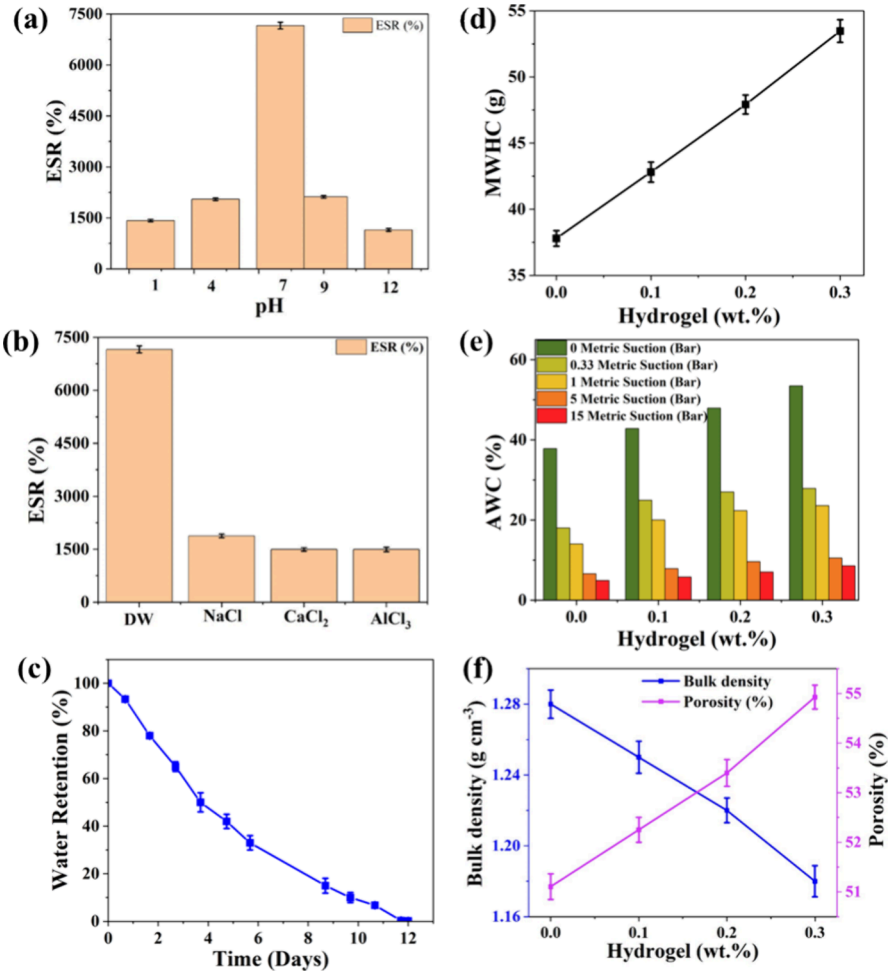



**Figure 7.** Equilibrium swelling ratio of the hydrogels in different (a) pH and (b) salt solutions and (c) water retention by the hydrogels in the open air over time; effects of the hydrogel as soil conditioner: (d) maximum water-holding capacity, (e) moisture retention at different metrics, and (f) bulk density and porosity of the soil [Hydrogel: H-11, Table S1].

0.9% $AlCl_3$) solutions at room temperature (18 ± 2 °C), as shown in Figure 7. The pH was adjusted by adding HCl and NaOH. The effect of the concentrations of RSC, CMTKG, and ECH on ESR and gel fraction were studied (details in SI).

An initial enhancement in ESR was observed as the pH increased from 1 to 7, followed by a decrease in ESR as the pH increased. At pH 7, the −COOH groups in the hydrogel matrix exist as $-COO^-$ ions, which repel neighboring groups, creating more space and allowing additional water molecules to be absorbed.[14] Consequently, an increase in ESR was observed at pH 7. However, when HCl was added to lower the pH, the concentration of $H^+$ ions increased, surrounding the $-COO^-$ groups. This reduced the charge repulsion between $-COO^-$ groups in the hydrogel matrix, leading to hydrogel shrinkage.[40] A similar effect was observed when the pH was raised above 7 by adding NaOH. NaOH introduces $Na^+$ ions, which reduce the charge density of the $-COO^-$ groups, resulting in a decrease in the ESR of the hydrogels.[41]

The ESR of hydrogels in different salt solutions is shown in Figure 7b. It was observed that the introduction of ions reduced the ESR significantly. This effect can be explained by the ionic strength of the cations, which affects the osmotic pressure between the outer and inner media, restricting the entry of water molecules into the hydrogels.[14] The higher the ionic strength, the greater the resistive forces to hinder the diffusion of water molecules; therefore, ESR decreases profoundly. Furthermore, it was observed that the ESR of the hydrogels in ionic solutions followed the order: NaCl > $CaCl_2$ > $AlCl_3$. This is because more $COO^-$ groups are needed for charge neutralization of $Al^{3+}$ compared to bivalent and monovalent cations, such as $Ca^{2+}$ and $Na^+$.

**3.4. Water Retention in Air and Maximum Water-Holding Capacity (MWHC).** Water retention is crucial for plants in acrid environments because it helps them survive harsh, dry conditions where water is scarce.[42] By retaining moisture in their tissues and soil, plants can continue to perform essential functions like photosynthesis and nutrient uptake.[43] Figure 7c shows that water retention gradually decreases with time. It took 12 days for the swollen hydrogel to become fully dry. Some weight loss was found due to the dissolution of the hydrogel.

Hydrogels with high water-holding capacity enhance water retention in soil and reduce water waste.[44] This ability to conserve water also prevents dehydration, reduces salt buildup, and allows plants to withstand extreme temperatures, ensuring their survival in arid or saline environments.[45] It was observed that the control sample held 38 g of water per 100 g of soil. As hydrogels were incorporated into the soil at concentrations of 0.1, 0.2, and 0.3%, the MWHC increased linearly to 43, 48, and 53 g of water per 100 g of soil, respectively, as depicted in


https://doi.org/10.1021/acsagscitech.5c00092
ACS Agric. Sci. Technol. XXXX, XXX, XXX−XXX

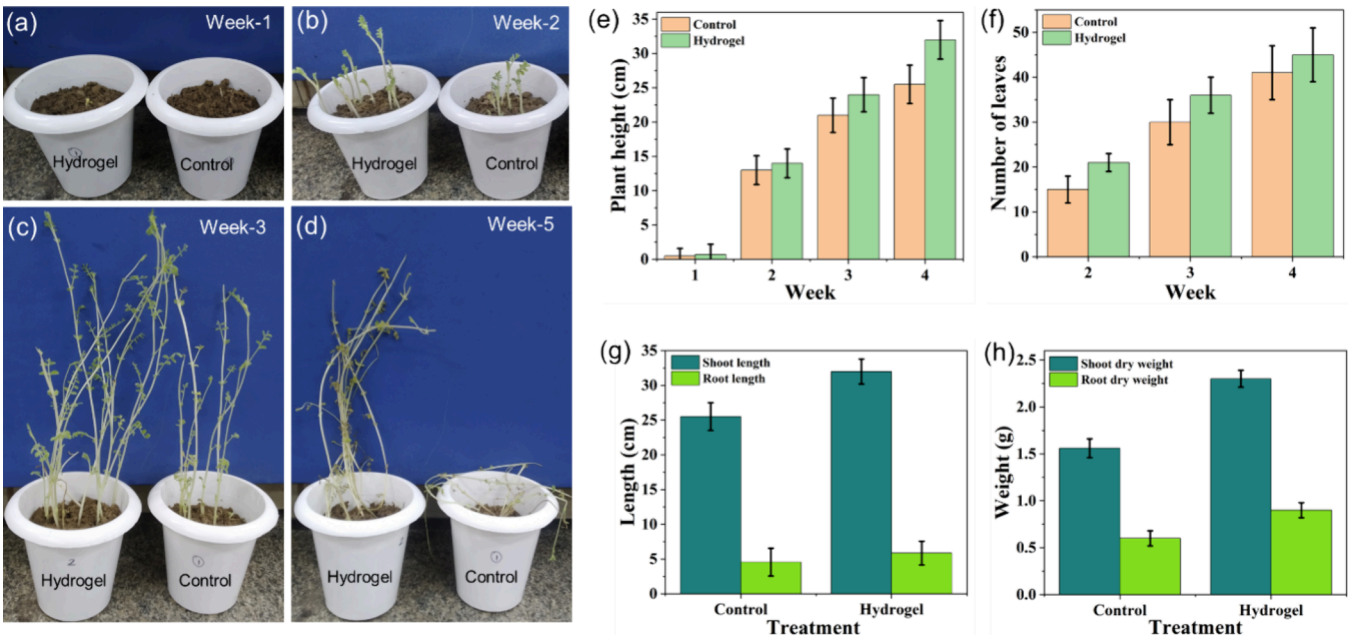



**Figure 8.** Effects of the hydrogel on the growth of chickpeas. Photograph of the plant at (a) week-1, (b) week-2, (c) week-3, and (d) week-5. Growth parameters: (e) plant height (cm), (f) number of leaves, (g) length of shoot and root, and (h) dry weight of shoot and root [H-11, Table S1].

Figure 7d. This corresponds to increases of 13, 24, and 33% in MWHC compared with the control sample. This suggests that hydrogel incorporation enhanced the water retention capacity in soil as it absorbs water.

**3.5. Available Water Content (AWC).** The available water content of soil is vital for irrigation planning and sustainable ecosystem function.[46] It was observed that as the percentage of the hydrogel increases, the AWC in the soil also increases linearly for all metric tensions, as depicted in Figure 7e. The water retention capacity in soil increases by the addition of the hydrogel as it helps retain water in it under pressure. Therefore, these hydrogels could play a vital role in enhancing crop production and survival rate under drought or water-stressed conditions.

**3.6. Particle Density, Bulk Density, and Porosity.** To understand the soil composition, soil properties, soil behavior, and determination of the best uses of the soil, particle density is important to measure. If any soil has a low particle density, this means that the soil contains high organic materials. Bulk density is also an important indicator of soil health. Bulk density controls the available water content, infiltration, soil porosity, aeration, and rooting depth. Bulk density helps assess soil structure, soil compaction, soil erosion, nutrient availability, and soil microorganism activity.[47] Soil porosity is a key indicator of available water. It is very important for plant growth as it controls the balance of water and oxygen to plant roots and soil organisms. It was observed that the bulk density of the soil samples decreases as the content of the hydrogel in the soil increases. The bulk densities are 1.25, 1.22, and 1.18 g/cm$^3$ for the soil containing 0.1, 0.2, and 0.3 wt % of hydrogels, respectively (Figure 7f). The decrease in the bulk density may be due to the presence of hydrophilic groups in the hydrogels. Hydrogels absorb large amounts of water, increasing the volume of hydrogel and soil and, as a result, decreasing soil bulk density. As the particle density decreases with the presence of the hydrogel, this is reflected in the porosity of the soil. The porosity of the control sample was calculated to be 51%, while the soil containing 0.1, 0.2, and 0.3% hydrogels had porosity values of 52, 53, and 55%, respectively. A well-aerated soil is essential for plant growth. As hydrogel-containing soil enhances soil porosity, it boosts plant growth, productivity, and biodiversity.

**3.7. Plantation under Water-Stressed Condition.** To further investigate the effects of hydrogel on plant growth under water-stressed conditions, the seeds of chickpeas were sown, and the survival of the plants under water-stressed conditions was assessed.[14] It was observed that the plants in the control soil wilted quickly. The hydrogel-treated plants survived for a longer duration under water-stressed conditions. After 25 days, the control plants wilted, while the hydrogel-treated plants survived for up to 32 days (Figure 8a–d). After 5 weeks, the shoot length and root length of the control plants were measured at 25.5 and 4.5 cm, respectively, whereas those of the 1% hydrogel-treated plants measured 32 and 5.8 cm. The number of leaves counted in the control group was 15, 30, and 41 after 2, 3, and 4 weeks, respectively, while the 1% hydrogel-treated plants had counts of 21, 36, and 45 at the same intervals. The heights of the control plants were recorded as 0.5, 13, 21, and 25.5 cm after 1, 2, 3, and 4 weeks, respectively, while the 1% hydrogel-treated plants measured 0.7, 14, 24, and 32 cm over the same periods. The shoot and root dry weights for the control plants were found to be 1.6 and 0.6 g, respectively, while the hydrogel-treated plants had dry weights of 2.3 and 0.9 g (Figure 8).

One-way ANOVA was used to analyze the statistical significance between the plant height and number of leaves of chickpeas at 1% CMTKG-RSC hydrogel application. The *p*-values in the *F*-test corresponding to the plant height (Table S4) and number of leaves (Table S5) of chickpeas are 0.769 and 0.631, respectively. The statistical analysis of one-way ANOVA shows that the growth of chickpeas with CMTKG-RSC hydrogel application compared to the control had no significant difference in plant growth, as expressed in the Supporting Information.


https://doi.org/10.1021/acsagscitech.5c00092
ACS Agric. Sci. Technol. XXXX, XXX, XXX−XXX

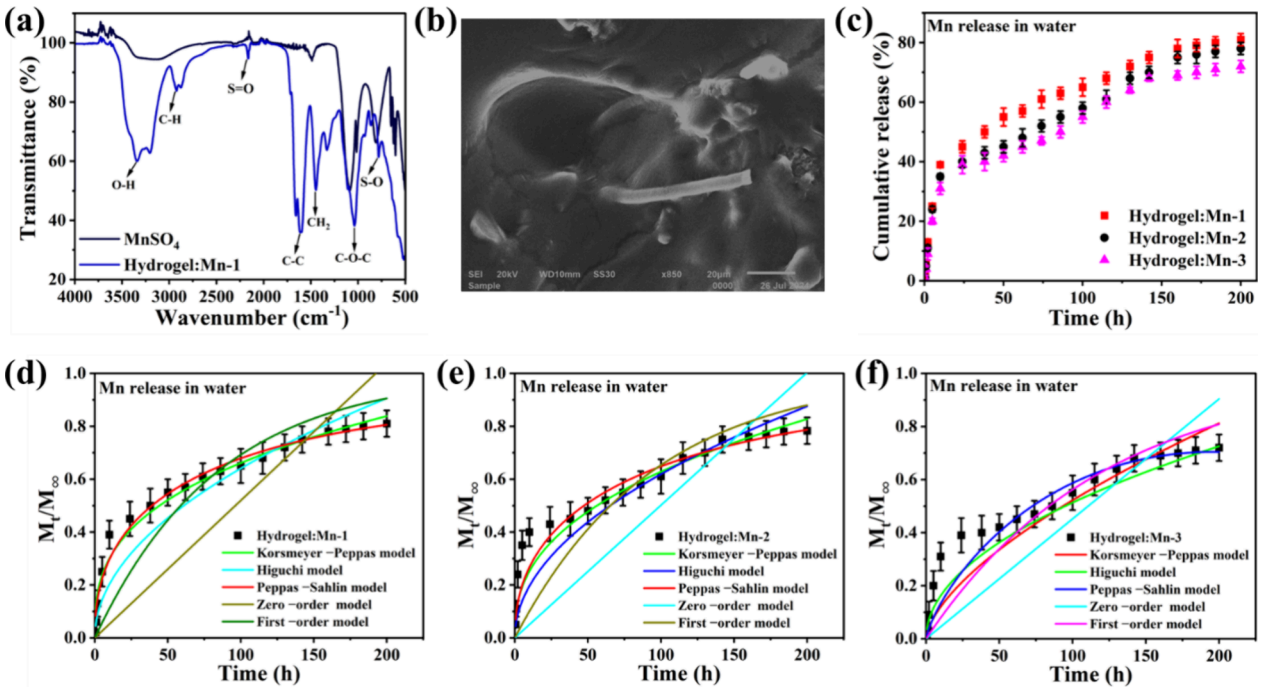



**Figure 9.** Mn-loaded hydrogel (Hydrogel: Mn-1): FTIR spectrum (a), SEM image (b), cumulative release of Mn (c), and Mn release kinetics. 0.1, 0.05, and 0.01 M of $MnSO_4$ were loaded into the hydrogels, labeled as Hydrogel:Mn-1 (d), Hydrogel:Mn-2 (e), and Hydrogel:Mn-3 (f).

**3.8. Mn-Loaded Hydrogels.** Manganese plays a key role in activating life-sustaining enzymes, e.g., dehydrogenase, decarboxylase, and oxidase enzymes, enhancing the disintegration of water in photosynthesis and respiration as well as effective nitrogen metabolism and assimilation.[48] It also accelerates the formation of chlorophyll, germination of plant seeds, and fruit ripening. In addition to that, Mn fertilizer can reduce Cd bioavailability in soil and enhance soil pH.[48] Mn has an antagonistic effect on Cd. As a result, Mn serves as a nutrient and reduces Cd absorption in plants.[48] Therefore, we incorporated Mn into the hydrogels and investigated how these Mn-loaded hydrogels release Mn. We studied the water absorption and water retention capacity of the hydrogels and their effects on the growth of blackberry seedlings.

Controlled-release fertilizers and micronutrients are widely used in agriculture to meet the demand for micronutrients and fertilizers over an extended period.[49] Research has been done to investigate the controlled release of micronutrients and fertilizers by loading them into hydrogels.[50] Here, we selected Mn as a proof-of-concept measurement as it is one of the essential micronutrients for plants, and Mn reduces the heavy metal, Cd, bioavailability in soil through an antagonistic process.[48] Mn was loaded into the hydrogels, and their application in the planting of blackberries was investigated. A randomized pot trial was performed on the growth of blackberries plantation. Blackberry was selected as the proof-of-concept trail. The maximum loading efficiencies of Hydrogel:Mn-1, Hydrogel:Mn-2, and Hydrogel:Mn-3 were 48.8, 25.4, and 5.3%, respectively.

FTIR spectra of $MnSO_4$ and Mn-loaded CMTKG-RSC hydrogel (Hydrogel:Mn-1) were investigated to confirm the presence of Mn in the hydrogel (Figure 9a). The FTIR spectrum of $MnSO_4$ shows a peak at 2113 $cm^{-1}$, which indicates S=O stretching, and a peak at 808 $cm^{-1}$, which indicates S−O stretching. These peaks were also identified in the FTIR spectrum of Hydrogel:Mn-1, where the peak at 2165 $cm^{-1}$ indicates S=O stretching, and the peak at 784 $cm^{-1}$ indicates S−O stretching. Therefore, it is clear that S−O and S=O groups are present in Hydrogel:Mn-1 The surface morphologies of the manganese-loaded hydrogels (Hydrogel:Mn-1) are shown in Figure 9b. It is noted that after loading of Mn, the surface morphology of the hydrogel preserved a similar structure and features as the pristine hydrogel.

**3.9. Manganese Release Kinetics in Water.** After 86 h, the cumulative release rates of manganese from Hydrogel:Mn-1, Hydrogel:Mn-2, and Hydrogel:Mn-3 were observed to be 50.2, 54.9, and 63.1%, respectively, as shown in Figure 9c. Hydrogel:Mn-3 revealed a maximum water absorption of 858.52%. Water absorption enhanced with a lower concentration of manganese. Hydrogel:Mn-1exhibited a minimum water absorption of 683.81%. This study reveals an inverse relationship between water absorption and manganese content in the hydrogel. This could be due to the electrostatic interaction between the $Mn^{2+}$ and $COO^-$ groups in the hydrogel, leading to shrinkage of the hydrogel matrix.

To understand the release behavior of Mn in water, the kinetics data were fitted with different models (Figure 9d), and the fitting parameters are presented in Table S2. It is noted that release behavior can be best described by the Peppas−Sahlin model. This model combines elements from two different mechanisms of Mn release: diffusion and erosion.[27] This suggests that the changes in the hydrogel matrix happen while releasing Mn. These changes might be due to the interaction between the Mn and the $-COO^-$ groups in the hydrogel matrix. From these findings, slow-release nutrients would be helpful for nutrient management in sustainable agricultural practices. Mn release from Hydrogel:Mn-1 follows the Peppas−Sahlin model by diffusion and erosion of hydrogels.

**3.10. Manganese Release Kinetics in Soil.** The cumulative release of Mn in the soil was studied, indicating that Mn releases gradually with time (Figure 10a). The kinetic


https://doi.org/10.1021/acsagscitech.5c00092
ACS Agric. Sci. Technol. XXXX, XXX, XXX−XXX

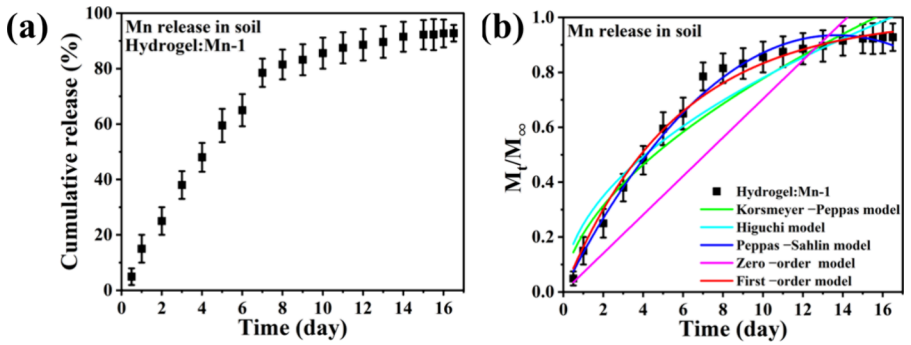



**Figure 10.** (a) Cumulative Mn release in soil. (b) Mn release kinetics models.

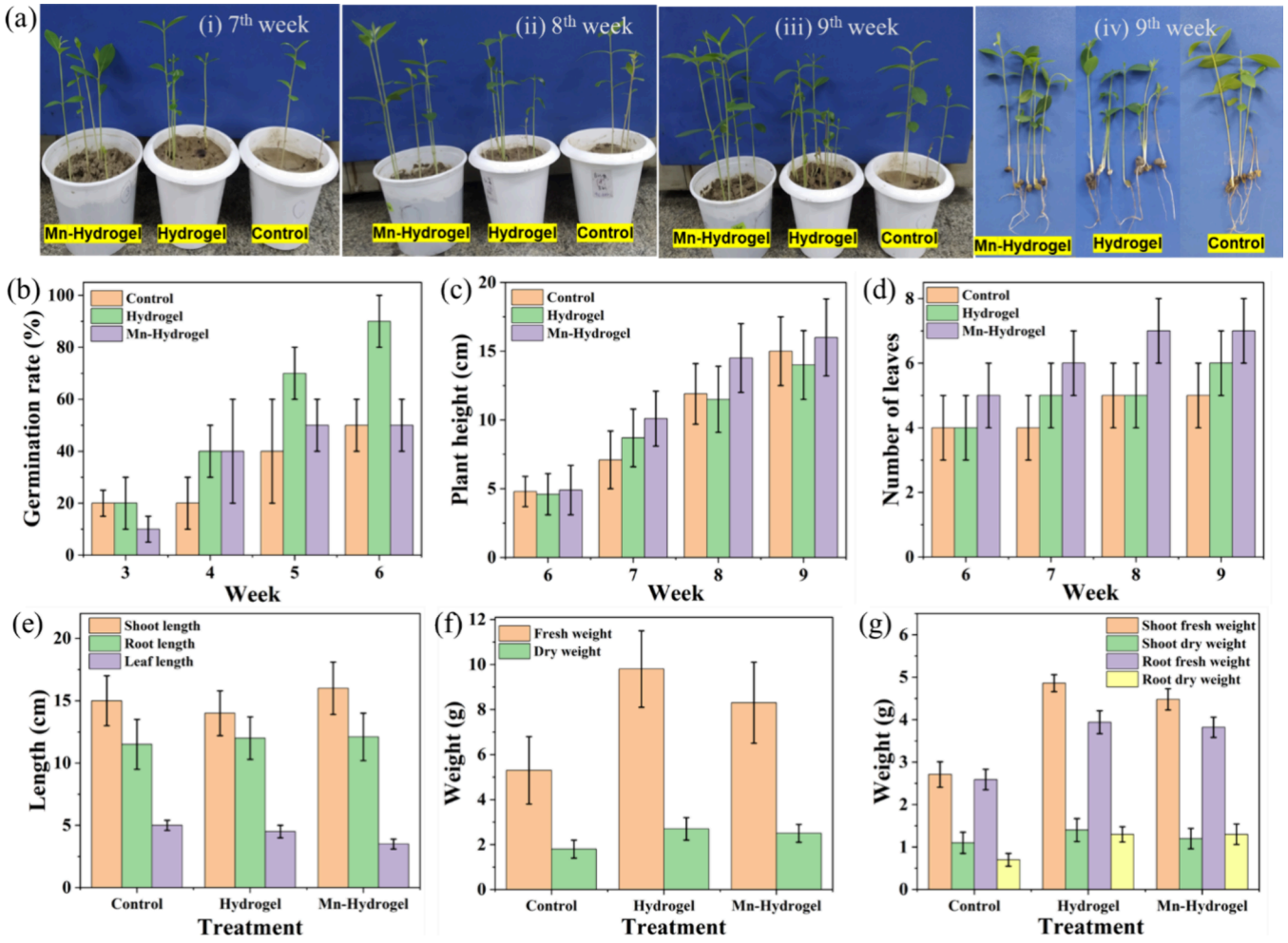



**Figure 11.** Effects of Mn-loaded hydrogel on the growth of blackberry: (a) plants growth in 7th, 8th, and 9th weeks, (b) germination index (%), (c) plant height (cm), (d) number of leaves, (e) lengths of shoot, root, and leaf, (f) fresh and dry weights of the plant, and (g) fresh and dry weights of shoot and root.

parameters for Mn release in soil are listed in Table S3. Different mathematical kinetics models were investigated, showing that the Peppas−Sahlin model and the first-order model were the best fit as the correlation coefficients ($R^2$) were 0.99 (Figure 10b). These findings indicate that Mn molecules are distributed uniformly into the hydrogel, and the release process is controlled by the Mn concentration and degradation or erosion.

**3.11. Blackberry Plantation Study.** The effects of the Mn-loaded hydrogel in soil on the growth of blackberries (random selection) were studied (Figure 11) and compared with the soils containing pristine hydrogel and control, respectively. It is observed that the germination rate of seeds was higher in the pristine hydrogel/soil mixture. However, plant heights and number of leaves were found maximum in Mn-loaded hydrogels at different weeks. Control groups are the lowest in all parameters. All plants were healthy and showed similar shapes and sizes of the shoot, root, and leaf. The shoot length, root length, and leaf length of blackberry plants were almost similar in the treatment of Hydrogel:Mn,


https://doi.org/10.1021/acsagscitech.5c00092
ACS Agric. Sci. Technol. XXXX, XXX, XXX−XXX

pristine hydrogel, and control soil. Fresh weight and dry weight of the plants were more in the treatment of the pristine hydrogel compared to the control and Hydrogel:Mn. The fresh weight of shoot and dry weight of root of blackberry plants were more in pristine hydrogel-treated plants compared to control and Hydrogel:Mn-treated plants. This study shows that plant growth is better in the presence of Mn-loaded hydrogel in soil. These hydrogels improve water retention and reduce irrigation frequency, producing excellent plant growth compared to the control.

The growth parameters of blackberry were analyzed statistically with one-way ANOVA. The *p*-values in the *F*-test corresponding to germination (Table S6), plant height (Table S7), and leaves' number (Table S8) were 0.615, 0.830, and 0.0723, respectively. Statistical analysis with one-way ANOVA showed no significant difference in blackberry growth at the treatment of the hydrogel and Mn-loaded hydrogel in comparison to the control.

In this present study, a biofriendly CMTKG superabsorbent biohydrogel was synthesized, embedding cellulose microfibers as a scaffold. We demonstrated that the gel fraction and swelling ratio can be controlled by formulating the ratios of RSC, CMTKG, ECH, and urea. Our method provides a unique structure in which cellulose microfibrils function as a structural framework, and their porous microchannels facilitate faster transport of water and plant nutrients. Furthermore, CMTKG, with a degree of substitution of 0.2, contributes carboxyl groups, enabling superior water retention and aiding in the controlled release of fertilizer. The ESR of the CMTKG-RSC hydrogel was achieved as high as 7722% in distilled water, with a 51.96% gel fraction. This prepared optimized CMTKG-RSC hydrogel retained water for more than 11 days in open air (~18 °C). The maximum water-holding capacity and bulk density of the optimized CMTKG-RSC hydrogel-incorporated soil were found to be 33% and 23% higher, respectively, compared to the control group. Furthermore, we demonstrated that the prepared hydrogel could be utilized as a plant nutrient carrier. As a proof of concept, Mn was loaded into the hydrogel, indicating 80% Mn release after 184 h in DI water, and its release kinetics in water were studied, which was found to follow the Peppas−Sahlin model, revealing Mn release from the CMTKG-RSC hydrogel matrix with polymer relaxation and Fickian diffusion. Mn release in soil indicated ~ 90% Mn release in soil after 12 days, and its release kinetics followed the Peppas−Sahlin model. These Mn-loaded hydrogels were applied to the blackberry's plantation, showing healthier plant growth performance compared to the control.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsagscitech.5c00092.

Formulations of the hydrogels with different ratios of biopolymers and cross-linker, energy-dispersive X-ray spectroscopy (EDX) spectra, determination of viscosity of polymer solution; determination of viscosity-average molecular weight; viscosity-average degree of polymerization; effects of the concentration of CMTKG on ESR and gel fraction; effects of the concentration of RSC on ESR and gel fraction; effects of cross-linking agent on ESR and gel fraction, kinetic parameters of Mn release in water, Kinetic parameters for Mn release in soil from different kinetics models, and one-way ANOVA for the responses of chickpea and blackberry plantation (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Authors

**Mohammad Amdad Ali** − *Illinois Materials Research Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States*; Email: mali85@illinois.edu

**Sudhir G. Warkar** − *Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India*; orcid.org/0000-0001-7305-282X; Email: sudhirwarkar@gmail.com

### Author

**Tarun Kumar Gayen** − *Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsagscitech.5c00092


### Notes

The authors declare no competing financial interest.

## ■ REFERENCES


(1) Baruah, M.; Gogoi, R.; Borgohain, A.; Karak, T.; Saikia, J. Biodegradable CMC/ZnO/Ni Cross-Linked Hydrogel Beads for Enhanced Plant Growth and Increased Soil Urease Activity. *ACS Appl. Polym. Mater.* **2024**, *6*, 13238−13252.

(2) Dimkpa, C. O.; Fugice, J.; Singh, U.; Lewis, T. D. Development of Fertilizers for Enhanced Nitrogen Use Efficiency−Trends and Perspectives. *Sci. Total Environ.* **2020**, *731*, No. 139113.

(3) Srivastav, A. L.; Patel, N.; Rani, L.; Kumar, P.; Dutt, I.; Maddodi, B. S.; Chaudhary, V. K. Sustainable Options for Fertilizer Management in Agriculture to Prevent Water Contamination: A Review. *Environment, Development and Sustainability.* **2024**, *26* (4), 8303−8327.

(4) Moradi, S.; Babapoor, A.; Ghanbarlou, S.; Kalashgarani, M. Y.; Salahshoori, I.; Seyfaee, A. Toward a New Generation of Fertilizers with the Approach of Controlled-Release Fertilizers: A Review. *Journal of Coatings Technology and Research.* **2024**, *21* (1), 31−54.

(5) Nandal, K.; Vaid, V.; Rahul; Saini, P.; Devanshi; Sharma, R. K.; Joshi, V.; Jindal, R.; Mittal, H. Synthesis and Characterization of κ-Carrageenan and Guar Gum-Based Hydrogels for Controlled Release Fertilizers: Optimization, Release Kinetics, and Agricultural Impact. *Ind. Crops. Prod.* **2025**, *225*, No. 120587.

(6) De, B.; Bera, M.; Bhattacharjee, D.; Ray, B. C.; Mukherjee, S. A Comprehensive Review on Fiber-Reinforced Polymer Composites: Raw Materials to Applications, Recycling, and Waste Management. *Prog. Mater. Sci.* **2024**, *146*, No. 101326.

(7) Hu, H.; Li, P.; Tong, Y. W.; Li, J.; He, Y. Reutilization Recovered Nutrients from Wastewater by Synthesizing Eco-Friendly Slow-Release Hydrogel Fertilizer for Sustainable Vegetable Production. *Chemical Engineering Journal* **2024**, *499*, No. 155943.

(8) Luo, Z.; Li, Y.; Pei, X.; Woon, K. S.; Liu, M.; Lin, X.; Hu, Z.; Li, Y.; Zhang, Z. A Potential Slow-Release Fertilizer Based on Biogas Residue Biochar: Nutrient Release Patterns and Synergistic Mechanism for Improving Soil Fertility. *Environ. Res.* **2024**, *252*, No. 119076.

(9) Abdelfattah, A.; Mostafa, H. Potential of Soil Conditioners to Mitigate Deficit Irrigation Impacts on Agricultural Crops: A Review. *Water Resources Management* **2024**, *38* (8), 2961−2976.

(10) Iqbal, D. N.; Tariq, Z.; Philips, B.; Sadiqa, A.; Ahmad, M.; Al-Ahmary, K. M.; Ali, I.; Ahmed, M. Nanocellulose/Wood Ash-Reinforced Starch-Chitosan Hydrogel Composites for Soil Conditioning and Their Impact on Pea Plant Growth. *RSC Adv.* **2024**, *14* (13), 8652−8664.

(11) Nyoo Putro, J.; Soetaredjo, F. E.; Santoso, S. P.; Irawaty, W.; Yuliana, M.; Wijaya, C. J.; Saptoro, A.; Sunarso, J.; Ismadji, S. Jackfruit Peel Cellulose Nanocrystal−Alginate Hydrogel for Doripenem Adsorption and Release Study. *Int. J. Biol. Macromol.* **2024**, *257*, No. 128502.
(12) Qin, X.; Zhao, Z.; Deng, J.; Zhao, Y.; Liang, S.; Yi, Y.; Li, J.; Wei, Y. Tough, Conductive Hydrogels Based on Gelatin and Oxidized Sodium Carboxymethyl Cellulose as Flexible Sensors. *Carbohydr. Polym.* **2024**, *335*, No. 121920.
(13) Sannino, A.; Maffezzoli, A.; Nicolais, L. Introduction of Molecular Spacers between the Crosslinks of a Cellulose-Based Superabsorbent Hydrogel: Effects on the Equilibrium Sorption Properties. *J. Appl. Polym. Sci.* **2003**, *90* (1), 168−174.
(14) Khushbu; Warkar, S. G.; Kumar, A. Synthesis and Assessment of Carboxymethyl Tamarind Kernel Gum Based Novel Superabsorbent Hydrogels for Agricultural Applications. *Polymer (Guildf)* **2019**, *182*, No. 121823.
(15) Zhang, L.; Li, Q.; Liang, Y.; Zhang, G.; Zou, J.; Fei, P.; Lai, W. Hydrogels Comprising Oxidized Carboxymethyl Cellulose and Water-Soluble Chitosan at Varied Oxidation Levels: Synthesis, Characterization, and Adsorptive toward Methylene Blue. *Int. J. Biol. Macromol.* **2024**, *277*, No. 134351.
(16) Lin, H.; Ning, L.; Dai, T.; Wei, W.; Ma, N.; Hong, M.; Wang, F.; You, C. Application of Bio-Based Nanopesticides for Pine Surface Retention and Penetration, Enabling Effective Control of Pine Wood Nematodes. *Chemical Engineering Journal* **2024**, *502*, No. 158009.
(17) You, C.; Ji, X.; Lin, H.; Ma, N.; Wei, W.; Long, L.; Ning, L.; Wang, F. Preparation of UV-Responsive Hydrogels Based on Nanocellulose and Their Utilization in Fungicide Delivery. *Environ. Sci. Nano* **2024**, *11* (4), 1442−1451.
(18) Akalin, G. O.; Pulat, M. Preparation and Characterization of κ-Carrageenan Hydrogel for Controlled Release of Copper and Manganese Micronutrients. *Polym. Bull.* **2020**, *77* (3), 1359−1375.
(19) Sharma, N.; Allardyce, B. J.; Rajkhowa, R.; Agrawal, R. Rice Straw-Derived Cellulose: A Comparative Study of Various Pre-Treatment Technologies and Its Conversion to Nanofibres. *Sci. Rep.* **2023**, *13* (1), 16327.
(20) Qin, X.; Lu, A.; Zhang, L. Gelation Behavior of Cellulose in NaOH/Urea Aqueous System via Cross-Linking. *Cellulose* **2013**, *20* (4), 1669−1677.
(21) Nishu; Liu, R.; Rahman, M. M.; Li, C.; Chai, M.; Sarker, M.; Wang, Y.; Cai, J. Catalytic Pyrolysis of Microcrystalline Cellulose Extracted from Rice Straw for High Yield of Hydrocarbon over Alkali Modified ZSM-5. *Fuel* **2021**, *285*, No. 119038.
(22) Lai, V. W. X.; Jamaluddin, J.; Mat Nasir, N. A. F.; Baharulrazi, N.; Majid, R. A.; Lai, J. C.; Mohd Bohari, S. P.; Chua, L. S.; Hasham, R.; Rahmat, Z.; Adrus, N. Extraction of Cellulose from Rice Straw for Regeneration of Hydrogels. *Environ. Qual. Manage.* **2022**, *32* (1), 333−341.
(23) Chen, C.; Deng, X.; Kong, W.; Qaseem, M. F.; Zhao, S.; Li, Y.; Wu, A. M. Rice Straws with Different Cell Wall Components Differ on Abilities of Saccharification. *Front Bioeng. Biotechnol.* **2021**, *8*, No. 624314.
(24) Li, S.; Chen, G. Agricultural Waste-Derived Superabsorbent Hydrogels: Preparation, Performance, and Socioeconomic Impacts. *Journal of Cleaner Production.* **2020**, *251*, No. 119669.
(25) Abakar, A.; Dadi Mahamat, A.; Donnot, A.; Tanguier, J.-L.; Benelmir, R. Physical and Chemical Characteristics of Rice Straw. *Research Journal of Applied Sciences, Engineering and Technology* **2020**, *17* (4), 115−121.
(26) Ahmed, M. S.; Maniruzzaman, M.; Al-Mamun, M. R.; Ali, M. A.; Badal, M. M. R.; Aziz, M. A.; Jafar Mazumder, M. A.; Hakeem, A. S.; Yousuf, M. A. Jute Stick-Derived Cellulose-Based Hydrogel: Synthesis, Characterization, and Methylene Blue Removal from Aqueous Solution. *ACS Omega* **2023**, *8* (50), 47856−47873.
(27) Lu, K.; Abouzeid, R.; Wu, Q.; Chen, Q.; Liu, S. Hydrogel Nanocomposite Based Slow-Release Urea Fertilizer: Formulation, Structure, and Release Behavior. *Giant* **2024**, *18*, No. 100270.
(28) Tang, X.; Zhang, Z.; Shi, H.; Wang, Y.; Yan, A.; Yang, F.; Xie, K. Synthesis and Characterization of Agarose-Bacterial Cellulose Hydrogel for Promoting Seed Germination by Improving Soil Water Retention. *Polym. Bull.* **2024**, *81* (9), 8215−8227.
(29) Čechmánková, J.; Skála, J.; Sedlařík, V.; Duřpeková, S.; Drbohlav, J.; Šalaková, A.; Vácha, R. The Synergic Effect of Whey-Based Hydrogel Amendment on Soil Water Holding Capacity and Availability of Nutrients for More Efficient Valorization of Dairy By-Products. *Sustainability (Switzerland)* **2021**, *13* (19), 10701.
(30) Blake, G. R. Bulk Density. In *Methods of Soil Analysis, Part 1: Physical and Mineralogical Properties, Including Statistics of Measurement and Sampling*; Wiley, 2015; vol *30*.
(31) Montesano, F. F.; Parente, A.; Santamaria, P.; Sannino, A.; Serio, F. Biodegradable Superabsorbent Hydrogel Increases Water Retention Properties of Growing Media and Plant Growth. *Agriculture and Agricultural Science Procedia* **2015**, *4*, 451−458.
(32) Priya, E.; Sarkar, S.; Maji, P. K. A Review on Slow-Release Fertilizer: Nutrient Release Mechanism and Agricultural Sustainability. *J. Environ. Chem. Eng.* **2024**, *12* (4), No. 113211.
(33) Guan, Y.; Yu, C.; Zang, Z.; Wan, X.; Naeem, A.; Zhang, R.; Zhu, W. Chitosan/Xanthan Gum-Based (Hydroxypropyl Methylcellulose-Co-2-Acrylamido-2-Methylpropane Sulfonic Acid) Interpenetrating Hydrogels for Controlled Release of Amorphous Solid Dispersion of Bioactive Constituents of Pueraria Lobatae. *Int. J. Biol. Macromol.* **2023**, *224*, 380−395.
(34) Pang, J. H.; Liu, X.; Wu, M.; Wu, Y. Y.; Zhang, X. M.; Sun, R. C. Fabrication and Characterization of Regenerated Cellulose Films Using Different Ionic Liquids. *J. Spectrosc.* **2014**, *2014*, No. 214057.
(35) Yan, J.; Liu, J.; Sun, Y.; Song, G.; Ding, D.; Fan, G.; Chai, B.; Wang, C.; Sun, L. Investigation on the Preparation of Rice Straw-Derived Cellulose Acetate and Its Spinnability for Electrospinning. *Polymers (Basel)* **2021**, *13* (20), 3463.
(36) Mohamed, R. R.; Fahim, M. E.; Soliman, S. M. A. Development of Hydrogel Based on Carboxymethyl Cellulose/Poly(4-Vinylpyridine) for Controlled Releasing of Fertilizers. *BMC Chem.* **2022**, *16* (1), 52.
(37) Trache, D.; Khimeche, K.; Mezroua, A.; Benziane, M. Physicochemical Properties of Microcrystalline Nitrocellulose from Alfa Grass Fibres and Its Thermal Stability. *J. Therm. Anal. Calorim.* **2016**, *124* (3), 1485−1496.
(38) Panaitescu, D. M.; Vuluga, D. M.; Paven, H.; Iorga, M. D.; Ghiurea, M.; Matasaru, I.; Nechita, P. Properties of Polymer Composites with Cellulose Microfibrils. *Mol. Cryst. Liq. Cryst.* **2008**, *484*, 452−464.
(39) Zhang, Y.; Sun, X.; Ye, Y.; Oguzlu, H.; Zhu, Y.; Zhu, J.; Le, K.; Yang, P.; Jiang, F. All-Cellulose Hydrogel with Ultrahigh Stretchability Exceeding 40000%. *Mater. Today* **2024**, *74*, 67−76.
(40) Zhang, W.; Mu, X.; Xu, Y.; Ma, G.; Lei, Z. Research Progress of Environment-Responsive Hydrogel Applications in Agriculture. *Express Polymer Letters.* **2024**, *18*, 193−213.
(41) Wang, Y.; Shaghaleh, H.; Hamoud, Y. A.; Zhang, S.; Li, P.; Xu, X.; Liu, H. Synthesis of a pH-Responsive Nano-Cellulose/Sodium Alginate/MOFs Hydrogel and Its Application in the Regulation of Water and N-Fertilizer. *Int. J. Biol. Macromol.* **2021**, *187*, 262−271.
(42) Patra, S. K.; Poddar, R.; Brestic, M.; Acharjee, P. U.; Bhattacharya, P.; Sengupta, S.; Pal, P.; Bam, N.; Biswas, B.; Barek, V.; Ondrisik, P.; Skalicky, M.; Hossain, A. Prospects of Hydrogels in Agriculture for Enhancing Crop and Water Productivity under Water Deficit Condition. *Int. J. Polym. Sci.* **2022**, *2022* (1), No. 4914836.
(43) Lodhi, R. S.; Das, S.; Das, P. Recent Advances in Polymer Hydrogels for Agricultural Applications. *In Novel Polymeric Materials for Environmental Applications* **2023**, 109−177.
(44) Tariq, Z.; Iqbal, D. N.; Rizwan, M.; Ahmad, M.; Faheem, M.; Ahmed, M. Significance of Biopolymer-Based Hydrogels and Their Applications in Agriculture: A Review in Perspective of Synthesis and Their Degree of Swelling for Water Holding. *RSC Advances.* **2023**, *13* (35), 24731−24754.
(45) Qin, C. C.; Abdalkarim, S. Y. H.; Zhou, Y.; Yu, H. Y.; He, X. Ultrahigh Water-Retention Cellulose Hydrogels as Soil Amendments


https://doi.org/10.1021/acsagscitech.5c00092
*ACS Agric. Sci. Technol.* XXXX, XXX, XXX−XXX

for Early Seed Germination under Harsh Conditions. *J. Clean Prod.* **2022**, *370*, No. 133602.
(46) Malik, R.; Khatri, K.; Saxena, R.; Warkar, S. G. Fabrication of Carboxymethyl Tamarind Kernel Gum-Based Hydrogel and Its Applicability in Different Types of Soils for Agronomy. *Int. J. Biol. Macromol.* **2024**, *280*, No. 135616.
(47) Panagos, P.; De Rosa, D.; Liakos, L.; Labouyrie, M.; Borrelli, P.; Ballabio, C. Soil Bulk Density Assessment in Europe. *Agric. Ecosyst. Environ.* **2024**, *364*, No. 108907.
(48) Tang, S.; Zhang, L.; Tong, Z.; Wu, Z.; Wang, H.; Zhan, P.; Shao, L.; Qing, Y.; Wu, Y.; Liu, J. Encapsulated Lignin-Based Slow-Release Manganese Fertilizer with Reduced Cadmium Accumulation in Rice (Oryza Sativa L.). *Int. J. Biol. Macromol.* **2024**, *262*, No. 130019.
(49) Mansouri, H.; Ait Said, H.; Noukrati, H.; Oukarroum, A.; Ben youcef, H.; Perreault, F. Advances in Controlled Release Fertilizers: Cost-Effective Coating Techniques and Smart Stimuli-Responsive Hydrogels. *Advanced Sustainable Systems.* **2023**, 7 (9), 2300149.
(50) Fu, J.; Yap, J. X.; Leo, C. P.; Chang, C. K.; Show, P. L. Polysaccharide Hydrogels for Controlling the Nutrient Release. *Separation and Purification Reviews.* **2024**, *53* (3), 276−288.


https://doi.org/10.1021/acsagscitech.5c00092
*ACS Agric. Sci. Technol.* XXXX, XXX, XXX−XXX

JMEPEG
https://doi.org/10.1007/s11665-025-11379-z

©ASM International
1059-9495/$19.00

ORIGINAL RESEARCH ARTICLE

# Consequence of Ultrasonic Vibration-Assisted Cold Metal Transfer Welding Parameters on Metallurgical and Mechanical Properties of Ferritic Stainless-Steel-Welded Joints

*Alka Sharma, Reeta Wattal, and M.S. Niranjan*

Submitted: 5 March 2025 / Revised: 11 April 2025 / Accepted: 13 April 2025

**This study investigated the impact of cold metal transfer and ultrasonic vibration cold metal transfer welding processes on AISI 409L ferritic stainless-steel joints using ER 308L filler metal. The use of ultrasonic vibrations significantly refined the primary ferrite grains, reducing the ferrite content in the weld metal by approximately 45% and achieving a more uniform distribution of alloying elements. The maximum microhardness value of 305.67 HV was observed at the weld center at 9 mm/sec, while ultrasonic vibrations increased peak hardness to 325.32 HV at a 25 $\mu$m amplitude. The softened coarse heat-affected zone exhibited a microhardness of 210-235 HV, which was nearly equivalent to the base metal microhardness. The ultimate tensile strength, % strain, and hardness of the base metal AISI 409L were 565.85 MPa, 23.57%, and 197 HV, respectively, after welding with ultrasonic vibration, the maximum tensile strength of 574.6 MPa and a joint efficiency of 101.54% were observed compared to cold metal transfer-welded joint. This improvement is attributed to the refined grain structure, which enhances strain distribution and delays fracture onset. Despite the benefits of increased welding speed up to 9 mm/sec in the welded joints, these still exhibited lower tensile properties due to larger dendrites and increased interdendritic spacing. It led to a reduced strain of 21.19% and a higher likelihood of premature failure. The novelty of ultrasonic vibration-assisted cold metal transfer welding to ferritic stainless-steel AISI 409L, particularly for use in aerospace, pressure vessels, and automotive exhaust systems.**

**Keywords** austenitic filler metal, cold metal transfer, grain refinement, percentage dilution

## 1. Introduction

Ferritic stainless steel (FSS) offers outstanding corrosion resistance, ductility, and impact resistance. The use of this material has expanded across various fields. The cost of stainless steel is heavily influenced by its alloy content. FSS, composed of a body-centered-cubic (BCC) structure, commonly does not contain nickel (Ni) and 11% to 30% chromium (Ref 1). It excels in less severe conditions, such as automotive exhaust systems and architectural finishes (Ref 2). The primary concern with this grade is the loss of impact toughness and ductility due to grain coarsening in the fusion zone (FZ) and HAZ (Ref 3). The grain hardening problem can be overcome by welding FSS with high speed at low heat input (HI) (Ref 4). Cold metal transfer (CMT) is an advanced form of the gas metal arc welding (GMAW) technique. In contemporary times, the automobile sector is successively moving from GMAW to CMT for joining thin sheets ($<5$ mm) due to minimal heat generation during the joining process. CMT presents a precise approach to material transfer and minimal HI by combining filler metal (FM) movement with the digital process control system. It controls wire retraction movement in short circuit mode, which results in droplet detachment from the FM (Ref 5). Low HI and controllable short circuits are two essential features of CMT. Due to low HI, minimum welding current is maintained during short circuits. (Ref 6). The CMT welding technique is efficient in joining fragile sheets without distortion or spatter-free and with minimum residual stresses of different materials (Ref 7). Wang et al. (Ref 8) discovered that the ultrasonic vibration-assisted (UVA) welding technique has become more popular due to improved weld structure which results in reduced porosity and increased penetration. Jose et al. (Ref 9) stated that UVA welding is a successful replacement for post-weld heat treatments of joints, which improves the mechanical properties and minimizes the residual stresses. Kumar et al. (Ref 10) reviewed different techniques of ultrasonic probes, from the side of the work material, the top of the work material, to the FM, and directly to the molten pool positions during welding for best results. He observed that due

This invited article is part of a special topical issue of the *Journal of Materials Engineering and Performance* on Advances in Dissimilar Welding. The issue was organized by Dr. Chandan Pandey, Indian Institute of Technology, Jodhpur; Dr. Sachin Sirohi, SRM Institute of Science and Technology Delhi, NCR Campus; Prof. Dariusz Fydrych, Gdańsk University of Technology; and Dr. Shailesh Mani Pandey, National Institute of Technology, Patna.

**Alka Sharma, Reeta Wattal,** and **M.S. Niranjan**, Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India. Contact e-mail: sharma.alka1001@gmail.com.



Published online: 09 May 2025

to UV, mechanical energy is converted into heat, which helps in fusion welding. Watanabe et al. (Ref 11) investigated ultrasonic vibrations (UV) directly on a molten pool through filler wire and observed that tensile and elongation of weld metal are increased due to equiaxed grains formation for FSS. Dai (Ref 12) conducted UV on the upper surface of AA7075 for TIG welding, improved weld penetration, and enhanced grain growth achieved in comparison with without vibration process. Morisada et al. (Ref 13) performed GTA welding using UV and investigated the consequences of vibrations on porosity, tensile strength, and grain structure. Refined grain structure and reduced porosity lead to outstanding tensile strength. Abdullah et al. (Ref 14) studied welded joints by ultrasonic peening on 304 SS using the TIG welding method. Improved fatigue strength, grain structures, and reduced tensile residual stresses observed with UV. Zhao et al. (Ref 15) introduced a tungsten needle with ultrasonic vibration from the front side of the weld pool, leading to grain refinement and enhanced tensile strength in the weldment.

Dissimilar-welded Inconel 718 with SS 304L using GTAW and nickel-based filler IN 82 was investigated. Optical and FESEM analyses revealed distinct microstructures in the base metals and weld zones, including the presence of carbides in interdendritic regions and unmixed zones. The IN 82 weld zone was identified as the weakest area at room temperature, while high-temperature tests showed improved tensile strength. Residual stress measurements indicated that the highest tensile residual stress was located 3 mm from the weld surface (Ref 16). Inconel 718 and SS304L using ERNiCrCoMo-1 and ASS 304L revealed that the Ni-based fillers prevented weld solidification cracking, while the austenitic filler did not. Vickers microhardness peaked at 271 HV for ERNiFeCr-2 and dropped to 156 HV for ASS 304L. The ultimate tensile strength reached a maximum of 649 MPa for ERNiCrCoMo-1 and a minimum of 404 MPa for ASS 304L. Charpy toughness values were 70 J for ERNiCrCoMo-1, 56 J for ERNiFeCr-2, and 32 J for ASS 304L, with the highest residual stress found 2 mm below the top weld face (Ref 17). The A-TIG-welded joint between P91 steel and 800HT was investigated, focusing on microstructure and mechanical properties influenced by activating flux composition. Mechanical tests revealed excellent properties: ultimate tensile strength of 585 $\pm$ 5 MPa, elongation of 38 $\pm$ 2%, impact toughness of 96 $\pm$ 5 J, and maximum microhardness of 345 $\pm$ 5 HV was observed (Ref 18). The joint of Inconel 617 and P92 steel using an Inconel 82 buttering layer revealed significant improvements in mechanical properties. The ultimate tensile strength was 620 $\pm$ 4 MPa, with elongation of 19 $\pm$ 4%. The Charpy impact toughness values were 138 $\pm$ 6 J for the buttering layer and 115 $\pm$ 4 J for the filler weld, demonstrating the benefits of the buttering layer in enhancing joint integrity and performance (Ref 19).

The phase composition of the weld zone remained consistent throughout the process. Fig. 1 presents an optical micrograph of different heat-affected zone (HAZ) regions in their as-welded state, illustrating the temperature ranges during the welding (Ref 20). The HAZ experiences substantial microstructural transformations during heating, including the dissolution of precipitates and the transformation to austenite. These fine precipitates play a crucial role in enhancing hardness and tensile strength by impeding dislocation movement, thereby contributing significantly to the high-temperature tensile properties of the welded joints. The resulting microstructure is complex, featuring phases such as martensite and ferrite. Additionally, the HAZ is characterized by significant thermal gradients, which lead to variations in residual stress and hardness. These factors ultimately reduce the mechanical properties of the welded joint and increase its susceptibility to Type IV cracking due to creep deformation.

Yuan et al. (Ref 21) positioned the ultrasonic probe behind the arc to stir the weld pool. Effective grain refinement with improved mechanical properties was achieved. Tian et al. (Ref 22) used ultrasonic peening on Al alloy in the CMT and showed that pores count is decreased due to increased Al-Si lamellar spacing. Ultrasonic vibration on Al alloy using TIG welding was investigated, and heterogeneous nucleation in the molten pool was promoted. Columnar grains transformed into equiaxed grains around the fusion zone due to UV (Ref 23). Fattahi et al. (Ref 24) implemented nanoparticles with UV on Al using TIG welding and concluded that grain structure transformed from coarse dendrite to fine globular shape and mechanical properties improved. Ning and Cong (Ref 25) use the UVA laser-engineered net shaping (UVA-LENS) method on low-carbon steel for SS fabrication. Porosity, residual stress, tensile strength, and grain structures of the fabricated samples were excellently improved due to the effect of vibrations compared to those without vibrations. Chen et al. (Ref 26) comparatively investigated the pulsed U GTAW (PU GTAW) and U GTAW on low-carbon steel. Better grain refinement was obtained in PU GTAW. In the past few years, the UVA process has accumulated more interest due to refining the grain structure of welded parts and enhanced consistency between the weld and substrate (Ref 27). Xie et al. fabricated an Al-Zn-Mg weld joint and observed UVA GMA displayed larger HAZ and fusion zone along with enhanced corrosion resistance of welding zone (WZ) in comparison with the traditional GMA (Ref 28). Chen et al. (Ref 29) presented UV to pulsed TIG and observed the superior microstructural enhancement and grain refinement in weld beads compared to conventional TIG. Fang et al. (Ref 30) performed a the-squeeze casting (RSC) UVA technique to construct Mg-Zn-Y alloy. Their findings demonstrated that employing frequency and amplitude high to UV had a notable impact on enhancing metallurgical properties of the alloy. Koli et al.(Ref 31) stated that nitrogen addition significantly enhances mechanical properties, with good tensile strength, microhardness, and compressive residual stress achieved at higher heat inputs, although with decreased ductility due to martensite development and heat accumulation. Most researchers have focused on welding SS409L stainless steel using techniques such as GMAW, gas tungsten arc welding (TIG), friction stir welding, and laser welding. However, there is a notable lack of published information on welding SS409L using CMT and ultrasonic vibration cold metal transfer (UCMT) techniques. An ultrasonic peening device was employed to introduce ultrasonic vibrations during CMT welding. The application of these vibrations enhanced reinforcement and weld depth while reducing the contact angle. The Al-Si eutectic lamellar spacing increased, and the pore count decreased. Fine equiaxed grains (450 nm) formed near the eutectic structures were observed due to grain refinement induced by acoustic streaming and cavitation, resulting in improved wear resistance and mechanical properties (Ref 32). UCMT with three welding currents (85, 90, and 95A) and two ultrasonic vibrational amplitudes (50 and 99 $\mu$m) was used for the analysis of the AISI304 weldments. The UCMT-welded joint showed significant enhancement in tensile strength and hardness values due to grain refinement from ultrasonic

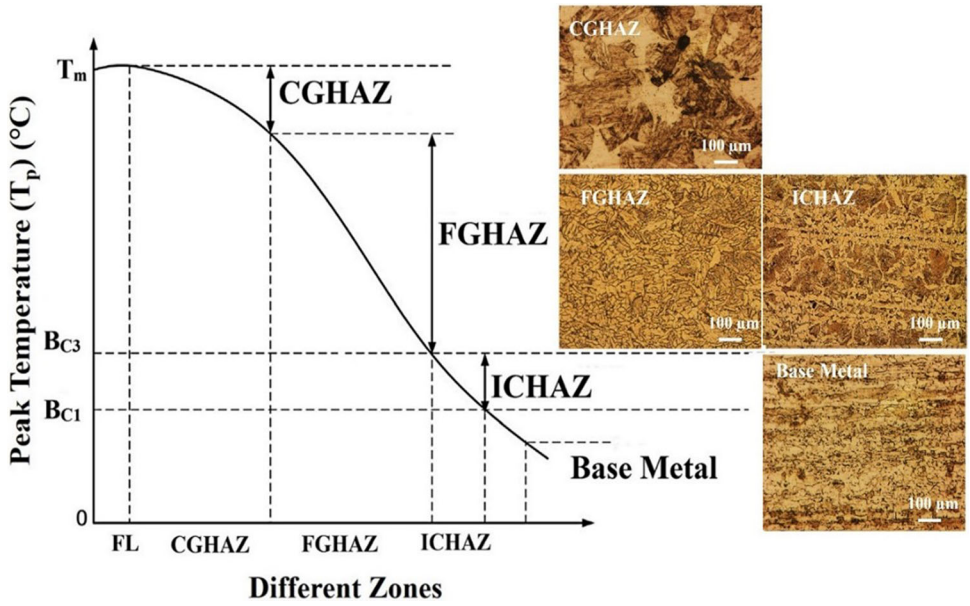



**Fig. 1** Impact of peak temperature on different regions of the HAZ (Ref 20)

**Table 1 Chemical configuration (wt.%) of base material and filler metal**

| Material | C | Mo | Mn | Si | S | P | Cr | Ni | N | Nb | Fe |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ER 308L | 0.029 | 0.073 | 1.42 | 0.42 | 0.012 | 0.022 | 19.15 | 10.03 | 0.01 | 0.059 | Balance |
| SS 409L | 0.017 | 0.006 | 0.840 | 0.560 | 0.006 | 0.024 | 11.240 | 0.407 | 0.021 | 0.040 | Balance |

vibrations. With a 99 $\mu$m vibration amplitude and 95A welding current, the UCMT joint achieved 100% weld joint efficiency, along with reduced residual stresses (Ref 33). A tungsten needle with ultrasonic vibration was inserted into the weld pool to stir the molten metal, refining grains and enhancing the tensile properties of Ti-6Al-4 V weldments. The ultrasonic vibration effectively inhibited columnar crystal growth and improved tensile strength, with significant effects when the needle was placed closer to the rear of the pool. The vibration did not alter the phase composition but affected grain orientation (Ref 15). The novelty of UCMT welding to ferritic stainless-steel AISI 409L, particularly for use in aerospace, pressure vessels, and automotive exhaust systems. By introducing ultrasonic vibrations during the welding process, the mechanical properties and metallurgical properties in the HAZ region were improved.

In this work, the impact of CMT and UCMT welding processes on AISI 409L ferritic stainless-steel joints using ER 308L filler metal was investigated and observed the mechanical i.e., tensile strength, microhardness, and metallurgical properties of weldments. To assess the effects of ultrasonic vibrations on weld quality and mechanical properties, overall joint performance was compared to CMT-welded joints.

## 2. Experimental Technique

### 2.1 Materials and Methods

In this work, AISI 409L of 3 mm thickness stainless steel was taken as substrate for investigation. Sample dimensions of 100 mm x 60 mm x 3 mm were prepared and fabricated in a butt joint with no spacing. The chemical compositions of 409L SS and 308 L filler metal are detailed in Table 1. The mechanical properties of base metal (BM) are revealed in Table 2. The BM was mechanically cleaned using emery paper before welding.

CMT and UCMT welding processes were used for this work. CMT is an advanced form of GMAW that offers precise heat control during welding. Its working principle revolves around a controlled short-circuiting process, where the filler wire retracts after each short circuit, minimizing heat input (Ref 34). This results in lower spatter, reduced distortion, and better control over the weld pool. CMT is highly effective for welding thin materials and dissimilar metals due to its minimal heat generation, which helps prevent burn-through and excessive warping (Ref 35). The UCMT was associated with a probe, which is linked to an ultrasonic welding machine generator box maintaining a vibration amplitude ranging from 20, 25, and 30 $\mu$m and a constant frequency of 20 kHz. The vibration amplitude measurement during welding using an ultrasonic welding machine is typically carried out through the machine's generator box, which controls and monitors the ultrasonic vibrations. The generator box is equipped with sensors and a control system that accurately measures the frequency and amplitude of the ultrasonic vibrations being transmitted to the welding tool and, consequently, to the workpiece.

The vibration amplitude is detected using piezoelectric transducers integrated within the ultrasonic welding system. These transducers convert electrical energy from the generator into mechanical vibrations. As the ultrasonic waves propagate through the welding tool to the material, the generator

**Table 2 Mechanical property of the SS 409L**

| Ultimate tensile strength, MPa | Yield strength, MPa | Hardness, HV | Strain, % |
|---|---|---|---|
| 565.85 | 436.48 | 197 | 23.57 |

**Table 3 Process parameters for CMT welding**

| Specimen No. | WFR, m/min | WS, mm/sec | GFR, *l/min* |
|---|---|---|---|
| Specimen 1 | 5 | 7 | 17 |
| Specimen 2 | | 8 | |
| Specimen 3 | | 9 | |
| Specimen 4 | | 10 | |

continuously monitors and displays the amplitude in real-time. Adjustments can be made directly through the generator box to ensure optimal vibration amplitude, critical for achieving proper weld penetration and grain refinement. This precise monitoring helps maintain the consistency and quality of the weld. A UCMT butt joint is created for the purpose of comparing it with a CMT butt joint. The contact-to-workpiece distance (CTWD) of 15 mm was established following extensive trial experiments. Experiments were performed using 80% argon and 20% $CO_2$ at a constant gas flow rate (GFR) of 17 L/min. Welding parameters of CMT and UCMT are shown in Table 3 and Table 4.

A TPS 400i CMT machine with an ultrasonic attachment was used to experiment. The ultrasonic probe makes contact with the width side of the workpiece to facilitate longitudinal vibrations, demonstrated in Fig. 2. The welded plate and tensile test specimen with their dimensions as per ASTM E8 standard is revealed in Fig. 3a-c. The optical and SEM with EDS image of the SS409L is shown in Fig. 4. The hardness of the weldment was analyzed by a Vicker hardness machine (IS:1500(Pt.1):2019) at a constant load of 500 grams for 15 s of dwell time. Tensile test was performed using WAW 600D model UTM with test method ASTM E8/E8M:2022 at a ram speed of 1 mm/min. Microstructure inspection necessitates extracting from the transverse direction of the weldment. These specimens were polished using a conventional metallographic technique and then etched by using an etching agent composed of 30 ml HCl, 120 ml $H_2O$, and 10 g $FeCl_3$. The optical Stereo zoom microscope and test method ASM Vol-9 were used to conduct macroscopic and microscopic studies, and the fractured surfaces were investigated by scanning electron microscope (SEM) model No: ZEISS, EVO/18. Grain particle size measurements in the WM zone for both samples were conducted using ImageJ software.

## 3. Results and Discussion

### 3.1 Microstructure Analysis

The dilution ratio ($D_L$%) is determined to assess the chemical composition of WM. Dilution is defined as the contribution of the BM into the weld pool. The critical importance of dilution is that the final weld material's chemical composition becomes unique, forming an intermediate alloy distinct from both the BM and the FM (Ref 36). The dilution ratio ($D_L$%) in welding is calculated using the geometric properties of the WM, such as the top reinforcement area (AT), root reinforcement area (AR), and BM fusion area (AB). To obtain the total WM area (AW) is represented by equation (1). Fig. 5 shows the macrostrcutre of the weld bead of CMT- and UCMT-welded joints. The $D_L$% can be determined using the following equation (2) (Ref 37).

$$A_W = A_T + A_B + A_R \quad \text{(Eq 1)}$$

$$D_L\% = A_B \div A_W \times 100 \quad \text{(Eq 2)}$$

$D_L$% was utilized to calculate the chemical components of WM using equation (3), where $C_W$%, $C_B$%, and $C_F$% represent the elements's mass fraction in the WM, BM, and FM, respectively (Ref 37).

$$C_W\% = C_B\% \times D_L\% + C_F\% \times (1 - D_L\%) \quad \text{(Eq 3)}$$

Table 5 presents the calculated chemical composition of WM, Chromium equivalent ($Cr_{eq}$), Nickel equivalent ($Ni_{eq}$), and $Cr_{eq}/Ni_{eq}$ values for both samples using equations (4) & (5) (Ref 37).

$$Cr_{eq} = Mo\% + Cr\% + 0.5(Nb\%) + 1.5(Si\%) \quad \text{(Eq 4)}$$

$$Ni_{eq} = 0.5(Mn\%) + 30\,(C\%) + Ni\% \quad \text{(Eq 5)}$$

Various constitution diagrams, like the Schaeffler and De Long accompanied by specific equations, were generated to anticipate the phases that can generate in SS during welding. These tools serve to figure out microstructural changes, providing valuable insight to improve welding procedures (Ref 37). Mostly constitution diagrams are used $Cr_{eq}$ and $Ni_{eq}$ to predict the phases in SS. In this investigation, the Schaeffler diagram is employed for phase prediction using equation (4) & (5). The welded joints, consisting of a ferrtic BM and austenitic FM, result in a mixed microstructure in WM. The $Cr_{eq}/Ni_{eq}$ greatly affects the development of ferrite ($\alpha$). When $Cr_{eq}/Ni_{eq}$ is below 1.35, solidification leads to formation of austenite, whereas $Cr_{eq}/Ni_{eq}$ is above 1.35 solidification leads to formation of ferrite. Cr acts as stabilizing element for ferrite while Ni favors the stabilization of austenite phase. Beyond Cr other elements such as manganese and silicon contribute to ferrite stabilization, and elements like cobalt, molybdenum, and copper play roles in stabilizing austenite. That's why $Cr_{eq}$ and $Ni_{eq}$ has been calculated. it was obsereved that, CMT and UCMT both have $(Cr/Ni)_{eq}$ value greater than 1.35 so austenitic-ferritic modes will appear in all weld metals. This solidification mode can be explained using a pseudobinary phase diagram as shown in Fig. 6, in which all WM belong to the two phases ($\delta + \gamma$) solidification (Ref 38). As the $(Cr/Ni)_{eq}$ increases, the stability of the austenite phase decreases, leading to the conversion of unstabilized austenite into martensite (38). As UCMT exhibited the highest Cr-Ni ratio, it is associated

**Table 4 Process parameters for UCMT welding**

| S. no | WFR, m/min | WS, mm/sec | GFR , *l/min* | Amplitude, $\mu$m | Frequency, kHz |
|---|---|---|---|---|---|
| 1 | 5 | 9 | 17 | 20 | 20 |
| 2 | | | | 25 | |
| 3 | | | | 30 | |

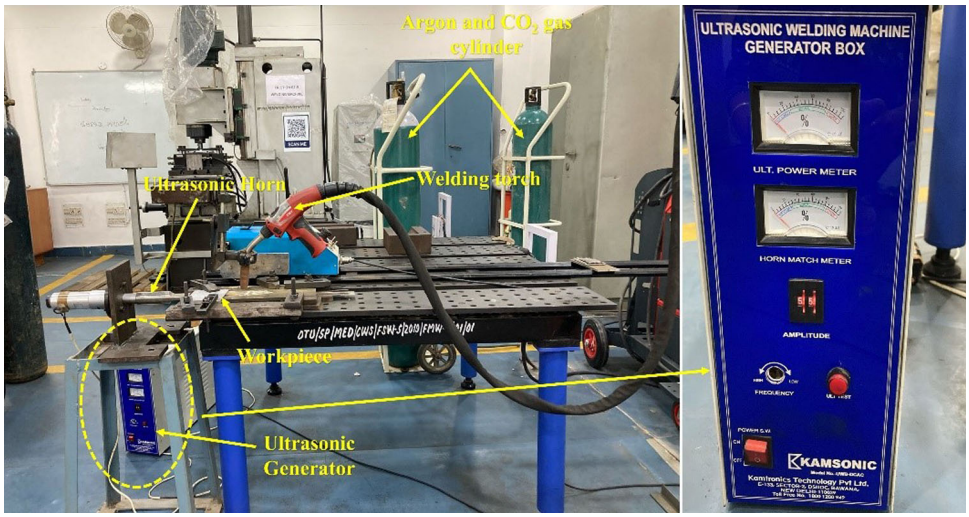



**Fig. 2** (a) Experimental setup of CMT process and (b) ultrasonic welding machine generator box

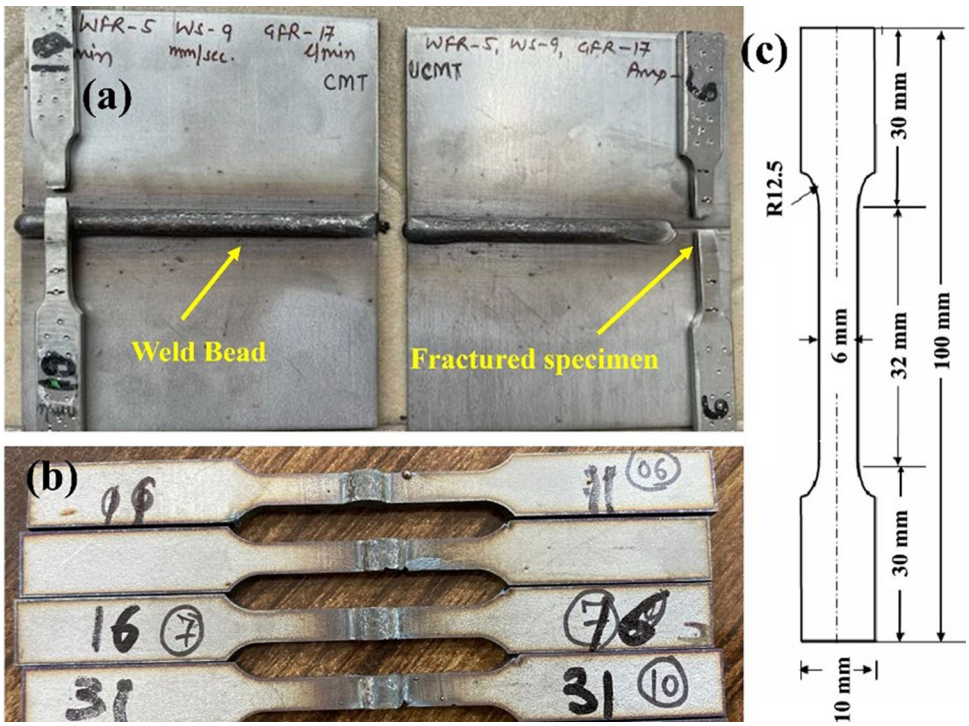



**Fig. 3** (a) Fabricated welded plate, (b) tensile specimen, and (c) tensile specimen dimension as per ASTM E8 standard

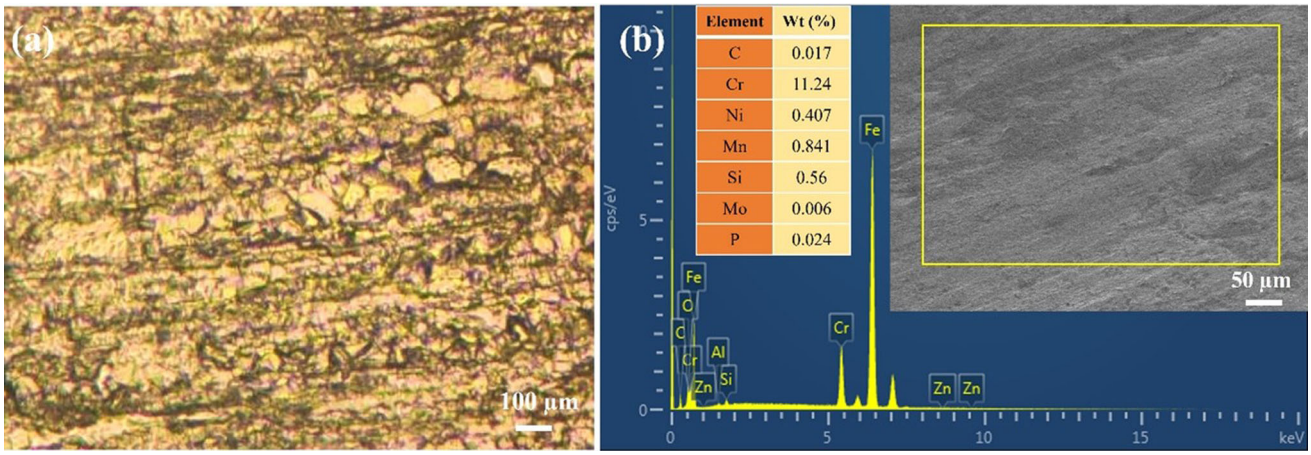



**Fig. 4** (a) Optical image of base metal and (b) SEM image with EDS of base metal

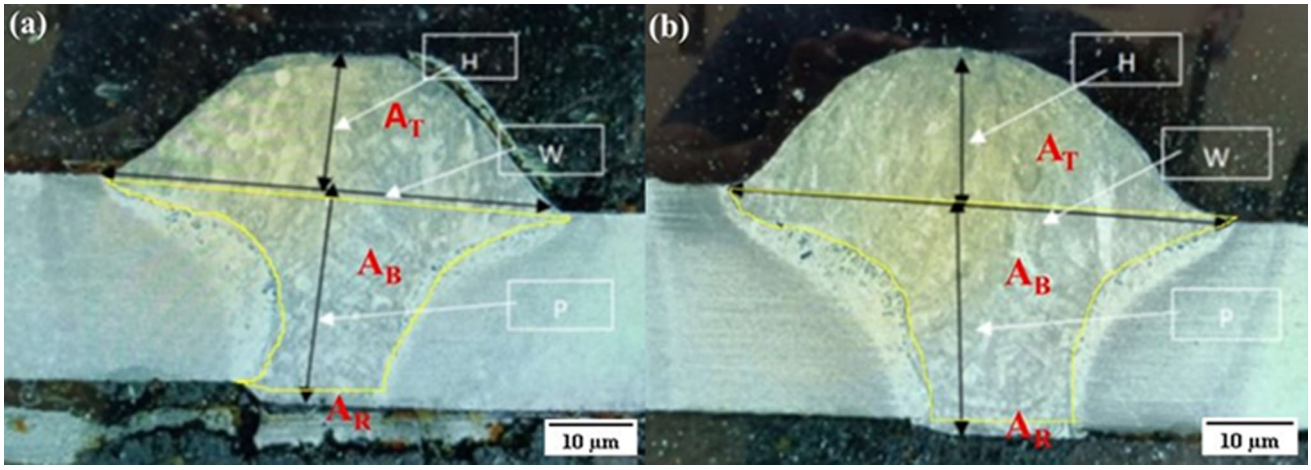



**Fig. 5** Macroimages exhibit differences in WM geometry between (a) CMT and (b) UCMT samples

**Table 5 Chemical composition of weld metals**

| Specimen | Cr | Mo | Si | Nb | Ni | C | Mn | $Cr_{eq}$ | $Ni_{eq}$ | $Cr_{eq}/Ni_{eq}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| CMT | 15.48 | 0.0419 | 0.4890 | 0.050 | 5.57 | 0.023 | 1.151 | 16.280 | 6.835 | 2.381 |
| UCMT | 15.42 | 0.0414 | 0.4859 | 0.050 | 5.49 | 0.023 | 1.146 | 16.219 | 6.753 | 2.401 |

with more martensite phase as shown in Fig. 7. Due to austenitic FM with ferritic BM, it resulted in the development of mixed phases comprising ferrite, austenite, and martensite in WM.

The stability and transformation of phases during cooling from liquid to atmospheric temperature can be explained using equilibrium phase diagrams. Lippold and Savage (1979) proposed using a pseudobinary phase diagram to describe the phase transformation in the welding of ternary Fe-Cr-Ni systems (Ref 39). As per values of $Cr_{eq}/Ni_{eq}$ ratio, the transformation sequence of WM follows the ferrite-austenite mode:

$$L \rightarrow L + F \rightarrow L + F + (F + A) \rightarrow F + A$$

F, L, and A represent ferrite, liquid, and austenite respectively. It was observed that when $(Cr/Ni)_{eq}$ is more than 1.35 so, the first phase formed from L is F. After solidification, austenite forms at ferrite grain boundaries (FGB) through peritectic and eutectic reactions. As the weld metal cools into the ferrite-austenite biphasic zone, ferrite gradually transforms into austenite in a diffusion-controlled process. During this transition, ferritic elements accumulate in the remaining ferrite, while austenitic elements decrease. Eventually, as diffusion slows, ferrite stabilizes at a lower temperature. Both weld metals showed a Cr/Ni ratio above 1.35, leading to ferrite being the primary phase during solidification. As the Cr/Ni ratio increases, the stability of austenite decreases, leading to the conversion of unstabilized austenite into martensite (Ref 38).

Therefore, UCMT samples, characterized by a high Cr/Ni ratio, exhibit a greater presence of martensite phase.

The weld metal of CMT- and UCMT-welded joint is revealed in Fig. 7a, c, and the HAZ microstructure, found between the weld metal (WM) and the BM, is depicted in Fig. 7(b) and (d) for CMT and UCMT, respectively. The HAZ of the CMT sample revealed the presence of long dendrites of significant size and columnar-shaped grains, which is shown in Fig. 7(b). Wei (Ref 40) noted that vibration accelerates the dispersion of molten weld material evenly, leading to uniform adjustments in weld pool temperature. Consequently, this process results in the formation of equiaxed grains. However, when vibration was introduced, smaller dendrites and equiaxed grains were observed instead. The ferrite grains expanded notably compared to those in the BM due to exposure to high temperatures. At sufficiently elevated temperatures, austenite transformed, forming a seamless layer of intergranular martensite along the FGB, adopting a network-like arrangement. A few intragranular martensite formations appeared as individual grains or strips nestled within the ferrite grains. Granular precipitates were detected within the ferrite grains. When exposed to elevated temperatures, the ferrite becomes supersaturated with C and N, facilitating precipitate formation. High cooling rates promote precipitate formation within the grains (Ref 39). The absence of noticeable precipitation in the HAZ can be attributed to the low chromium (Cr) content of this alloy and the inclusion of titanium (Ti). Ti serves to bind carbon and nitrogen, thereby impeding the development of M23C6 carbides and carbonitrides. It prevents their precipitation within the HAZ, ensuring its clarity from such formations.

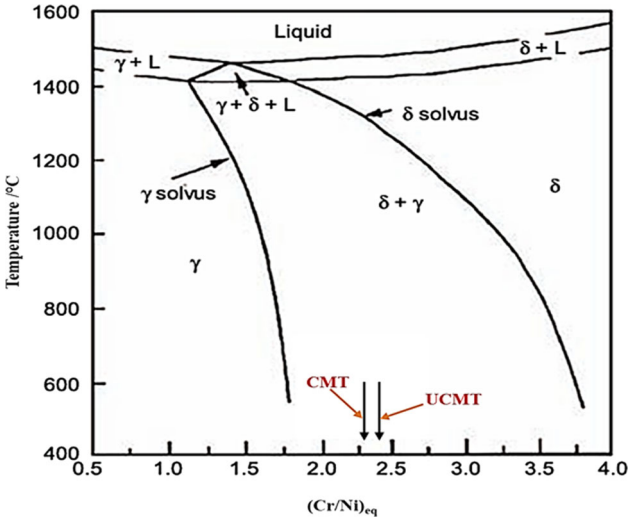



**Fig. 6** A pseudobinary diagram for predicting equilibirium microstructure in stainless steel (Ref 39)

The CMT welding of SS409L, particularly when enhanced with ultrasonic vibration, provides a unique microstructure. When combined with ER308L filler, the microstructure within different regions of the HAZ, FGHAZ, CGHAZ, and IGHAZ undergoes significant changes. These changes can be directly linked to the thermal cycles experienced during welding, and they are critical in assessing the overall performance of the weld.

In the case of CMT welding of SS409L with ER308L filler, the FGHAZ typically exhibits a refined grain structure due to the controlled heat input. The optical microstructure in this region is characterized by small, equiaxed grains. The smaller grain size contributes to enhanced mechanical properties, such as increased strength and toughness. This is because smaller

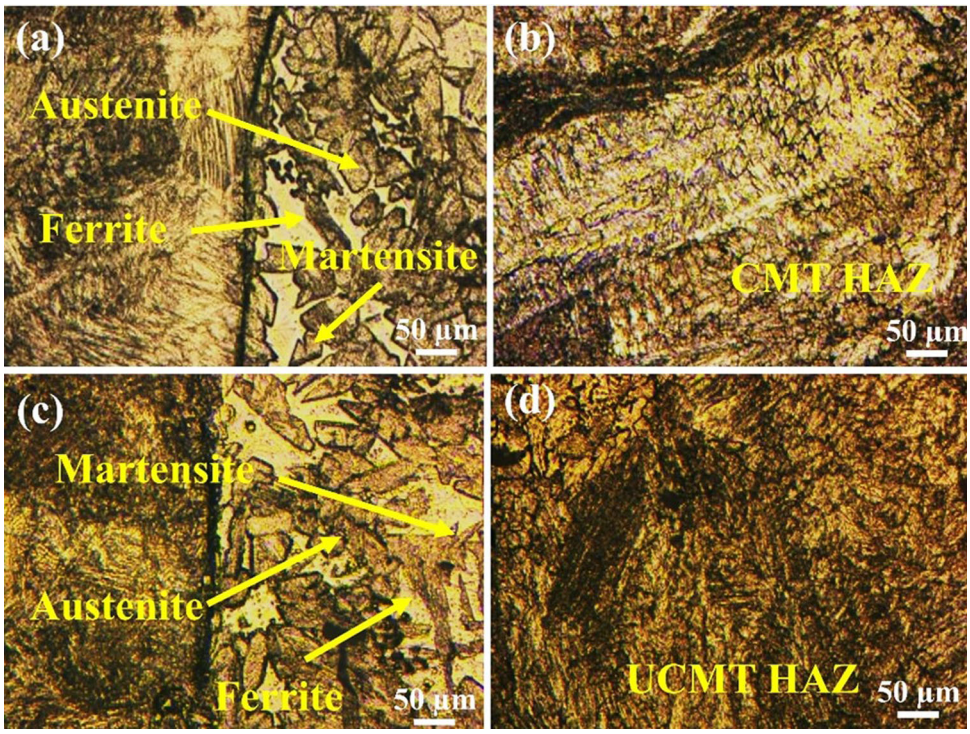



**Fig. 7** Microstructure of CMT samples: (a) weld metal of CMT and (b) CMT_HAZ and (c) weld metal of UCMT and (d) UCMT_HAZ

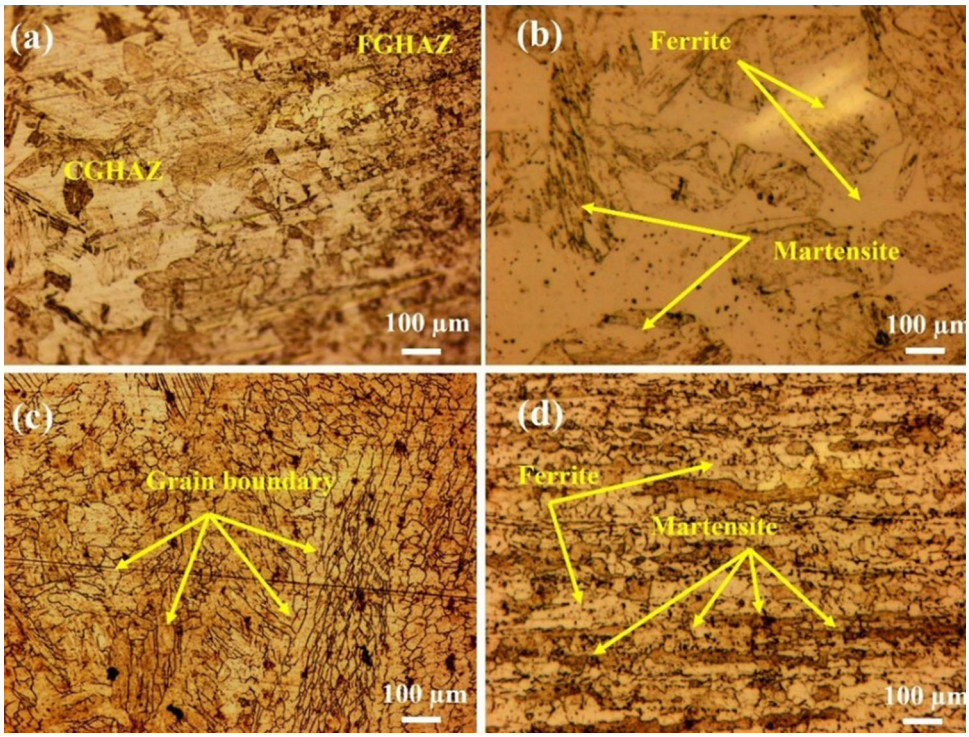



**Fig. 8** Optical micrograph of CMT-welded joint of different HAZ zones: (a) welded region, (b) CGHAZ, (c) FGHAZ, and (d) ICHAZ

grains provide more grain boundary area to hinder dislocation movement, which improves the material's resistance to deformation. In CMT welding, the FGHAZ generally contains a mixture of martensite and ferrite. The ferrite phase forms due to the high chromium content in SS409L, which stabilizes ferrite at elevated temperatures. Martensite forms as the steel cools rapidly after welding, particularly from the austenite phase (Ref 41). Optical micrograph of CMT-welded joint of different HAZ zone was revealed in Fig. 8(a-d). The presence of martensite contributes to the overall hardness of the FGHAZ, although excessive martensite can lead to brittleness. The CGHAZ experiences the highest temperatures during welding, typically just below the melting point of the material. As a result, the grains in this region grow significantly larger, which can be detrimental to the mechanical properties as demonstrated in Fig. 8b. The optical microstructure of the CGHAZ in CMT welding of SS409L is dominated by large, columnar grains of ferrite, and martensite as shown in Fig. 8. These coarse grains are less effective at impeding dislocation movement, leading to reduced strength and toughness. In SS409L welded using CMT with ER308L filler, the CGHAZ may also exhibit an uneven austenite grain structure with minimal precipitation of carbides or other phases (Ref 42). This lack of precipitates can further weaken the material, making the CGHAZ more prone to cracking, especially under high stress or in corrosive environments. The coarsening of grains in the CGHAZ can also lead to a reduction in hardness, as larger grains are typically softer and more ductile. While UCMT, the grain growth in the CGHAZ can be mitigated. Grain coarsening in the CGHAZ adversely affects the mechanical properties of the weld by decreasing strength and toughness due to the formation of larger, less stable grains. Ultrasonic vibrations can mitigate this issue by promoting nucleation during solidification and disrupting the growth of larger grains. It leads to a refined microstructure, enhancing mechanical properties such as tensile strength and toughness, ultimately improving the performance and reliability of the welded joint. The FGHAZ is located immediately adjacent to the fusion line, where the metal experiences a moderate thermal cycle that leads to grain refinement as illustrated in Fig. 8c. High-frequency oscillations are introduced into the molten pool by ultrasonic vibrations, which causes mechanical agitation in the liquid metal. Larger grain structures are broken up by this agitation, which makes it possible for smaller grains to nucleate more uniformly. Ultrasonic vibrations effectively limit the size of the grains by interfering with the natural grain growth process in the CGHAZ, where high temperatures can cause coarse grains to grow (Ref 43). The vibrations improve heat distribution and accelerate cooling rates by improving the molten material's mixing. In the heat-affected zone, this leads to the formation of finer, more uniform grains, which directly improves mechanical qualities like strength and toughness. Additionally, by promoting the formation of equiaxed grains, ultrasonic vibrations lessen the propensity for columnar grain formation, which is frequently seen in the CGHAZ. A more balanced and dependable weld results from this microstructure refinement in the CGHAZ, which lowers the possibility of flaws like cracking and enhances the welded joint's overall performance. The IGHAZ is the region within the HAZ that experiences temperatures in the intercritical range, between the Bc1 and Bc3 temperatures. In this zone, the steel undergoes partial transformation, resulting in a dual-phase microstructure comprising both austenite and ferrite as shown in Fig. 8d. Upon cooling, the austenite may transform into martensite, while the ferrite remains stable. The optical microstructure of the IGHAZ in CMT welding of SS409L shows a mixture of ferrite, martensite, and possibly some retained austenite (Ref 44). The IGHAZ is particularly susceptible to variations in mechanical

properties due to its heterogeneous microstructure. The presence of both hard martensite and softer ferrite can lead to localized stress concentrations, which may decrease the strength of the weld. Furthermore, the IGHAZ may also experience the development of new austenite grains during the thermal cycle, leading to the generation of precipitates and refined dislocations. This alteration in the microstructure can further reduce the material's resistance to cracking.

UCMT welding significantly enhance the grain refinement process in the FGHAZ and CGHAZ of SS409L. The introduction of these vibrations increases energy input into the weld pool, promoting uniform temperature distribution and prolonging the liquid state of the material, which facilitates better mixing. This dynamic environment creates numerous nucleation sites due to the mechanical vibrations generating dislocations and defects, leading to the formation of finer grains during solidification. The oscillations disrupt the growth of larger grains, allowing for a more refined microstructure. By mitigating the effects of thermal gradients, the ultrasonic vibrations also promote uniform cooling rates, which are crucial for achieving finer grain structures. These vibrations enhance bonding between materials, reducing elemental segregation and resulting in a more homogeneous microstructure.

High-frequency ultrasonic waves impart mechanical energy, and ultrasonic vibration in UCMT welding significantly refines the ferrite grains in the weld metal, particularly in ferritic stainless steel AISI 409L. Rapid oscillations caused by the ultrasonic vibrations in the molten pool prevent large grains from growing naturally and promote the formation of smaller, more uniform grains. By encouraging improved nucleation and grain boundary migration, this effect inhibits the development of coarse columnar grains that are commonly observed in welds. It leads to a finer and more equiaxed grain structure in the weld metal, which directly enhances the mechanical properties. By improving heat distribution in the molten pool, ultrasonic vibrations help regulate cooling rates and stop harmful phases from forming. The UCMT welding further refines the grain structure in the FGHAZ. The vibrations promote dynamic recrystallization, leading to a more uniform and fine-grained microstructure. This refined microstructure not only enhances the mechanical properties but also reduces the likelihood of residual stress formation, which can otherwise lead to cracking. The vibrations introduce mechanical energy that disrupts the normal grain growth process, leading to a finer grain structure even in this high-temperature region. The refined grains in the CGHAZ improve the toughness and ductility of the welded joint, reducing the risk of failure. Additionally, the presence of ultrasonic vibrations can help in the redistribution of elements that might otherwise segregate, leading to a more homogeneous microstructure. In UCMT welding, the IGHAZ benefits from the enhanced grain refinement and homogenization of phases. The ultrasonic vibrations help to stabilize the austenite phase, reducing the extent of martensitic transformation upon cooling. This results in a tougher and more ductile IGHAZ, with a lower likelihood of cracking. Additionally, the refined microstructure in the IGHAZ helps to enhance the strength of the weldment, making it more resistant to failure. Due to the high temperature, the ferrite grain size increased considerably compared to that in the BM (Ref 45). When the temperature rose sufficiently, the austenite converted into intragranular martensite layer revealed in Fig. 9a-d , which formed a network pattern along the FBG. A small amount of intragranular martensite appeared as isolated grains within the ferrite grains and observed granular precipitates within the ferrite grains. According to Lippold and Kotecki (Ref 46), these precipitates are likely carbonitrides, nitrides, and carbides formed due to the supersaturation of nitrogen and carbon in the ferrite phase, caused by the rapid decrease in the solubility of these elements as the temperature dropped (Ref 47). The UCMT process, which combines the low heat input of CMT with ultrasonic vibrations, results in a refined microstructure within the WFZ and HAZ. The ultrasonic vibrations promote dynamic recrystallization, leading to smaller and more uniform grain sizes, which enhances the joint efficiency. The microstructure in the WFZ typically exhibits a mix of martensite and ferrite phases. The martensite, formed during rapid cooling, contributes to increased hardness but can also introduce brittleness. The high chromium content in SS409L stabilizes the ferrite phase during CMT welding by promoting the formation of austenite and minimizing the transformation to martensite. This stabilization enhances the weld's overall performance by improving corrosion resistance and maintaining ductility. As a result, the welded joint exhibits better mechanical properties, making it more suitable for applications requiring reliable performance under challenging conditions.

Fig. 10a-d revealed the SEM images of CGHAZ, and FGHAZ of CMT and UCMT weldment of SS409L. In the FGHAZ, the average grain size was smaller than the BM. In this area, the content of martensite was significantly lower, and instead of forming at the grain boundaries, martensite was dispersed. This change occurred because the fine-grain zone is farther from the weld center, resulting in a lower temperature compared to regions closer to the weld. The reduced temperature hindered the formation of high-temperature austenite, which in turn limited the transformation of austenite into martensite. Additionally, the lower temperature in this zone prevented the formation of intragranular precipitates within the grains, unlike what occurs in regions exposed to higher temperatures. This zone's microstructural characteristics reflect its distance from the weld center and the associated temperature gradient, influencing the phase transformations and precipitate formation. In CMT and UCMT welding, δ-ferrite formation and fusion line macrosegregation significantly influence the weld's microstructure and mechanical properties. δ-ferrite forms due to ferrite-stabilizing elements like chromium and molybdenum in stainless steels, stabilizing at high temperatures (Ref 48). It partially transforms into austenite during cooling, but some δ-ferrite may remain, enhancing weldability by reducing hot cracking. However, retained δ-ferrite can cause embrittlement. UCMT welding, with its ultrasonic vibrations, increases cooling rates and promotes grain refinement, reducing δ-ferrite content compared to CMT welds. Fusion line macrosegregation, caused by the uneven distribution of alloying elements like chromium and nickel, affects mechanical properties, corrosion resistance, and crack susceptibility. In CMT, slower solidification promotes greater macrosegregation. UCMT mitigates this through ultrasonic vibrations, which enhance element mixing and solidification uniformity, resulting in a more homogeneous weld with improved performance.

### 3.2 Microhardness Analysis

The Vickers hardness of CMT and UCMT welds was measured in the transverse direction of the welding, as shown in Fig. 11(a,b). The hardness values in the weld zones ranged

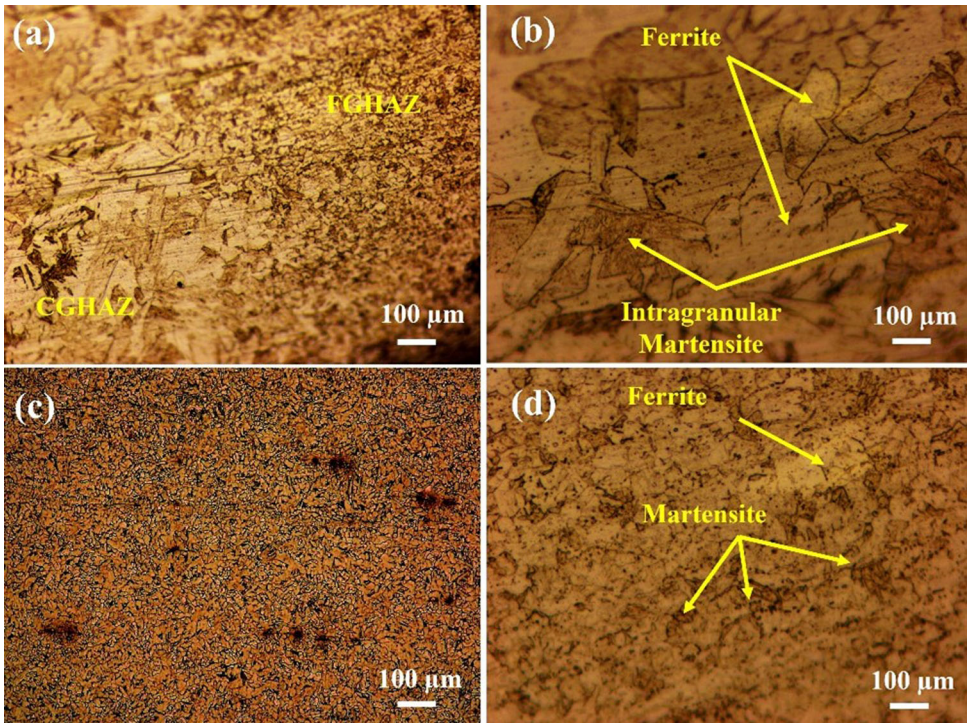



**Fig. 9** Optical micrograph of UCMT-welded joint of different HAZ zones: (a) welded region, (a) CGHAZ, (c) FGHAZ, and (d) ICHAZ

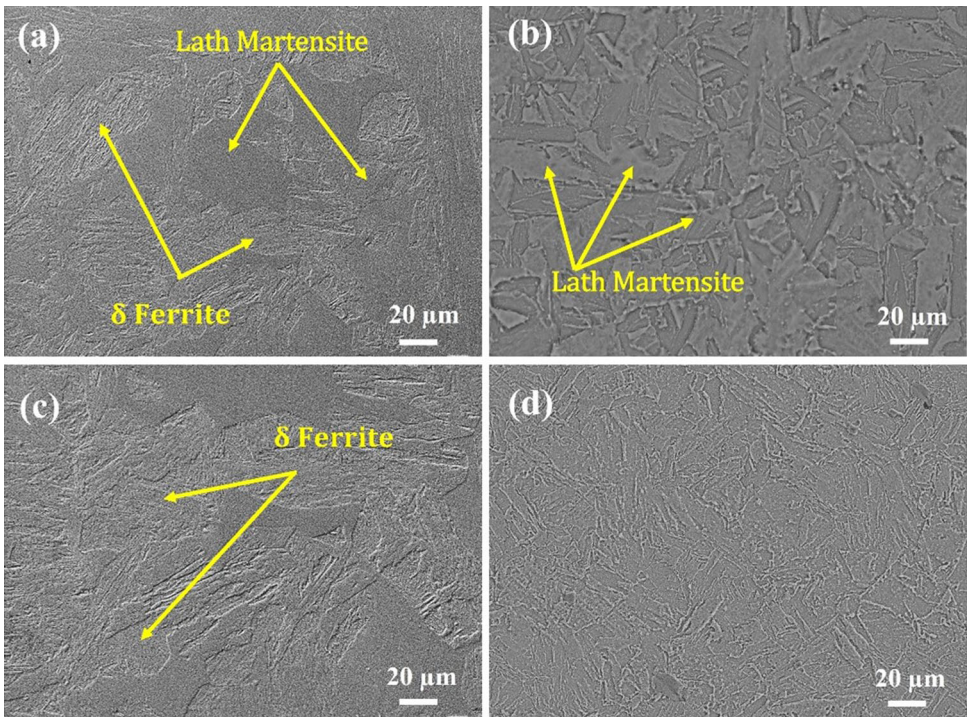



**Fig. 10** SEM images of welded joints of SS409L: (a) CGHAZ_CMT, (b) FGHAZ_CMT, (c) CGHAZ_UCMT, and (d) FGHAZ_UCMT

from 270-310 HV for CMT and 290-330 HV for UCMT-welded joints. The increased hardness in the CMT- and UCMT-welded zones is mainly due to the microstructural transform from a ferritic to a martensitic structure. Fig. 11(a,b) illustrates the hardness variation across the weld, from the fusion line to the weld center and from the top surface to the root face. In the CMT weld fusion zone, hardness was generally uniform, with a slightly lower value at the weld center. This uniformity is due to the consistent phase distribution throughout the fusion zone, while the slower solidification rate at the weld center caused greater grain growth and less martensitic transformation, resulting in reduced hardness. The maximum microhardness value of 305.67 HV was observed at the weld center in the CMT joint at 9 mm/sec, while the UCMT joint achieved a peak

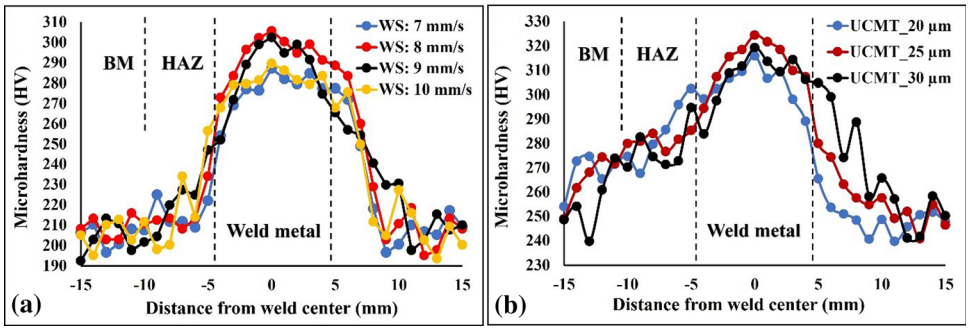



**Fig. 11** Hardness variaton from the weld center of CMT- and UCMT-welded joints at different paramaters

hardness of 325.32 HV at a 25 $\mu$m amplitude. The softened coarse heat-affected zone (SCHAZ) exhibited a microhardness of 210-235 HV, which was nearly equivalent to the base metal microhardness. This similarity in microhardness between the SCHAZ and BM is consistent with their comparable microstructures. The slight difference in microhardness values can be attributed to the presence of segregation bands, which Emygdio et al. (Ref 49) have identified as areas where a higher concentration of carbon is dissolved, leading to a localized increase in hardness. The ICHAZ underwent incomplete austenitization during the heating process, resulting in a predominantly ferrite and martensitic structure upon cooling. This microstructure is associated with the maximum hardness value observed in UCMT weldment. The high hardness in the ICHAZ caused by the coexistence of a soft matrix and a dispersed hard phase of martensite. Jang et al. (Ref 50) also noted that an increase in the concentration of carbide aggregates within the microstructure further elevates local microhardness, contributing to the significant hardness found in the ICHAZ. The FCHAZ exhibited high microhardness values. This is primarily due to the combination of fine austenite and martensitic structure. The presence of these microstructural features enhances the hardness of the FCHAZ. The CGHAZ displayed the largest austenite grain size, along with a coarse martensitic microstructure. Despite these characteristics, the CGHAZ showed a decrease in microhardness compared to the FCHAZ. This reduction in hardness is likely due to the coarsening of the microstructure, which generally leads to a softer material (Ref 51).

The UCMT weld zone showed maximum hardness near the center line, which decreased toward the HAZ region, as shown in Fig. 11(b). A significant variation in hardness along the weld thickness was observed, with hardness decreasing from the root to top surface. The hardness of the UCMT weld was primarily affected by the dilution of the BM in the weld pool and the cooling rate. The significant temperature difference between the weld pool and the BM, coupled with the larger area for heat dissipation, led to a shorter solidification time. UCMT demonstrates superior microhardness compared to CMT due to its finer grain size. It was observed that increased Cr/Ni content in the WM promotes more excellent martensite formation, reducing grain size and consequently enhancing hardness. The retained heat and auto-tempering effect significantly impact hardness in CMT and UCMT welds. CMT welds experience localized auto-tempering, softening the martensitic structures in the HAZ and reducing hardness. In contrast, UCMT welds benefit from more uniform temperature distribution due to ultrasonic vibrations, promoting grain refinement and minimizing heat retention. This results in a more consistent hardness profile across the weld, particularly in the WM and HAZ, with less softening than CMT welds. The base metal's dilution in the weld pool was most concentrated near the fusion line. Multiple thermal cycles during welding and microstructural changes in the weld fusion zone led to hardness variations in both CMT and UCMT welds. The higher hardness in the heat-affected zones (HAZs) compared to the base metal is due to grain refinement and martensitic transformation. The weld metal (WM) shows a 55% higher hardness, and the HAZ displays a 43% increase in hardness compared to the base metal using the CMT technique. In the case of UCMT, the hardness value for WM exhibits a 64.62% increase compared to the BM, while the HAZ displays 52% growth from the BM. The microhardness value of UCMT reached its peak, which was attributed to vibrations contributing to increased dilution. Higher Cr/Ni values were observed, facilitating an enhanced formation of martensite, which is a challenging phase (Ref 52). So, the maximum hardness is achieved in UCMT due to the increased formation of martensite. Due to its brittle nature, martensite reduces toughness while increasing hardness, so its presence in the HAZ has an impact on the weld's hardness. The controlled cooling and refined grain structure used in ultrasonic vibration-assisted CMT welding of ferritic stainless steel help to prevent excessive martensite formation (Ref 53). It preserves increased hardness while improving toughness. Vibrations and process parameters maximize the welded joint's hardness. The increased formation of martensite limits grain growth, leading to a smaller average grain size in the WM. This relationship between grain size and microhardness was explained by the Hall-Petch equation. In comparing the microhardness between the WM, HAZ, and base metal in CMT and UCMT welds, distinct variations arise due to grain size and martensitic transformation. In CMT welds, the WM generally exhibits lower microhardness compared to the HAZ due to larger grain sizes and fewer martensitic transformations. The HAZ, particularly the FGHAZ, often shows higher microhardness due to refined grain structures and localized martensitic transformation. The base metal typically maintains consistent hardness depending on its original condition. However, in UCMT welds, ultrasonic vibrations refine grain structures across the weld zones, leading to enhanced microhardness in the WM and HAZ. The vibrations promote grain refinement and suppress martensitic growth, contributing to more uniform hardness distribution and improved mechanical properties across the welded joint.

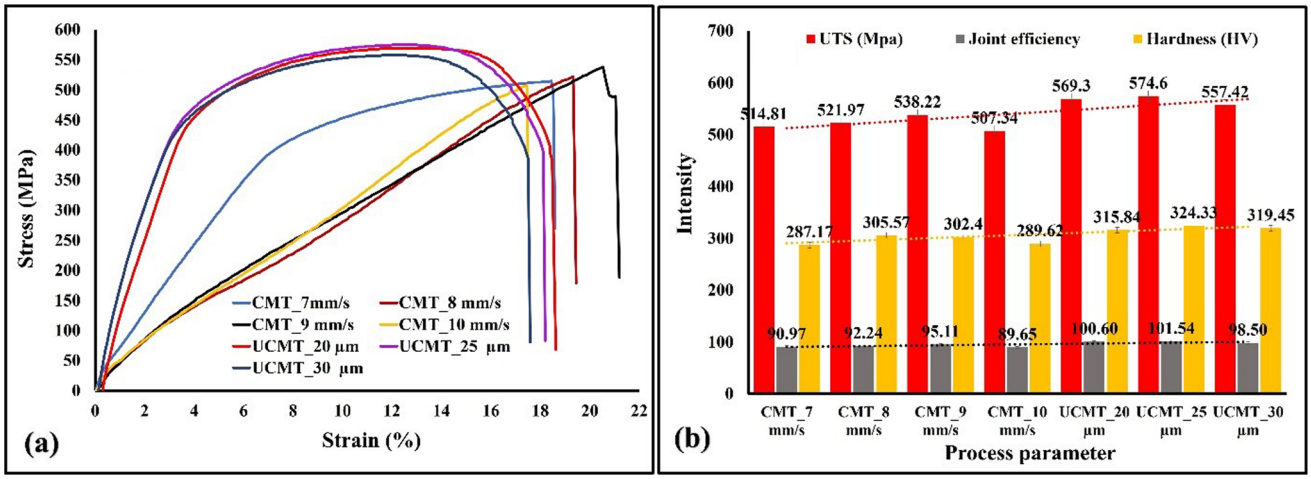



**Fig. 12** (a) Stress strain diagram of CMT- and UCMT-welded joints and (b) mechanical characteristics of the CMT and UCMT weldment at different process parameters

**Table 6 Hardness value of the CMT-welded joint at the weld center with 95% confidence interval (CI)**

| Welding speed mm/s | Wire feed rate | Microhardness, HV | | | Mean MH, HV | Standard Deviation, SD | Standard error, SE | 95% confidence Interval, CI | |
|---|---|---|---|---|---|---|---|---|---|
| | | Specimen-1 | Specimen-2 | Specimen-3 | | | | Minimum | Maximum |
| 7 | 5 | 287.63 | 288.67 | 285.41 | 287.17 | 1.58 | 0.91 | 285.38 | 288.96 |
| 8 | | 306.52 | 304.96 | 305.23 | 305.57 | 0.83 | 0.48 | 304.63 | 306.51 |
| 9 | | 301.64 | 303.46 | 302.11 | 302.40 | 0.94 | 0.55 | 301.33 | 303.47 |
| 10 | | 289.65 | 290.72 | 288.49 | 289.62 | 1.12 | 0.64 | 288.36 | 290.88 |

### 3.3 Tensile Properties

The stress strain diagram and mechanical properties of the CMT and UCMT weldments are revealed in Fig. 12a-b. The UTS and % strain of the base metal were observed at 565.85 MPa and 23.57%. The maximum joint efficiency (95.11%) of the CMT-welded joints was observed at the welding speed of 9 mm/sec with constant wire feed rate and GFR, while the maximum joint efficiency (101.54%) of the UCMT weldment was perceived at a vibration amplitude of 25 $\mu$m as shown in Fig. 12. The enhanced tensile strength in the UCMT-welded samples can be attributed to the microstructural changes induced by the ultrasonic vibrations during the CMT. In these samples, the grain structure was more refined, leading to a more uniform distribution of grains throughout the welded joint. Refined grains are known to exhibit better strain distribution, reducing the likelihood of localized stress concentrations that can lead to premature failure. In the UCMT samples, this refinement resulted in a smaller strain gap within the grains, meaning that the grains could deform more uniformly, which, in turn, delayed the onset of fracture. It is a significant advantage, as it allows the welded joint to undergo more extensive deformation before breaking, thereby improving its overall elongation. In the CMT, welded samples exhibited larger, coarser grains, which are more prone to fracturing under stress.

During this work, three statistical responses were investigated, i.e., SD, SE, and 95% CI for both CMT- and UCMT-welded joints revealed in Tables 6-9. The SD provides the deviation of the experimental values from the average values, and it may be analyzed as SD $= = \left[\sum (X_i - M)^2/(N-1)\right]^{1/2}$ (Ref 54, 55). The SE was used to analyzed how to close the predicted values to the experimental values, and SE can be analyzed as SE = SD/$N^{1/2}$, where M is the mean, and N is the observation. Different phases in the microstructures of both samples were predicted using the Schaeffler diagram. The UCMT sample consists of martensite and ferrite phases, leading to higher hardness value, as shown in Table 7. The CMT sample comprises ferrite, austenite, and martensite phases. In the UCMT weld metal, a higher percentage of martensite leads to increased UTS. The tensile strength of the weld metal is closely tied to grain size, with the Hall-Petch equation indicating that smaller grains typically enhance yield strength. Tensile tests revealed that CMT-welded samples had lower tensile properties compared to UCMT samples, primarily due to the formation of larger dendrites and increased interdendritic spacing in the CMT weld's fusion zone. These microstructural characteristics weaken the material by providing paths of least resistance where fractures can propagate. The thermal cycles during CMT welding of SS409L significantly influence martensite development in the FGHAZ and IGHAZ. Rapid

**Table 7 Hardness value of the UCMT-welded joint at the weld center with 95% CI**

| Process parameters | welding speed, mm/s | Microhardness, HV | | | Mean MH, HV | SD | SE | 95% CI | |
|---|---|---|---|---|---|---|---|---|---|
| | | Specimen 1 | Specimen 2 | Specimen 3 | | | | Minimum | Maximum |
| UCMT_20 $\mu$m | 9 | 315.52 | 317.72 | 314.29 | 315.84 | 1.74 | 1.00 | 313.88 | 317.81 |
| UCMT_25 $\mu$m | | 323.18 | 325.28 | 324.49 | 324.33 | 1.06 | 0.61 | 323.12 | 325.52 |
| UCMT_30 $\mu$m | | 320.15 | 318.59 | 319.45 | 319.45 | 0.79 | 0.46 | 318.56 | 320.35 |

**Table 8 UTS of the CMT-welded joint at 95% CI with different process parameters**

| Welding speed, mm/s | Wire feed rate | Tensile Strength, MPa | | | Mean UTS, Mpa | SD | SE | 95% CI | |
|---|---|---|---|---|---|---|---|---|---|
| | | Sample−1 | Sample−2 | Sample−3 | | | | Minimum | Maximum |
| 7 | 5 | 514.32 | 515.89 | 514.23 | 514.81 | 0.93 | 0.54 | 513.76 | 515.87 |
| 8 | | 523.62 | 520.45 | 521.84 | 521.97 | 1.59 | 0.92 | 520.17 | 523.77 |
| 9 | | 540.25 | 536.83 | 537.58 | 538.22 | 1.80 | 1.04 | 536.19 | 540.25 |
| 10 | | 509.34 | 506.74 | 505.94 | 507.34 | 1.78 | 1.03 | 505.33 | 509.35 |

**Table 9 % strain of the CMT weldment at 95% CI with different process parameters**

| Welding speed, mm/s | Wire feed rate | % Strain | | | Mean % Strain | SD | SE | 95% CI | |
|---|---|---|---|---|---|---|---|---|---|
| | | Specimen-1 | Specimen-2 | Specimen-3 | | | | Minimum | Maximum |
| 7 | 5 | 18.96 | 17.56 | 19.25 | 18.59 | 0.90 | 0.52 | 17.57 | 19.61 |
| 8 | | 20.63 | 19.23 | 18.47 | 19.44 | 1.10 | 0.63 | 18.20 | 20.68 |
| 9 | | 21.36 | 20.85 | 21.36 | 21.19 | 0.29 | 0.17 | 20.86 | 21.52 |
| 10 | | 16.67 | 18.56 | 17.36 | 17.53 | 0.96 | 0.55 | 16.45 | 18.61 |

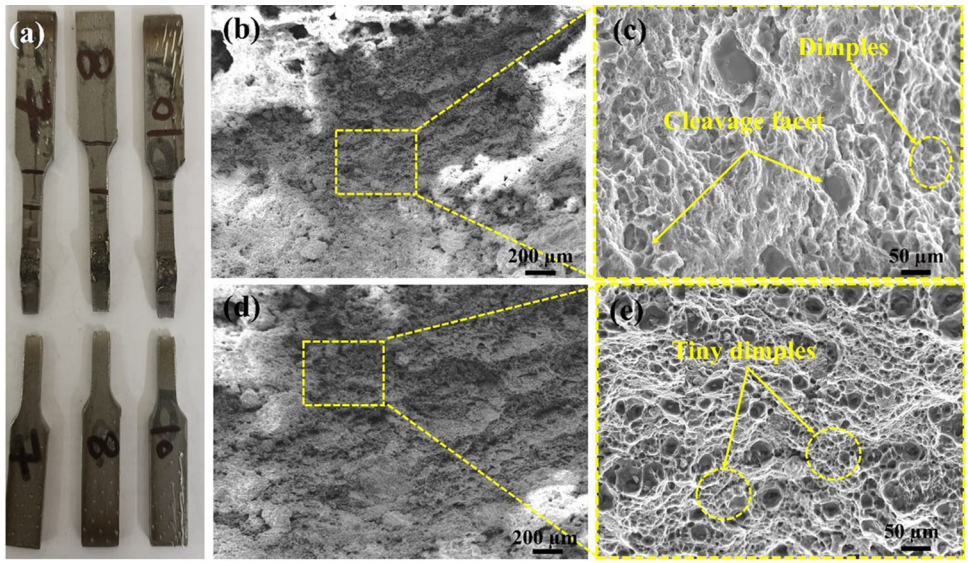



**Fig. 13** SEM images of fractured samples: (a) fractured tensile specimen, (b, c) CMT-welded joint, and (d, e) UCMT-welded joints

heating and cooling cycles promote austenite formation, which can transform into martensite upon cooling, especially in the FGHAZ due to higher cooling rates. In contrast, slower cooling in the IGHAZ may limit martensite formation. This differential impact affects the mechanical properties, with the FGHAZ potentially exhibiting increased hardness and reduced toughness compared to the IGHAZ (Ref 56). The multiple thermal cycles involved in the welding process, particularly in CMT,

exacerbate these effects by contributing to grain growth and the coarsening of microstructures. The maximum tensile strength of CMT- and UCMT-welded joints was observed as 538.22 MPa and 574.6 MPa, demonstrated in Table 8 and Fig. 12. While the maximum % strain at 95% CI was observed at 21.19% at the welding speed of 9 mm/sec as shown in Table 9, the UTS of the CMT-welded joints increases with increasing welding speed up to 9 mm/sec. After this, UTS decreases when the welding speed is increased.

### 3.4 Fractography of CMT- and UCMT-Welded Joints

A shear plane typically forms at an angle of nearly 45° to the tensile axis dueing the tensile load, creating a cup and cone fracture. This fracture pattern occurs as the shear plane merges with the existing crack, leading to the final failure of the material (Ref 57). The fractured SEM images of tensile samples as shown in Fig. 13(a-d) reveal key differences in the fracture characteristics of the CMT- and UCMT-welded joints. In the CMT-welded fracture surface, as shown in Fig. 13, a mixed-mode fracture is observed, characterized by the coalescence of microvoids. This fracture surface includes both cleavage facets and elongated ductile dimples. Cleavage facets are indicative of brittle fracture, where the material breaks along specific crystallographic planes, while the elongated ductile dimples suggest a more ductile failure mechanism where microvoids form, grow, and eventually coalesce to cause the material to fracture.

Fig. 13(a-e) displays fracture surfaces of CMT- and UCMT-welded joints. SEM images show mixed-mode fractures, including ductile dimples formed by microvoid coalescence and transgranular cleavage facets (Ref 51). Grain coarsening with longer tempering durations enhances microvoid formation while reducing cleavage area. Heterogeneous dimple size was also observed, highlighting a transition from brittle to ductile fracture behavior. His presence of these features indicates that the base metal experienced a combination of brittle and ductile failure modes. For the CMT-welded specimen, as shown in Fig. 13(c), the fracture zone exhibits a significant heterogeneity in the size of the dimples. The fracture surface reveals larger dimples and more extensive cleavage facets compared to the UCMT-welded samples. The larger size of the dimples and cleavage facets indicates a more brittle fracture behavior in the CMT weld, which suggests that the weld zone experienced less ductility and more abrupt failure. This heterogeneity in the dimple size and the presence of large cleavage facets point to an uneven distribution of stress and strain within the weld, leading to localized brittle failure (Ref 58, 59). The fracture zone of the UCMT-welded specimen, depicted in Fig. 13(d, e), reveals similarities to that of the CMT weld but with notable differences. The UCMT weld fracture surface also contains both ductile dimples and cleavage facets, but the area of cleavage facets is smaller, and there is a higher degree of heterogeneity in the dimple size compared to the BM. The smaller cleavage facet area suggests that the UCMT weld exhibited a more ductile behavior than the CMT weld, although still less ductile than the BM (Ref 60, 61). The increased heterogeneity in dimple size in the UCMT weld indicates variability in the local mechanical properties. The tensile test results for all samples showed that the fractures occurred primarily in the BM, although the fracture surfaces of the BM differed from those of the samples taken from the CMT and UCMT weldments. The gauge section of the transverse tensile specimens included the weld fusion zone, HAZ, and the BM. Each of these zones has distinct mechanical properties and strengths, resulting in an asymmetrical strain distribution along the gauge length during tensile testing. As the significant elongation occurred predominantly in the BM zone, the effective strain rate in this zone was higher compared to that in the BM specimens despite the crosshead speed being consistent in all tests. This difference in strain dispersal led to variations in the fracture surface morphology between the BM and the welded specimens (Ref 62). The fractured surfaces of the tensile-tested specimens also revealed essential insights into the failure mechanisms. For instance, the occurrence of cleavage facets, microvoids, and tiny dimples across the fracture surfaces indicates the complex interplay between ductile and brittle failure modes. The maximum UTS of the specimens was influenced by these microstructural features, with the BM generally showing better performance due to its more uniform microstructure. In contrast, the welded specimens, mainly the CMT weld, exhibited lower ductility compared to UCMT.

## 4. Conclusions

In the present investigation, CMT- and UCMT-welded joints were successfully fabricated via different process parameters, and the following conclusions were drawn.

- The maximum microhardness value of 305.67 HV was observed at the weld center in the CMT joint at 9 mm/sec, while the UCMT joint achieved a peak hardness of 325.32 HV at a 25 $\mu$m amplitude. The SCHAZ exhibited a microhardness of 210-235 HV, which was nearly equivalent to the base metal microhardness
- The UCMT welding process significantly refines the grain structure in various zones of the weld, particularly in the FGHAZ and CGHAZ, through dynamic recrystallization. This refinement enhances mechanical properties, such as tensile strength and microhardness.
- The UCMT-welded joint demonstrated superior mechanical performance, achieving a maximum tensile strength of 574.6 MPa and a joint efficiency of 101.54%, compared to the CMT-welded joint, which had a tensile strength of 538.22 MPa and a joint efficiency of 95.11%.
- Ultrasonic vibrations in UCMT also contribute to a more homogeneous microstructure by redistributing elements and stabilizing the austenite phase, particularly in the IGHAZ. This leads to a reduction in martensitic transformation and the formation of a balanced microstructure in the WFZ, combining the hardness of martensite with the toughness of ferrite, ultimately optimizing the weld's overall mechanical properties.
- The UCMT welds displayed a more favorable balance between ductility and brittleness, with smaller cleavage facets and higher heterogeneity in dimple size compared to the CMT welds, indicating better ductile behavior and less abrupt failure.

## References


1. K.A. Cashell and N.R. Baddoo, Ferritic stainless steels in structural applications, *Thin-Walled Struct.*, 2014, **83**, p 169–181. https://doi.org/10.1016/j.tws.2014.03.014
2. Y. Chen, S. Sun, T. Zhang, X. Zhou, and S. Li, Effects of post-weld heat treatment on the microstructure and mechanical properties of laser-welded NiTi/304SS joint with Ni filler, *Mater. Sci. Eng. A*, 2020, **771**, 138545. https://doi.org/10.1016/j.msea.2019.138545
3. L. Shao, X. Zhang, Y. Chen, L. Zhu, S. Wu, Q. Liu, and T. Wang, Why do cracks occur in the weld joint of Ti-22Al-25Nb alloy during post-weld heat treatment?, *Front. Mater.*, 2023, **10**, p 1135407. https://doi.org/10.3389/fmats.2023.1135407
4. H. Zhang, H. Liu, and H. Kuai, Stress intensity factor analysis for multiple cracks in orthotropic steel decks rib-to-floorbeam weld details under vehicles loading, *Eng. Fail. Anal.*, 2024, **164**, 108705. https://doi.org/10.1016/j.engfailanal.2024.108705
5. J. Liu, J. Xu, K. Paik, P. He, and S. Zhang, In-situ isothermal aging TEM analysis of a micro Cu/ENIG/Sn solder joint for flexible interconnects, *J. Mater. Sci. Technol.*, 2024, **169**, p 42–52. https://doi.org/10.1016/j.jmst.2023.06.020
6. S. Selvi, A. Vishvaksenan, and E. Rajasekar, Cold metal transfer (CMT) technology - An overview, *Def. Technol.*, 2017, **14**(1), p 28–44. https://doi.org/10.1016/j.dt.2017.08.002
7. N. Yuvaraj and J.G. Roy, Enhancement of microstructural and mechanical properties of ultrasonic vibration-assisted cold metal transfer welding of 304 stainless steel, *J. Mater. Eng. Perform.*, 2022 https://doi.org/10.1007/s11665-022-06877-3
8. Y. Wang, C. Yu, H. Lu, and J. Chen, Research status and future perspectives on ultrasonic arc welding technique, *J. Manuf. Process.*, 2020, **58**(June), p 936–954. https://doi.org/10.1016/j.jmapro.2020.09.005
9. M.J. Jose, S.S. Kumar and A. Sharma, Vibration assisted welding processes and their influence on quality of welds, *Sci. Technol. Weld. Join.*, 2016, **21**(4), p 243–258. https://doi.org/10.1179/1362171815Y.0000000088
10. S. Kumar, C. S. Wu, G. K. Padhy, and W. Ding, "Application of ultrasonic vibrations in welding and metal processing: A status review," vol. 26, pp. 295–322, 2017, https://doi.org/10.1016/j.jmapro.2017.02.027
11. T. Watanabe, M. Shiroki, A. Yanagisawa, and T. Sasaki, Technology Improvement of mechanical properties of ferritic stainless steel weld metal by ultrasonic vibration, *J. Mater. Process.*, 2010, **210**, p 1646–1651. https://doi.org/10.1016/j.jmatprotec.2010.05.015
12. W. Dai, Effects of high-intensity ultrasonic-wave emission on the weldability of aluminum alloy 7075–T6, *Mater. Lett.*, 2003, **57**, p 2447–2454. https://doi.org/10.1016/S0167-577X(02)01262-4
13. Y. Morisada, H. Fujii, F. Inagaki, and M. Kamai, Development of high frequency tungsten inert gas welding method, *Mater. Des.*, 2013, **44**, p 12–16. https://doi.org/10.1016/j.matdes.2012.07.054
14. A. Abdullah, M. Malaki, and A. Eskandari, Strength enhancement of the welded structures by ultrasonic peening, *Mater. Des.*, 2012, **38**, p 7–18. https://doi.org/10.1016/j.matdes.2012.01.040
15. G. Zhao, Z. Wang, S. Hu, S. Duan, and Y. Chen, Effect of ultrasonic vibration of molten pool on microstructure and mechanical properties of Ti-6Al-4V joints prepared via CMT + P welding, *J. Manuf. Process.*, 2020, **52**, p 193–202. https://doi.org/10.1016/j.jmapro.2020.01.045
16. N. Kumar, P. Kumar, R.S. Vidyarthy et al., Role of Microstructure Evolution During Welding on Mechanical Properties and Residual Stresses of the Inconel 718 and Austenitic Stainless Steel 304L Dissimilar Weld Joint, *Metall. Mater. Trans. A*, 2024, **55**, p 4543–4570. https://doi.org/10.1007/s11661-024-07571-w
17. N. Kumar, P. Kumar, and C. Pandey, Influence of filler materials on GTAW dissimilar welds: Inconel 718 and austenitic stainless steel 304L, *Arch. Civ. Mech. Eng.*, 2024, **24**, p 231. https://doi.org/10.1007/s43452-024-01042-0
18. V. Bhanu, J. Manoj, A. Gupta et al., Microstructural evolution and mechanical behavior of activated tungsten inert gas welded joint between P91 steel and Incoloy 800HT, *Int. J. Adv. Manuf. Technol.*, 2024, **134**, p 223–244. https://doi.org/10.1007/s00170-024-14117-2
19. A. Kumar, S. Sirohi, M. Singh, and D. Fydrych, Microstructure and mechanical properties of a dissimilar metal welded joint of Inconel 617 and P92 steel with Inconel 82 buttering layer for AUSC boiler application, *Int. J. Press. Vessels Pip.*, 2024, **209**, 105196. https://doi.org/10.1016/j.ijpvp.2024.105196
20. M.M. Chandan Pandey, P.K. Mahapatra, and N. Saini, Effect of normalization and tempering on microstructure and mechanical properties of V-groove and narrow-groove P91 pipe weldments, *Mater. Sci. Eng. A*, 2017, **685**, p 39–49.
21. T. Yuan, S. Kou, and Z. Luo, Grain refining by ultrasonic stirring of the weld pool, *Acta Mater.*, 2016, **106**, p 144–154. https://doi.org/10.1016/j.actamat.2016.01.016
22. Y. Tian, J. Shen, S. Hu, Z. Wang, and J. Gou, E ff ects of ultrasonic vibration in the CMT process on welded joints of Al alloy, *J. Mater. Process. Tech.*, 2018, **259**, p 282–291. https://doi.org/10.1016/j.jmatprotec.2018.05.004
23. Q.C.S. Lin and C.Y.C. Fan, Effect of Ultrasound on Heterogeneous Nucleation in TIG Welding of Al – Li Alloy, *Acta Metall. Sin.*, 2016, **29**(12), p 1081–1088. https://doi.org/10.1007/s40195-016-0483-1
24. M. Fattahi, A. Ghaheri, N. Arabian, F. Amirkhanlu, and H. Moayedi, Jo ur l P re of, *J. Mater. Process. Tech.*, 2020 https://doi.org/10.1016/j.jmatprotec.2020.116672
25. F. Ning and W. Cong, Microstructures and mechanical properties of Fe-Cr stainless steel parts fabricated by ultrasonic vibration-assisted laser engineered net shaping process, *Mater. Lett.*, 2016, **179**, p 61–64. https://doi.org/10.1016/j.matlet.2016.05.055
26. C. Chen, C. Fan, X. Cai, S. Lin, C. Yang and Y. Zhuo, Microstructure and mechanical properties of Q235 steel welded joint in pulsed and un-pulsed ultrasonic assisted gas tungsten arc welding, *J. Mater. Process. Technol.*, 2020, **275**(92), 116335. https://doi.org/10.1016/j.jmatprotec.2019.116335
27. S. Zhou, G. Ma, W. Dongjiang, D. Chai, and M. Lei, Ultrasonic vibration assisted laser welding of nickel-based alloy and Austenite stainless steel, *J. Manuf. Process.*, 2018, **31**, p 759–767. https://doi.org/10.1016/j.jmapro.2017.12.023
28. W. Xie, T. Huang, C. Yang, C. Fan, S. Lin, and W. Xu, Comparison of microstructure, mechanical properties and corrosion behavior of Gas Metal Arc (GMA) and Ultrasonic-wave-assisted GMA (U-GMA) welded joints of Al–Zn–Mg alloy, *J. Mater. Process. Technol.*, 2020, **277**, p 116470. https://doi.org/10.1016/j.jmatprotec.2019.116470
29. C. Chen, C. Fan, X. Cai, S. Lin, C. Yang, and Y. Zhuo, Microstructure and mechanical properties of Q235 steel welded joint in pulsed and un-pulsed ultrasonic assisted gas tungsten arc welding, *J. Mater. Process. Technol.*, 2020, **275**, 116335. https://doi.org/10.1016/j.jmatprotec.2019.116335
30. X. Fang, S. Wu, S. Lü, J. Wang, and X. Yang, Microstructure evolution and mechanical properties of quasicrystal-reinforced Mg-Zn-Y alloy subjected to ultrasonic vibration, *Mater. Sci. Eng. A*, 2017, **679**, p 372–378. https://doi.org/10.1016/j.msea.2016.10.035
31. Y. Koli, J.G. Roy, N. Yuvaraj Vipin, and G. Vedabouriswaran, Microstructural evolution, ferrite content, and grain size study in the coalescence of high nitrogen austenitic stainless steel thick plates using cold metal transfer welding process, *J. Mater. Eng. Perform.*, 2023, **33**(9), p 4422–4434. https://doi.org/10.1007/s11665-023-08247-z
32. Y. Tian, J. Shen, Hu. Shengsun, Z. Wang, and J. Gou, Effects of ultrasonic vibration in the CMT process on welded joints of Al alloy, *J. Mater. Process. Technol.*, 2018, **259**, p 282–291. https://doi.org/10.1016/j.jmatprotec.2018.05.004
33. N. Yuvaraj, J.G. Roy, and Vipin, Enhancement of Microstructural and Mechanical Properties of Ultrasonic Vibration-Assisted Cold Metal Transfer Welding of 304 Stainless Steel, *J. Mater. Eng. Perform.*, 2022, **31**, p 8497–8511. https://doi.org/10.1007/s11665-022-06877-3
34. T. Madsa, P. Jattakul, S. Pansa-nga et al., Investigation of CMT Welding AISI H13 Hot Work Tool Steel on Weld Bead Geometry, Microstructure, and Mechanical Performance, *Arab. J. Sci. Eng.*, 2024, **49**, p 10943–10959. https://doi.org/10.1007/s13369-023-08547-5
35. G. Posch, K. Chladil, and H. Chladil, Material properties of CMT—metal additive manufactured duplex stainless steel blade-like geometries, *Weld World*, 2017, **61**, p 873–882. https://doi.org/10.1007/s40194-017-0474-5
36. Y.L. Sun et al., Effects of dilution on alloy content and microstructure in multi-pass steel welds, *J. Mater. Process. Technol.*, 2019, **265**, p 71–86. https://doi.org/10.1016/j.jmatprotec.2018.09.037
37. S.K. Gupta, A.R. Raja, M. Vashista, and M.Z.K. Yusufzai, Effect of heat input on microstructure and mechanical properties in gas metal arc welding of ferritic stainless steel, *Mater. Res. Express*, 2019, **6**, p 036516. https://doi.org/10.1088/2053-1591/aaf492

38. S. N. Kane, A. Mishra, and A. K. Dutta, 2016 "Preface: International Conference on Recent Trends in Physics (ICRTP 2016)," J. Phys. Conf. Ser., https://doi.org/10.1088/1742-6596/755/1/011001
39. M. Mukherjee and T.K. Pal, Influence of Heat Input on Martensite Formation and Impact Property of Ferritic-Austenitic Dissimilar Weld Metals, *J. Mater. Sci. Technol.*, 2012, **28**(4), p 343–352. https://doi.org/10.1016/S1005-0302(12)60066-8
40. W. Reitz, A Review of: 'Welding Metallurgy and Weldability of Stainless Steel,' *Mater. Manuf. Processes*, 2006, **21**(2), p 219–219. https://doi.org/10.1080/10426910500476747
41. W. Bingbo, Unidirectional dendritic solidification under longitudinal resonant vibration, *Acta Metall. Mater.*, 1992, **40**(10), p 2739–2751. https://doi.org/10.1016/0956-7151(92)90344-E
42. S.S.M. Tavares, C.R. Rodrigues, C.A.S. Oliveira, C.B. Woyames, and J. Dille, Influence of Heat Treatments on Microstructure and Toughness of 9% Ni Steel, *J. Mater. Eng. Perform.*, 2018, **27**, p 1530–1536. https://doi.org/10.1007/s11665-018-3257-2
43. X.-Q. Zhao, T. Pan, Q.-F. Wang, H. Su, C.-F. Yang, and Q.-X. Yang, Effect of Tempering Temperature on Microstructure and Mechanical Properties of Steel Containing Ni of 9%, *J. Iron. Steel Res. Int.*, 2011, **18**, p 47–51. https://doi.org/10.1016/S1006-706X(11)60064-2
44. S. Sirohi, A. Gupta, C. Pandey, R.S. Vidyarthy, K. Guguloth, and H. Natu, Investigation of the microstructure and mechanical properties of the laser welded joint of P22 and P91 steel, *Opt. Laser Technol.*, 2022, **147**, 107610. https://doi.org/10.1016/j.optlastec.2021.107610
45. H.C. Ferreira, F.J.M. Boratto, and V.T.L. Buono, Determination of low levels of retained austenite in low-carbon high-manganese steel using x-ray diffraction, *Mater. Sci. Eng. A*, 2015, **628**, p 110–115. https://doi.org/10.1016/j.msea.2015.01.019
46. A. Ebrahimian and S.S.G. Banadkouki, Mutual mechanical effects of ferrite and martensite in a low alloy ferrite-martensite dual phase steel, *J. Alloys Compd.*, 2017, **708**, p 43–54. https://doi.org/10.1016/j.jallcom.2017.02.287
47. J.C. Lippold and D.J. Kotecki, *Welding Metallurgy and Weldability of Stainless Steel; John Wiley & Sons: Hoboken*, NJ, USA, 2005
48. W.M. Van, D. Nolan, and J. Norrish, Mitigation of sensitisation effects in unstabilised 12% Cr ferritic stainless steel welds, *Mater. Sci. Eng. A*, 2007, **464**, p 157–169.
49. H.S. Surkar, A. Kumar, S. Sirohi et al., A dissimilar welded joint of grade 92 steel and AISI 304L steel obtained using IN82 buttering and IN617 filler: relationship of microstructure and mechanical properties, *Archiv. Civ. Mech. Eng*, 2024, **24**, p 109. https://doi.org/10.1007/s43452-024-00920-x
50. P.G.Z. Emygdio, K.A. Santos, A. Zeemann, and O.R. Mattos, Behavior of 9Ni steel in the presence of aqueous solutions containing CO2 and H2S, *Corrosion*, 2016, **73**, p 303–314. https://doi.org/10.5006/2198
51. J.-I. Jang, J.-B. Ju, B.-W. Lee, D. Kwon, and W.-S. Kim, Micromechanism of local brittle zone phenomenon in weld heat-affected zones of advanced 9% Ni steel, *J. Mater. Sci. Lett.*, 2001, **20**, p 2149–2152. https://doi.org/10.1023/A:1013784600394
52. S. Wang, X. Chen, K. Luo, H. Zhou, R. Li, P. He, and S. Zhang, The design of low-temperature solder alloys and the comparison of mechanical performance of solder joints on ENIG and ENEPIG interface, *J. Market. Res.*, 2023, **27**, p 5332–5339. https://doi.org/10.1016/j.jmrt.2023.11.066
53. Q.J. Sun, S.B. Lin, C.L. Yang, and G.Q. Zhao, Rapid Communication: Penetration increase of AISI 304 using ultrasonic assisted tungsten inert gas welding, *Sci. Technol. Weld. Join.*, 2009, **14**(8), p 765–767. https://doi.org/10.1179/136217109X12505932584772
54. J. Li, J. Shen, S. Hu, Y. Liang, and Q. Wang, Microstructure and mechanical properties of 6061/7N01 CMT+P joints, *J. Mater. Process. Technol.*, 2019, **264**, p 134–144. https://doi.org/10.1016/j.jmatprotec.2018.09.011
55. H. Mehdi and R.S. Mishra, Effect of friction stir processing on mechanical properties and heat transfer of TIG welded joint of AA6061 and AA7075, *Def. Technol.*, 2021, **17**(3), p 715–727. https://doi.org/10.1016/j.dt.2020.04.014
56. H. Mehdi and R.S. Mishra, Influence of friction stir processing on weld temperature distribution and mechanical properties of TIG-welded joint of AA6061 and AA7075, *Trans. Indian Inst. Met.*, 2020, **73**, p 1773–1788. https://doi.org/10.1007/s12666-020-01994-w
57. C. Pandey and M.M. Mahapatra, Effect of heat treatment on microstructure and hot impact toughness of various zones of P91 welded pipes, *J. of Materi Eng and Perform*, 2016, **25**, p 2195–2210. https://doi.org/10.1007/s11665-016-2064-x
58. C. Pandey, N. Saini, M.M. Mahapatra, and P. Kumar, Study of the fracture surface morphology of impact and tensile tested cast and forged (C&F) Grade 91 steel at room temperature for different heat treatment regimes, *Eng. Failure Anal.*, 2017, **71**, p 131–147.
59. C. Pandey, M.M. Mahapatra, P. Kumar, F. Daniel and B. Adhithan, Softening mechanism of P91 steel weldments using heat treatments, *Arch. Civil Mech. Eng.*, 2019, **19**(2), p 297–310.
60. C. Yuhua, M. Yuqing, L. Weiwei, and H. Peng, Investigation of welding crack in micro laser welded NiTiNb shape memory alloy and Ti6Al4V alloy dissimilar metals joints, *Opt. Laser Technol.*, 2017, **91**, p 197–202. https://doi.org/10.1016/j.optlastec.2016.12.028
61. S. Govind, S.S. Ghule, C. Satish, C. Robert, K. Ajay, Y.B. Suresh, and K. Rakesh, Investigation of conventional and ultrasonic vibration-assisted turning of hardened steel using a coated carbide tool, *Front. Mech. Eng.*, 2024, **10**, p 1–11.
62. J. Zhou, J. Shen, Hu. Shengsun, G. Zhao, and Q. Wang, Microstructure and mechanical properties of AISI 430 ferritic stainless steel joints fabricated by cold metal transfer welding, *Mater. Res. Express*, 2019, **6**(11), 116536

**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.

Water Resources Management
https://doi.org/10.1007/s11269-025-04222-6

# Deciphering the Sustainable Stormwater Management Strategies for Urban Areas: a Review

**Shivani Yadav[1] · Saurav Ambastha[1] · Harsh Pipil[1] · Anil Kumar Haritash[1] · Krishna R. Reddy[2]**

Received: 4 June 2024 / Accepted: 16 April 2025
© The Author(s), under exclusive licence to Springer Nature B.V. 2025

## Abstract

The imperative concern of water scarcity has prompted the exploration of different strategies to meet the escalating demands of the growing population. The constrained accessibility to freshwater, vital for both potable and non-potable usage, has intensified the pressure on the existing water resources. The deteriorating quality of the available freshwater poses a significant threat contributing to adverse health effects on the environment and living beings. Notably, a substantial volume of rainwater reaches the Earth's surface as wet precipitation, and its interception before natural drainage along the terrain holds potential for reuse. The current study entails the different techniques such as gross pollutant traps, bioswales, and raingardens that can effectively mitigate the stormwater runoff impurities, encompassing organic matter (BOD), floating debris (leaves, paper, plastic), dissolved pollutants (nutrients like N and P, heavy metals, hydrocarbons) and suspended particles (sand, silt). The treatment systems not only contribute towards the preservation and treatment of stormwater runoff but also yield ancillary benefits, including better air quality, carbon sequestration, improved urban aesthetics, and a healthier ecology. The ensuing discussion in this paper delves into the nuanced management and treatment of stormwater runoff in urban areas of developing nations through the application of different techniques.

## Highlights

- The current paper describes the recent development and application of different stormwater treatment techniques.
- Treatment of stormwater runoff can bridge the demand and supply gap of water.
- Dependence on scarce freshwater resources can be reduced by proper management of stormwater runoff.
- Treated stormwater runoff can be used efficiently for non-potable applications.

**Keywords** Stormwater runoff · Stormwater management · WSUD · Bioretention basin · Nutrients

Extended author information available on the last page of the article

Published online: 30 April 2025

## 1 Introduction

The rapidly increasing population poses a threat in the coming future leading to pressure on scarce freshwater resources, sanitation system, and causing other environmental issues (Ezeh et al. 2012). The increasing population is leading to an increase in urbanisation which causes the replacement of forests and vegetation by impervious roads, pavements, houses, buildings, and rooftops. The exponential surge in impervious earth's surface due to urbanisation has also increased the stormwater runoff volume (Hogan and Walbridge 2007), and subsequent decrease in groundwater recharge due to poor infiltration. In addition, the non-uniform downpours lead to excessive rainfall over different parts of the world and the rest receives scanty rainfall (Ghosh et al. 2016). In ancient times, people used to collect and store stormwater from roofs, terraces and specifically designed paved surfaces for potable and non-potable use during the dry season (Gikas and Angelakis 2009). When the stormwater falls over various land use, its interaction with the earth's surface and infrastructure causes sweeping of deposited sand/silt, clay, minerals, heavy, and other organic impurities leading to the addition of suspended as well as dissolved impurities in the stormwater runoff (Pipil et al. 2022). As a consequence of the presence of impurities, the water quality of receiving urban water bodies has deteriorated slowly, and it impacts both human life and the aquatic ecosystem (Björklund et al. 2018). Though rainwater is the cleanest source of water, it also gets contaminated when it comes in contact with atmospheric impurities (Gonçalves et al. 2003). Proper management of stormwater runoff has the potential to minimise its wastage by preventing the mixing of pollutants with stormwater runoff, and its disposal into sewage and river, while, promoting its interception and treatment (Karami et al. 2025; Liu et al. 2025). Reusing the treated stormwater could reduce the dependence on other surface water resources further lessening groundwater extraction, and help in meeting the demand for water in many developing countries (Pipil et al. 2022). Efforts were made to reduce the contamination of stormwater runoff at the source which is a back-breaking practice, and decentralised water treatment units possess the challenge of providing treatment units/facilities in limited urban spaces (Brown and Farrelly 2009). Stormwater treatment is a significant treatment approach implemented in developed countries but they lack awareness as well as implementation in many developing countries.

Considering the different aspects mentioned above, the current paper is an attempt to provide a comprehensive overview of stormwater runoff contamination and sources of contamination. The review paper also draws attention to various stormwater runoff treatment techniques, and recent advancements in the field. Bibliometric analysis was also conducted to examine the awareness of treatment and management practices of stormwater runoff in different geographical regions. Moreover, the paper entails the dynamics at play in stormwater management, and factors affecting the efficiency and performance of treatment units. The paper provides current insight into stormwater management and its treatment, the knowledge and understanding can help policy makers, urban planners, and researchers across the globe for the implementation of such techniques in developed and developing nations.

## 2 Research Design

The current study focuses on stormwater management techniques and methods with an emphasis on reducing the cost of treatment using sustainable treatment methods. Such techniques will help in mitigating the demand and supply gap in urban areas. The treatment of contaminants originating from various land uses in urban areas such as highways, residential areas, etc. is also addressed in the present paper. Such techniques may include bioswales for maintaining the green belts which helps in attenuation of air and noise pollution. The paper also entails the sustainable management of stormwater in urban areas to ensure water security and capacity building, combat climate change, and help in building city resilience. The literature survey was conducted from the year 2000 to 2023 in Scopus considering the keywords stormwater runoff, stormwater treatment, bioretention systems, WSUD, green infrastructure, wetlands, bioswales, etc. in which the most relevant papers were considered while the rest of the irrelevant papers were removed. The literature survey also included different geographical regions which highlights the various treatment techniques from different climatic regions which were addressed in the present paper.

## 3 Bibliometric Analysis

A comprehensive bibliometric analysis was performed on strategies to manage stormwater runoff. Bibliometric analysis provides a basis to unveil the article publication trend, journal performance, and collaborations (Donthu et al. 2021). The study was based on the data retrieved from Scopus for the duration year 2000 to 2023 which included the keywords "stormwater" OR "runoff" OR "stormwater treatment" OR "bioretention systems" OR "WSUD" OR "LID" OR "green infrastructure", searched under "Article title and Abstract" which provided a total of 828 articles. This provides the basis on which the progress in the field of stormwater management can be estimated in the last few decades in various countries. Over a period of last two decades, it can be seen that the popularity and attention towards stormwater runoff management has increased significantly (Fig. 1). Until 2006, less than 10 articles were published which increased significantly to 90 articles in year 2019. However, a dip in the number of articles published can be seen in few years between 2000 and 2023. Country/territory-wise publication of articles between 2000 and 2023 is shown in Fig. 2. Most of the articles (509) were published by the researchers in United States of America alone. The significant number of articles published in the defined period by the countries follows the trend as Australia, China, United Kingdom, Canada, France, and Germany which have published 64, 47, 40, 35, 19, and 17 articles, respectively (Fig. 2). However, India has published 9 number of articles on stormwater runoff management from 2000 to 2023 which is a significant number among all the developing nations.

## 4 Stormwater Contamination Source

The increase in urbanisation has resulted in an increase in impervious land cover in terms of roads, rooftops, etc., which reduces the infiltration of rainwater into the ground. This leads to the generation of high volume of stormwater runoff from various point and non-point

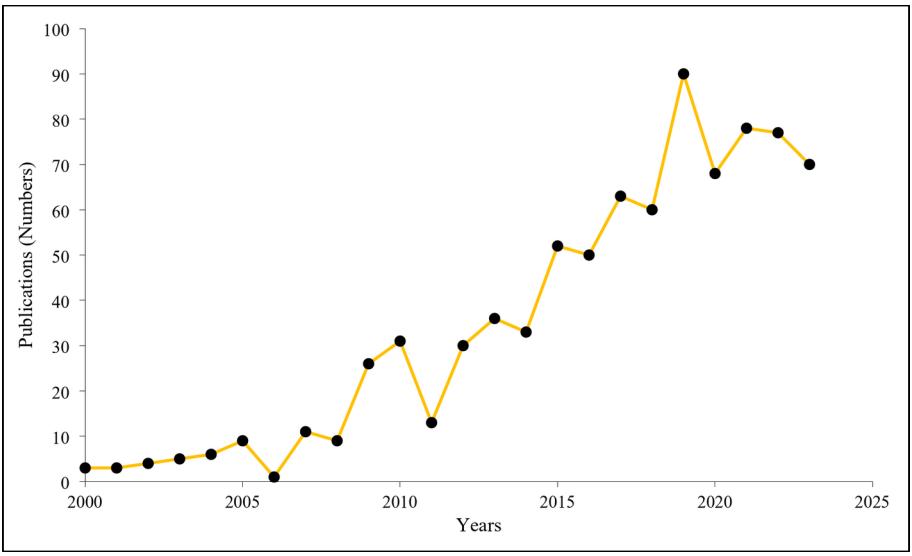



**Fig. 1** Yearly publications on stormwater management

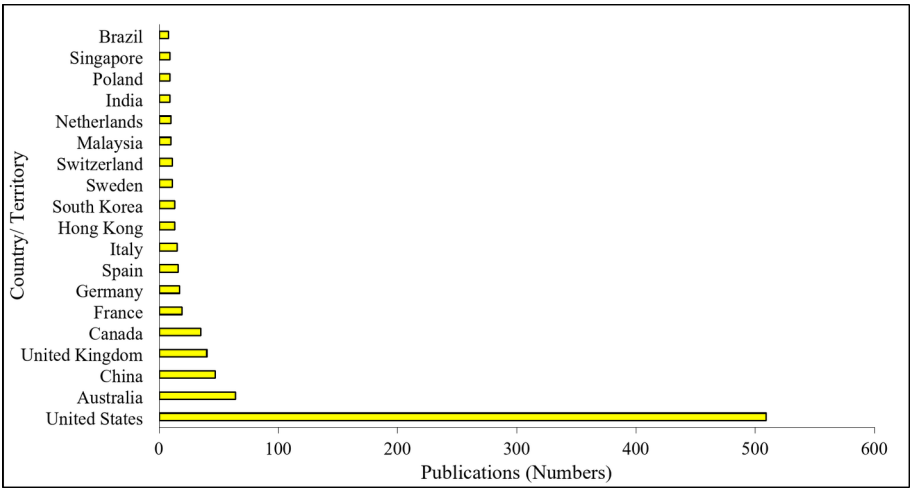



**Fig. 2** Country/territory-wise publications on stormwater management

sources of contamination which ultimately mixes and degrades the quality of the receiving surface water. The atmospheric deposition of pollutants due to anthropogenic activity and atmospheric washing during rainfall events leads to atmospheric pollutants scavenging. Followed by it, the runoff ultimately mixes with the deposited contaminants on the earth's surface during the dry period. A summary of contamination of rainwater and stormwater runoff through different polluting agencies has been described in Fig. 3.

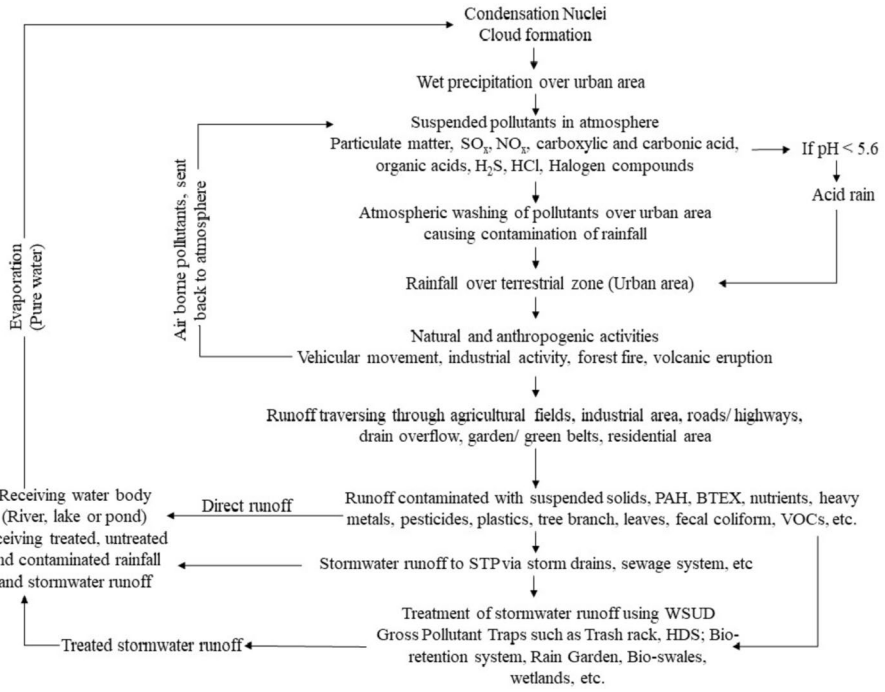



**Fig. 3** Sources of stormwater runoff contamination and their fate in the environment

### 4.1 Atmospheric Pollutants

Rainfall is considered the cleanest source of water until it comes in contact with atmospheric pollutants released either from natural or anthropogenic sources. The downpour of first rainfall scavenges and dissolves all the atmospheric impurities with it, triggering the contamination of the rainfall thus changing its quality, chemical composition, and pH, before reaching earth's surface (Al-Khashman 2005). The presence of sulphur dioxide, oxides of nitrogen, carboxylic and carbonic acid, and other acidic precursors in the atmosphere can lead to a decrease in the pH of rainfall, which when reaches pH <5.6 causes acid raid (Migliavacca et al. 2005). Volcanic eruption and forest fires are the natural sources of $CO_2$, particulate matter, nitrogen, and sulphur species such as sulphur dioxide ($SO_2$), hydrogen sulphide ($H_2S$), hydrogen chloride (HCl), halogen compounds, and aerosols (Kumar et al. 2019). These pollutants originate from local or remote anthropogenic activities, such as vehicular movement, industrial activities, and vehicular emission causing terrestrial, naval, and aerial pollution, releasing particulate matter, $NO_3^-$, $SO_4^{2-}$; carcinogenic species like polycyclic aromatic hydrocarbon (PAH), and persistent organic pollutants (POP); heavy metals, hydrocarbons, BTEX compounds (such as benzene, ethylbenzene, and toluene), polychlorinated dibenzo-para-dioxins (PCDD) and polychlorinated dibenzofurans (PCDF) (Adeyanju and Manohar 2017). Many of these pollutants are also on the list of criteria pollutants as suggested by the United States Environmental Protection Agency (United States Environmental Protection Agency 1983). Heavy metals such as arsenic (As), cadmium (Cd), chromium (Cr), copper (Cu), mercury (Hg), nickel (Ni), lead (Pb), selenium (Se), zinc

(Zn) and vanadium (V) are also released into the atmosphere through industrial activities (Mishra et al. 2019). Other pollutants that contribute to air pollution are volatile organic compounds (VOCs) originating from fuel, refrigerants, paints, and solvents; non-methane volatile organic compounds (NMVOCs) such as ketone and aldehydes (Liu et al. 2018). These pollutants stay in the atmosphere of the urban area until they are washed out by wet precipitation. Such pollutants have also been reported as contaminants in rainfall in many researches done earlier (Table 1).

As the rainfall reaches the earth's surface, it comes in contact with the pollutants that are already deposited on the various earth's surfaces during dry period under the effect of gravity. Due to the presence of impervious surfaces, it traverses its path through roads, pavements, highways, buildings, rooftops, sheds, canopy, trees, open spaces, and green belts, from where stormwater runoff sweeps and mixes floating, suspended, and dissolved impurities in it.

### 4.2 Vehicular Movement

The vehicular movement and its presence on the road cause the dispersion and re-deposition of contaminants over the paved surfaces and roads. The pollutants released from the exhaust pipe of internal combustion (IC) engines are deposited on the road, dripping/leakage of automotive fluid, vehicle washing, road accidents, and abrasion of pavement materials also contribute towards the deposition of the pollutants (Markiewicz et al. 2017). The pollutants contributing from vehicular movement included total suspended solids (TSS), PAH, BTEX, BOD, COD, nitrogen oxide ($NO_x$), carbon monoxide, sulphur dioxide ($SO_2$), and trace metals (Huber et al. 2016). The exhaust of vehicles equipped with catalytic converters that help in reducing air pollution has been identified as the source of Rh, Pd, and Pt (Rauch et al. 2005). The leakage and seepage from the internal combustion (IC) engines emit PAH, whereas petrol refill stations are prominent hotspot sources of BTEX (Liu et al. 2018). Vehicle washing releases various chemicals adhered to the vehicle's top or underbody, while rain falling on the vehicle also performs a similar type of action, and ultimately adds to the contamination of urban stormwater runoff. The tyre and brakes are also the sources of heavy metals such as Zn, Cu, Cd, Ni, and Pb, along with PAH and microplastics with rubber-derived quinones such as phenylenediamines (Cao et al. 2022). The tyres and brakes wear out as material on their application and get deposited over the road with time. The abrasion of road material over a period of time also serves as the source of TSS along with

**Table 1** Various pollutants that result in contamination of wet precipitation as reported by published studies

| Pollutant | Reference |
|---|---|
| Suspended Solids | Pipil et al. 2022; Gunawardena et al. 2013; |
| Nutrients:<br>N, P, Fe, Zn | Baker et al. 2007; Anderson and Downing 2006; |
| Heavy metals:<br>Fe, Cr, Mn, Ni, Cu, Zn, Cd, Pb, Hg | Gunawardena et al. 2013 Bąk et al. 2019; Lynam et al. 2014; |
| Organic Pollutants:<br>Polychlorinated dibenzo-p-dioxins (PCDD) and polychlorinated dibenzofurans (PCDF);<br>Benzene, Toluene, Ethylbenzene, and Xylenes (BTEX) | Mohamed et al. 2020; Kirkok et al. 2020; Yu et al. 2022; |

the deposition of other suspended solids during dry spells. The presence of roadside dust also contributes to TSS in urban stormwater, while heavy metals, PAH, and microplastics were also found in the roadside-suspended solid materials and road dust (Müller et al. 2020). They also get pulverised between the tyre and pavement due to moving heavy vehicles that get dispersed and act as the non-point source of TSS pollution. Microplastics also originated from the pavement material bitumen, and the paint used for road making (Horton et al. 2017). Moreover, the degree of abrasion is higher for concrete pavement than bitumen-asphalt pavement resulting in a higher concentration of heavy metals in road dust. The flushing of the pavement surfaces, roadside barriers, kerbs, delineators, and signage by rainfall takes along the contaminants from their surfaces which were deposited there during the dry season/or spell of no rainfall. The roads having more braking, acceleration, and deceleration contributed more heavy metal concentration along the road as compared to the other locations having similar traffic volume (Huber et al. 2016). Hence, the stormwater runoff gets contaminated with pollutants originating from the operation of vehicles, re-fuelling, traffic volume, and its intensity; pollutants originating from the exhaust, its maintenance; wear and tear of its components, etc.

### 4.3 Industrial Activities

Pollutants are emitted from the various industrial activities taking place in industry involved in production and manufacturing, vehicular movement for supply chain management, open storage units, etc. The pollutants from the industrial area when released into the air, come to the earth's surface as dry deposition under gravity or as wet deposition as a result of a downpour. The washing of pollutants deposited over the ground surface, trees, road services, sheds, etc. remains either in suspension or in dissolved form in the stormwater runoff (Müller et al. 2020). The contribution of pollutants originating from the industry depends upon the type of industry, the technology involved, and its associated activity. The poor handling of raw materials as well as finished products and improper management of their storage in open spaces can also contribute to stormwater runoff contamination (Ahmed et al. 2021). For instance, a study reported the presence of heavy metals like Cr, Co, Ni, Cu, Zn, As, Cd, Mn, Sn, Pb, V in road-side deposited sediments in the intensely industrialised area (Jeong et al. 2020). The storage of industrial raw materials in open spaces releases chemicals that get mixed with stormwater runoff. In addition, the wastewater treatment facilities, their management, improper handling/storage of sludge in open spaces, and their washing by rainfall can also contribute to stormwater runoff pollution. The dust from cement manufacturing industries consists of $NO_x$, $SO_2$, VOCs, mercury, toxic dioxins, furans, and PCBs. These pollutants originate from the raw materials, clinker, pulverisation, and storage and handling and get deposited on the nearby land, tree leaves, and over buildings thus contaminating the stormwater runoff (Mohamad et al. 2021). Literature studies have reported the majority of the pollutants originate from vehicle movement, vehicle service areas, and parking facilities.

### 4.4 Construction Activities

Construction practices vary from the construction of small dwelling units to the development of infrastructures like massive dams and never-ending roads. The activities involved during the construction phase such as cleaning, grading, stock piling near the site, and

excavation are known to produce dust, sand, and silt which when comes in contact with the stormwater, becomes resuspended and mobilises towards the drain and receiving water bodies. These resuspended silt and sand from the construction site carry along the organic matter, metals, oil and grease, nutrients, and bacteria, which mix with the sewer system and are finally discharged into the river or nearby receiving water body (Müller et al. 2020). The disposal of pollutants into the water body is toxic to the aquatic ecosystem thus harming the aquatic life (Bang et al. 2020). During the construction phase, the natural vegetation canopy is stripped and the ground stabilising material is removed which increases the runoff volume, and carries along the soil because of erosion and scouring, contributing it to the receiving water body. The intensity of rainfall, land use, gradient and slope of the land, type of soil, soil properties, etc. are the primary drivers of soil erosion (Gholami et al. 2021). TSS and turbidity are the most documented parameters in stormwater water runoff originating majorly from construction sites all over the globe. In a study conducted by Pipil et al. (2022), the construction activity taking place for the development of new infrastructure under the Central Vista Project in New Delhi, India, has reported high TSS in stormwater runoff when compared to the runoff sample collected from other land use such as residential, industrial and institutional area.

### 4.5 Drain Overflow

Modern cities are developed concerning the future expansion and scope of their inhabitants. Few cities have developed advanced sewage systems for the conveyance of the grey water whereas few other cities in developing nations rely on the conventional sewage system earlier designed to serve for lesser population. India being a developing nation has unauthorised settlements in the form of slums, colonies, clusters, etc. which is a home to a large population of India. The stormwater runoff commencing from these unauthorised localities contributes to total suspended solids (TSS), $BOD_5$, and faecal coliform as these areas do not have a properly laid underground sewage network (Jamwal et al. 2008). They rely upon the open sewage systems which when receiving high-intensity rainfall, tend to overflow and mix with the stormwater runoff causing its contamination.

### 4.6 Gardening and Agricultural

Residents around their dwellings in urban areas prefer to have gardening area maintained by them. Town planners also as a part of town aesthetics and ecological point of view, prefer to provide the green belt and open spaces for recreational activities, to preserve nature, and agricultural activities, etc. In the outskirts of the city, green belts are provided which surround the urban area and are being used for agricultural activities to fulfil the demands and needs of the urban population. These areas are generally applied with nutrients such as phosphate and nitrates to help plants and crops grow more effectively and enhance agricultural yield. Once the infiltration capacity of the gardening area and agricultural field is fulfilled, the rainwater excess becomes stormwater runoff that will carry along the suspended solids, nutrients applied to the field (dominated by nitrogen and phosphate), and other organic matter present in the field. India is a large country that has 3/4 of its population directly or indirectly dependent on agricultural activities and has a significant contribution (≈ 20%) to its gross domestic production (GDP) (Ministry of Agriculture & Farmers Welfare 2022).

The extent of agricultural activity and the production of stormwater runoff in a developing country like India can be understood. The gardening and agricultural activities such as cutting, pruning, chipping, and cropping of leaves and branches produce organic waste that also contains the nutrients P and N, which might get deposited on the ground or nearby area and come in contact with the stormwater runoff leading to deterioration of its quality. Pesticides are being used tremendously in agricultural fields to prevent pests, weeds, and insects. These have the potential to leach out to contaminate groundwater as suggested by field and column studies done by Katagi (2013). Pesticides such as organophosphates, organochlorines including diazinon, bifenthrin, chlordane, DDT (dichlorodiphenyltrichloroethane), DDD (dichlorodiphenyldichloroethane), and DDE (dichlorodiphenyldichloroethylene), and chlorpyrifos have been reported in the surface runoff originating from the fields (Samuel et al. 2012).

## 5 Treatment of Stormwater Runoff

The stormwater runoff cannot be treated at the source from where it is originating, such as buildings, roads, highways, and residential areas, due to limited availability of space. In earlier days, stormwater runoff was considered to be wastewater, and limited emphasis was given to its treatment and reusability (Ekka et al. 2021). Stormwater runoff was removed from the urban area as soon as possible through a network of stormwater drains. In urban areas, stormwater gets contaminated as it traverses its path over various impervious surfaces from different land-use originating from non-point sources. However, various stormwater treatment techniques have been employed for watershed management. The pollutant load of stormwater can be significantly reduced in urban areas by adopting the techniques popularly known as water-sensitive urban design (WSUD) in different configurations including gross pollutant trap (GPT), and bioretention basins such as bioswales, raingarden, and wetlands.

### 5.1 Gross Pollutant Trap (GPT)

The visible and tangible solids arising from the urban environment have been termed as trash, debris, floatable, gross pollutants, rubbish, or solid waste which composes the urban litter and becomes an integral part of the stormwater runoff and thus pose a threat to human health and ecology (Armitage 2007). As stormwater runoff traverses its path along the urban area, it keeps the urban pollutant floating or as a bed load in the conveyance system. Such pollutant loads are not only aesthetically unattractive but are also devastating to the natural equilibrium and ecology of urban areas, and thus significantly reduce the hydraulic performance of urban conveyance systems (Ab Ghani et al. 2011). Moreover, when these floating and suspended pollutants in stormwater runoff merge with natural waterways, become a threat to aquatic life in which fishes can get entangled, suffocate, or ingest the litter while searching for food. Hence, GPT is an engineered trap system designed to remove suspended as well as floating impurities from stormwater runoff via interception and retention of solids present in stormwater runoff. Generally, it can remove the litter of diameter $d > 5.0$ mm, and hence, the clear opening between the screens/mesh for removal of floating and suspended pollutants is greater than 5.0 mm (Madhani and Brown 2015). They are installed on the existing conveyance system with an object to prevent the entry of floating and suspended

impurities into the next treatment unit. Their design demands the careful study of local hydrological processes, catchment area, pollutant load, and size of conveyance system, and depends particularly on the cost of its operation and construction. Many developed countries have adopted this technique for the improvement of the quality of stormwater runoff with different configurations. Table 2 provides the summary of GPT configuration used for improving the quality of stormwater runoff in different locations around the globe.

### 5.2 Bioretention Basins

Bioretention systems are also known as bio-filtration systems comprising of media substrate made up of sand, gravel, slag, loam, etc. which is planted with suitable plant species. Bioretention systems involve bioswales, raingardens, and wetlands whose objective is to reduce the stormwater peak flow through volume attenuation, removal of pollutants, and groundwater recharge using stormwater runoff. In a study conducted by Henderson et al. (2007) bioretention system removed 63 to 77% of nitrogen and 85 to 94% of phosphorus when the vegetated column was used with varying substrate types (gravel, sand, loam). A study conducted by Du et al. (2021) in China showed an average reduction of 20.8–93.3%, 11.6–54.8%, and 9.8–47.3% in TSS, TN, and TP, respectively, from stormwater runoff. Zhang et al. (2011) observed that plant species can remove 59 to 83% and 28 to 71% of N

**Table 2** Different GPT configuration for pollutant removal from stormwater runoff

| Type of GPT configuration | Country | Removal of pollutant | Reference |
|---|---|---|---|
| Hydrodynamic Deflective Separation (HDS) | Malaysia | Plastic, paper, vegetation, sediments | Sidek et al. 2016 |
| Catch Basin Insert | Australia | Vegetation, plastic, newspaper, cardboard, food and drink package, cans, clothes, etc. | Alam et al. 2017 |
| Gross Pollutant Trap with bioretention | Sweden | 100–300 µm sized rubber, bitumen, and microplastics | Lange et al. 2021 |
| Gross Pollutant Trap with bioretention and sand filter | Sweden | 20–100 µm sized microplastics | Lange et al. 2022a |
| Gross Pollutant Trap followed by bioretention system | Sweden | Pb (> 76%), Cu (79%), Zn (94%) when GTP and bioretention treatment train was used | Lange et al. 2022b |
| Gross Pollutant Trap | Australia | TSS, TN and TP removal | Nichols and Lucke 2016 |
| Gross Pollutant Trap | Malaysia | Floatable litter load | Asfi et al. 2023 |
| Gross Pollutant Trap with biofilter and sand filter | Sweden | polycyclic aromatic hydrocarbons (PAH), petroleum hydrocarbons (PHC), nonylphenol (NP), 4-t-octylphenol (OP), bisphenol A (BPA), | Beryani et al. 2023 |

and P, respectively, and suggested that the removal efficiency of pollutants in a bioretention system largely depends upon the substrate composition, type of vegetation, and symbiotic relationship with bacteria.

### 5.2.1 Bioswales

One of the oldest and most commonly implemented in-situ stormwater technique is bioswales. They are also known as vegetated swales, grassed swales, bio-filters, and filter strips (Ahiablame et al. 2012). It is a technique for the removal of pollutants, improving stormwater runoff quality, and flow volume reduction, and it is a step towards making urban cities more resilient. It involves several processes such as infiltration, filtration, sedimentation, and biological processes through which it treats stormwater runoff (Stagge et al. 2012). The bioswales consist of trenches that are provided with engineered media for infiltration of stormwater runoff and macrophytes for the improved removal efficiency of sediments, nutrients, and heavy metals. The macrophytes or vegetation provided are native to the location such that they can survive during the dry season and help in landscaping (Pineau et al. 2021). Non-native species can be used in bioswales and bioretention systems as long as they are resistant to climatic conditions. Also, while drought is a limit, in other climates, winter can have a much greater impact on plant survival than drought (Pineau et al. 2021). It thus, helps in enhancing the urban biodiversity. The stormwater treatment efficiency of bioswales largely depends upon the healthy vegetation (Mazer et al. 2001). A study conducted by Backstrom (2003) in Sweden, swale removed 70% of TSS from stormwater runoff. Deletic and Fletcher (2006) conducted a study in Australia with a 65 m long swale at a longitudinal slope of 1.6% and observed that average TSS, TN, and TP removal rate was of order 69%, 56%, and 46%, respectively. In a study undertaken by Ismail et al. (2010) in Malaysia, demonstrated upto 96% TSS removal from stormwater runoff and suggested that the removal efficiency of bioswales depends upon substrate particle size, vegetation density, and surface slope. Table 3 illustrates the different studies that reported towards removal of pollutants using bioswales.

### 5.2.2 Raingarden

Raingarden is a stormwater management system that treats stormwater runoff through filtration via gravel media, sorption, and denitrification if anaerobic conditions are present at the bottom. Raingarden is a low-lying land that consists of gravel media whose size increases from top to bottom that helps in the physical removal of floating and suspended impurities, and they are planted with native vegetation to sustain in the local climate. The processes and mechanisms that take place in raingarden include evapotranspiration, exfiltration, infiltration, adsorption, and cation exchange for the removal of pollutants from stormwater runoff. At the bottom of the raingarden, the perforated pipe collects the partially treated stormwater runoff and sends it to the next treatment unit. The ponding in the raingarden provides the detention to stormwater runoff which is treated by both native and non-native vegetation to remove dissolved impurities through sorption, moreover, it helps in the beautification of the landscape, infiltration of treated stormwater runoff, replenishment of groundwater, flood management and peak flow attenuation (Pineau et al. 2021). In a study undertaken by Du et al. (2021), raingardens have achieved an average reduction of 20.8–93.3%, 11.6–54.8%,

**Table 3** Pollutant removal study from stormwater runoff by different bioretention systems

| Type of bioretention system | Pollutant removal efficiency | Study Location | Reference |
|---|---|---|---|
| Bioswales | 90%, 45%, 38%, and 80% removal of TSS, TN, TP, and Zn, respectively. | USA | Stagge 2006 |
| | 70% and 66% removal of TSS and Zn, respectively. | Sweden | Backstrom 2003 |
| | Average removal varied between 57 to 88%, 40 to 72%, and 12 to 67%, respectively for TSS, TN and TP. | Australia | Fletcher et al. 2002 |
| | 83.5% and 81.3% removal of TSS and TP, respectively. | China | Zhao et al. 2016 |
| | TSS and Zn removal varied between 65 to 71% and 30 to 60%, respectively. | USA | Stagge and Davis 2006 |
| Raingarden | > 92% heavy metals; 80% P; 60–80% ammonium; 65–75% TKN removal | USA | Davis et al. 2001 |
| | Higher removal of $NH_3$, TN than TP, TKN, $NO_3^-$, organic-N.Removal of Cu, Pb, Zn > 5, 5, 10 µg/L, respectively | USA | USA |
| | 90% removal of Naphthalene toluene | USA | Hong et al. 2006 |
| | 85% and 65% removal of Phosphorous | USA | Hsieh et al. 2007 |
| | 89–100% $PO_4^{3-}$, 84–100% atrazine, 42–63% removal of nitrate. | USA | Yang et al. 2010 |
| | Up to 87% $NO_3^-$, major reduction in P, lesser reduction in TSS, Zn. | New Zealand | Trowsdale and Simcock 2011 |
| | System with top-soil show 69.0%, 71.4% removal for Cu, Zn respectively;Sand and sand-top soil mix removed: Cu upto 83.3%, Zn upto 94.5%, Pb upto 97.3% | New Zealand | Good et al. 2012 |
| | Nitrate 91% removal, $PO_4^{3-}$ 99% removal, Atrazine 90% removal | USA | Yang et al. 2013 |
| | Cd: >0.2–2.5 mg/kg;<br>Cu: 3–290 mg/kg;<br>Pb: <1–310 mg/kg;<br>Zn: 11–3,900 mg/kg | Australia | Al-Ameri et al. 2018 |
| | 52–89% removal of TN | Colombia | Fajardo-Herrera et al. 2019 |
| | Average reduction in $PO_4^{3-}$ is 93.6 to 96%<br>Average reduction TP is 77 to 90.1% | USA | Wadzuk et al. 2021 |
| | 59% TSS removal;<br>79% TP removal;<br>69% Zn removal | USA | Davis 2007 |
| | Approximately 84%, 35%, 31% and 32% reduction of $NH_3$, $NO_3^-$, TKN, and TN, respectively. | USA | Dietz and Clausen 2005 |
| | 43%, 70%, 64%, and 87% reduction in Cu, Pb, Zn, and Phosphorus, respectively. | USA | Chisholm 2008 |
| | Average reduction of 20.8–93.3%, 11.6–54.8%, and 9.8–47.3% in TSS, TN, and TP, respectively. | China | Du et al. 2021 |

and 9.8–47.3% in TSS, TN, and TP, respectively. Trowsdale and Simcock (2011) conducted a study on raingarden using *Pennisetum clandestinum*, a sub-tropical grass in New Zealand, revealing $NO_3^-$ removal up to 87% but lesser reduction in TSS and Zn. Also, another study conducted in New Zealand by Good et al. (2012), reported heavy metal reduction for Cu upto 83.3%, Zn upto 94.5%, and Pb upto 97.3% using raingardens. Raingardens have also been found efficient in nutrient removal. A study conducted in USA using raingarden columns showed 89–100% $PO_4^{3-}$ and 42–63% nitrate removal (Yang et al. 2010). Table 3 summarises the various bioretention systems towards pollutant removal from stormwater runoff.

## 6 Contaminant Removal in Bioretention System

The selection of appropriate plant species, and their specific roles in stormwater treatment, is a significant aspect of bioretention system design. Plants, especially those with dense root systems, can absorb and utilize nutrients (Pipil et al. 2021; Jain et al. 2024). The property to assimilate or uptake nutrients differs significantly from species to species. In a study, *Typha latifolia*, *Juncus effusus*, and *Cyperus papyrus* were investigated for phosphorus and nitrogen removal capabilities (Bebba et al. 2019). *Juncus effusus* was found as the maximum nitrogen removal efficient (95%) and *Typha latifolia* as the best phosphorous removal (95%), thus displaying the maximum elimination rates. Whereas, a study by Keizer-Vlek et al. (2014) observed around 92% removal efficiency for phosphorous using *Iris pseudacorus*. Five mono-cultures, *Agrostis alba* (red top), *Canna generalis* 'Firebird' (canna lily), *Carex stricta* (tussock sedge), *Iris ensata* 'Rising Sun' (Japanese water iris), *Panicum virgatum* (switchgrass) were assessed towards the treatment of runoff from commercial nurseries (Spangler et al. 2019). *Panicum virgatum* was found most efficient and removed up to 64.7% phosphorous and up to 82.4% nitrogen, while other species remediation rates ranged from 25 to 50% towards removal of phosphorous and nearly 13 −50% removal of nitrogen.

The effectiveness of metal removal in bioretention depends on the type of metal, plant species metal tolerance, accumulation potential, and ability to thrive in the local environment as well as the sorption potential of growth substrate. In the bioretention system, the sorption potential of growth substrate plays a major role in the removal of metals from stormwater whereas plants play a secondary role, exhibiting different metals/metalloid uptake mechanisms. The latter improves the quality of stormwater via processes such as phytoextraction, rhizofiltration, phytodegradation, phytostabilization, etc. (Dagenais et al. 2018). The efficiency of *Dracaena marginata* in a bioretention system was evaluated against a non-vegetated system for the treatment of stormwater containing various metals like Pb, Mg, Ca, Cu, Fe, Cd, Al, Cr, K, Na, Ni, and Zn (Vijayaraghavan and Praveen 2016). The presence of vegetation significantly enhanced the overall removal performance of metals with a remediation efficiency of 99.8% for all the metals post 245 mm of rainfall. Plants enhance the microbial activity in the soil, endorsing processes like metal reduction, precipitation, or transformation (Hsieh and Davis 2005). Micro-organisms in association with plant roots, interact with metals, causing immobilization or transformation of metals into less toxic forms. Plants secrete compounds called chelators or exudates from their roots, which bind to metals and form complexes that are least toxic, and highly stable, thus improving the bioavailability as well as mobility of metals for enhanced plant uptake from the soil (Agar-

wal et al. 2023). Nevertheless, vegetation also boosts biodiversity, providing habitat for wildlife, and promoting the ecological health as well as aesthetics of urban environments.

## 7 Bioretention Media

Bioretention media are components, specially designed to manage and treat stormwater runoff by facilitating filtration, infiltration, pollutant removal, and ascertaining the stability of the bioretention systems. Conventional bioretention media are typically composed of soil, sand, and organic matter mixture which removes suspended solids, and dissolved contaminants, and supports macrophytes and healthy root zone (Brown and Hunt 2011). Conventional media reported effective remediation of stormwater runoff via either sorption or filtering impurities, suspended solids, sediments, heavy metals, nutrients (nitrogen and phosphorus), and organic matter (Shrestha et al. 2018). This helps improve water quality as runoff passes through the bioretention system.

Soil acts as a filter in bioretention systems and helps to bind and retain nutrients especially nitrogen and phosphorous when stormwater runoff flows through it. Nutrients are either retained or captured through processes like adsorption, ion exchange, or indirectly by altering the soil biome including the soil microbial activity (Dagenais et al. 2018). The sediments settle, trap, and accumulate the nutrients within the sediment layer, thus reducing the nutrient load. The efficiency of nutrient removal in bioretention systems is governed by factors such as hydraulic loading rates, vegetation health, soil conditions, and maintenance practices. Organic matter enhances the nutrient-holding capacity and microbial activity. Different studies have reported the effectiveness of additive substrates in bioretention systems including fly ash, biochar, vermiculite, zero-valent iron, zeolite, and others (Sang et al. 2019). The use of organic matter additives like compost increases the water-holding capacity as well as improves nutrient retention, whereas peat moss enhances the cation exchange capacity of media (Sharma and Malaviya 2021).

## 8 SWOT Analysis of Treatment Methods

The current study focuses into the management and treatment of stormwater runoff in urban areas of developing nations through the application of different stormwater runoff treatment techniques as discussed above. The environmental performance of various treatment methods could be hampered because of certain factors regulating the treatment and pollutant removal efficiency from stormwater runoff. Considering the various parameters and factors governing the treatment units towards the applicability, planning, and execution of these stormwater runoff treatment techniques, the strength, weakness, opportunity, and threat (SWOT) analysis has been illustrated in Table 4.

**Table 4** SWOT analysis of various stormwater treatment units

| Type of treatment unit | Strength | Weakness | Opportunity | Threat |
|---|---|---|---|---|
| Gross Pollutant Trap (GPT) | • Eco-friendly<br>• Works under the action of gravity;<br>• Can be retrofitted in existing stormwater drains;<br>• Work as a pre-treatment unit | • Doesn't not remove particle sizes less than 5.0 mm;<br>• Requires regular maintenance;<br>• Poor maintenance leads to foul smell;<br>• Limited removal of dissolved impurities | • Transfer of technology from developed nation to developing nation (SDG 9)<br>• Design criteria can be developed as a code of practice; | • Perilous to the person cleaning it;<br>• Resuspension and remobilisation of removed impurities; |
| Bioswales | • Low cost of construction;<br>• Augments the biodiversity;<br>• Helps in reducing the peak runoff flow;<br>• Infiltration from unlined sides helps in groundwater recharge;<br>• Sustainable technique since no energy is required;<br>• Better aesthetic than gutter and kerbs; | • Poor hydraulic conductivity<br>• Fails to treat stormwater runoff during high-intensity rainfall;<br>• Steep slopes require small check dams; | • Developing countries lack such a technique thus providing more scope for its application;<br>• Different vegetation can be used that are resistant to dry season;<br>• Proper design can store larger stormwater runoff volume which can delay peak flow volume; | • Wilting of vegetation during the dry period;<br>• During high intensity, stormwater runoff can bypass them;<br>• Earthen check dam may erode with the flow;<br>• Can be a mosquito breeding ground if the correct slope is not provided;<br>• Can produce bad odour; |
| Raingardens | • Sustainable treatment process since it does not require energy;<br>• Cheaper to construct;<br>• Increases groundwater level;<br>• Augment biodiversity;<br>• Prevent flood through runoff detention;<br>• Specialised supervision is not required;<br>• Easy to operate and maintain;<br>• Can be installed in smaller land areas; | • Poor performance of media due to its clogging;<br>• Cannot effectively handle larger stormwater runoff volume originating from catchment;<br>• Less effective in controlling floods;<br>• It requires land which is expensive in urban areas;<br>• Takes time to establish fully; | • Introduction of new species which can survive in dry season;<br>• Can be installed in family houses;<br>• Plants other than wetland vegetation can be used; | • Wilting of vegetation during dry season;<br>• Can be a mosquito breeding site;<br>• Stormwater runoff can bypass raingardens during high discharge as a result of high-intensity rainfall;<br>• Filter media may clog; |

## 9 Conclusion

Based on the current review of the literature, it can be concluded that stormwater management and its treatment methods are sustainable and effective solutions for pollutant removal and increasing the water capacity of urban areas. However, the economic analysis of the infrastructure and treatment cost that can occur towards the treatment of stormwater runoff is rare and limited. If the economic analysis can be provided, then cost-benefit analysis, life cycle cost, break-even period, and net asset value can be calculated for its application in developing nations where the shortage of funds remains for capital investment. The criti-

cality, outcomes, and significance during the development, maintenance, and operation of various treatment techniques by developed nations can be shared with developing nations to overcome unforeseen difficulties through the transfer of techniques. Stormwater runoff simulation models can be adopted for predicting the flow volume against which the infrastructure of optimum capacity can be developed through model testing and validation with real-world problems. Technical aspects such as design parameters, removal of pollutants, type of filter material, vegetation type, etc. have been addressed well in the literatures. However, it is recommended to include social, geological, and economic aspects along with technical aspects of stormwater treatment and management. It is recommended to conduct more studies in developing nations as these studies are limited in such countries. Also, implementing such methods and techniques for stormwater runoff management can help in attaining United Nations (UN) Sustainable Development Goals (SDG), such as SDG 6 (Clean Water and Sanitation), SDG 11 (Sustainable Cities and Communities), and SDG 14 (Life Below Water).

**Acknowledgements** The authors are grateful to Department of Environmental Engineering at DTU for providing the opportunity to write this review paper.

**Author Contributions** Shivani Yadav: Conceptualization, compilation and draft writing and editing.
Saurav Ambastha: Conceptualization, compilation and draft writing and editing.
Harsh Pipil: Conceptualization, compilation and draft writing and editing.
Anil Kumar Haritash: Conceptualization, editing, supervision, and reviewing.
Krishna R. Reddy: Conceptualization, supervision, and reviewing.

**Funding** No funding was received for current study.

## Declarations

**Conflict of interest** The authors have no conflicts of interest to declare that are relevant to the content of this article.

**Human and Animal Rights** Authors declare that no Human and Animals were involvement in the research work.

## References

Ab Ghani A, Azamathulla HM, Lau TL, Ravikanth CH, Zakaria NA, Leow CS, Yusof MAM (2011) Flow pattern and hydraulic performance of the REDAC gross pollutant trap. Flow Meas Instrum 22(3):215–224. https://doi.org/10.1016/j.flowmeasinst.2011.02.004

Adeyanju AA, Manohar K (2017) Effects of vehicular emission on environmental pollution in Lagos. Sci-Afric J Sci Issues Res Essays 5(4):34–51

Agarwal P, Vibhandik R, Agrahari R, Daverey A, Rani R (2023) Root exudates on the soil microbial diversity and biogeochemistry of heavy metals. Appl Biochem Biotechnol: 1–21. https://doi.org/10.1007/s12010-023-04465-2

Ahiablame LM, Engel BA, Chaubey I (2012) Effectiveness of low impact development practices: literature review and suggestions for future research. Water Air Soil Pollut 223:4253–4273. https://doi.org/10.1007/s11270-012-1189-2

Ahmed J, Thakur A, Goyal A (2021) Industrial Wastewater and Its Toxic Effects. In: Shah MP (ed) Biological Treatment of Industrial Wastewater, The Royal Society of Chemistry, pp 1–14. https://doi.org/10.1039/9781839165399-00001

Al-Ameri M, Hatt B, Le Coustumer S, Fletcher T, Payne E, Deletic A (2018) Accumulation of heavy metals in stormwater bioretention media: A field study of Temporal and Spatial variation. J Hydrol 567:721–731. https://doi.org/10.1016/j.jhydrol.2018.03.027

Al-Khashman OA (2005) Ionic composition of wet precipitation in the Petra region. Jordan Atmospheric Res 78(1–2):1–12. https://doi.org/10.1016/j.atmosres.2005.02.003

Alam MZ, Anwar AF, Sarker DC, Heitz A, Rothleitner C (2017) Characterising stormwater gross pollutants captured in catch basin inserts. Sci Total Environ 586:76–86. https://doi.org/10.1016/j.scitotenv.2017.01.210

Anderson KA, Downing JA (2006) Dry and wet atmospheric deposition of nitrogen, phosphorus and silicon in an agricultural region. Water Air Soil Pollut 176(1):351–374. https://doi.org/10.1007/s11270-006-9172-4

Armitage N (2007) The reduction of urban litter in the stormwater drains of South Africa. Urban Water J 4(3):151–172. https://doi.org/10.1080/15730620701464117

Asfi MF, Razak SBA, Zulkifli MH, Sharip Z (2023) Spatio-temporal factors influencing coliform and *Escherichia coli* contamination in gross pollutant trap and wetland Inlet, Putrajaya, Malaysia. Int J Energy Water Resour 7:487–498. https://doi.org/10.1007/s42108-022-00233-w

Backstrom M (2003) Grassed Swales for stormwater pollution control during rain and snowmelt. Water Sci Technol 48(9):123–132. https://doi.org/10.2166/wst.2003.0508

Bąk Ł, Szeląg B, Sałata A, Studziński J (2019) Modeling of heavy metal (Ni, Mn, Co, Zn, Cu, Pb, and Fe) and PAH content in stormwater sediments based on weather and physico-geographical characteristics of the catchment-data-mining approach. Water 11(3):626. https://doi.org/10.3390/w11030626

Baker AR, Weston K, Kelly SD, Voss M, Streu P, Cape JN (2007) Dry and wet deposition of nutrients from the tropical Atlantic atmosphere: links to primary productivity and nitrogen fixation. Deep Sea Res Part I 54(10):1704–1720. https://doi.org/10.1016/j.dsr.2007.07.001

Bang KW, Joo JC, Kim JH, Kang E, Choi J, Lee JM, Kim Y (2020) Application of bottom Ash as filter media for construction site runoff control. Water 12(4):990. https://doi.org/10.3390/w12040990

Bebba AA, Labed I, Zeghdi S, Messaitfa A (2019) Purification performance of *Typha latifolia*, *Juncus effusus* and *Papyrus cyperus* in arid climate: influence of seasonal variation. J Water Chem Technol 41:396–401. https://doi.org/10.3103/S1063455X19060092

Beryani A, Flanagan K, Viklander M, Blecken GT (2023) Performance of a gross pollutant Trap-Biofilter and sand filter treatment train for the removal of organic micropollutants from highway stormwater (Field Study). Sci Total Environ 900:165734. https://doi.org/10.1016/j.scitotenv.2023.165734

Björklund K, Bondelind M, Karlsson A, Karlsson D, Sokolova E (2018) Hydrodynamic modelling of the influence of stormwater and combined sewer overflows on receiving water quality: benzo (a) pyrene and copper risks to recreational water. J Environ Manage 207:32–42. https://doi.org/10.1016/j.jenvman.2017.11.014

Brown RR, Farrelly MA (2009) Challenges ahead: social and institutional factors influencing sustainable urban stormwater management in Australia. Water Sci Technol 59(4):653–660. https://doi.org/10.2166/wst.2009.022

Brown RA, Hunt WF (2011) Impacts of media depth on effluent water quality and hydrologic performance of undersized bioretention cells. J Irrigat Drain Eng 137(3):132–143. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000167

Cao G, Wang W, Zhang J, Wu P, Zhao X, Yang Z et al (2022) New evidence of rubber-derived Quinones in water, air, and soil. Environ Sci Technol 56(7):4142–4150. https://doi.org/10.1021/acs.est.1c07376

Chisholm H (2008) An analysis of the efficacy of rain gardens for the protection of water resources in Annapolis Royal, NS. Dissertation, University of Guelph

Dagenais D, Brisson J, Fletcher TD (2018) The role of plants in bioretention systems; does the science underpin current guidance? Ecol Eng 120:532–545. https://doi.org/10.1016/j.ecoleng.2018.07.007

Davis AP (2007) Field performance of bioretention: water quality. Environ Eng Sci 24(8):1048–1064. https://doi.org/10.1089/ees.2006.0190

Davis AP, Shokouhian M, Sharma H, Minami C (2001) Laboratory study of biological retention for urban stormwater management. Water Environ Res 73(1):5–14. https://doi.org/10.2175/106143001X138624

Deletic A, Fletcher TD (2006) Performance of grass filters used for stormwater treatment—a field and modelling study. J Hydrol 317(3–4):261–275. https://doi.org/10.1016/j.jhydrol.2005.05.021

Dietz ME, Clausen JC (2005) A field evaluation of rain garden flow and pollutant treatment. Water Air Soil Pollut 167:123–138. https://doi.org/10.1007/s11270-005-8266-8

Dietz ME, Clausen JC (2006) Saturation to improve pollutant retention in a rain garden. Environ Sci Technol 40(4):1335–1340. https://doi.org/10.1021/es051644f

Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM (2021) How to conduct a bibliometric analysis: an overview and guidelines. J Bus Res 133:285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Du YE, Hou JM, Ma HL, Liu QC, Wang XH, Zhang ZA et al (2021) Spatial pattern optimization of LID facility based on SWMM. China Water Wastewater 37(19):120–125

Ekka SA, Rujner H, Leonhardt G, Blecken GT, Viklander M, Hunt WF (2021) Next generation Swale design for stormwater runoff treatment: A comprehensive approach. J Environ Manage 279:111756. https://doi.org/10.1016/j.jenvman.2020.111756

Ezeh AC, Bongaarts J, Mberu B (2012) Global population trends and policy options. Lancet 380(9837):142–148. https://doi.org/10.1016/S0140-6736(12)60696-5

Fajardo-Herrera RJ, Valdelamar-Villegas JC, Mouthon Bello J (2019) A rain garden for nitrogen removal from storm runoff in tropical cities. Revista De Ciencias Ambientales 53(2):132–146. https://doi.org/10.15359/rca.53-2.7

Fletcher TD, Peljo L, Fielding J, Wong TH, Weber T (2002) The performance of vegetated swales for urban stormwater pollution control. In: Global Solutions for Urban Drainage, pp 1–16. https://doi.org/10.1061/40644(2002)51

Gholami V, Sahour H, Amri MAH (2021) Soil erosion modeling using erosion pins and artificial neural networks. CATENA 196:104902. https://doi.org/10.1016/j.catena.2020.104902

Ghosh S, Vittal H, Sharma T, Karmakar S, Kasiviswanathan KS, Dhanesh Y et al (2016) Indian summer monsoon rainfall: implications of contrasting trends in the Spatial variability of means and extremes. PLoS ONE 11(7):0158670. https://doi.org/10.1371/journal.pone.0158670

Gikas P, Angelakis AN (2009) Water resources management in Crete and in the Aegean Islands, with emphasis on the utilization of non-conventional water sources. Desalination 248(1–3):1049–1064. https://doi.org/10.1016/j.desal.2008.10.021

Gonçalves FLT, Andrade MF, Forti MC, Astolfo R, Ramos MA, Massambani O, Melfi AJ (2003) Preliminary Estimation of the rainfall chemical composition evaluated through the scavenging modeling for northeastern Amazonian region (Amapa State, Brazil). Environ Pollut 121(1):63–73. https://doi.org/10.1016/S0269-7491(02)00209-9

Good JF, O'Sullivan AD, Wicke D, Cochrane TA (2012) Contaminant removal and hydraulic conductivity of laboratory rain garden systems for stormwater treatment. Water Sci Technol 65(12):2154–2161. https://doi.org/10.2166/wst.2012.135

Gunawardena J, Egodawatta P, Ayoko GA, Goonetilleke A (2013) Atmospheric deposition as a source of heavy metals in urban stormwater. Atmos Environ 68:235–242. https://doi.org/10.1016/j.atmosenv.2012.11.062

Henderson C, Greenway M, Phillips I (2007) Removal of dissolved nitrogen, phosphorus and carbon from stormwater by biofiltration mesocosms. Water Sci Technol 55(4):183–191. https://doi.org/10.2166/wst.2007.108

Hogan DM, Walbridge MR (2007) Best management practices for nutrient and sediment retention in urban stormwater runoff. J Environ Qual 36(2):386–395. https://doi.org/10.2134/jeq2006.0142

Hong E, Seagren EA, Davis AP (2006) Sustainable oil and grease removal from synthetic stormwater runoff using bench-scale bioretention studies. Water Environ Res 78(2):141–155. https://doi.org/10.2175/106143005X89607

Horton AA, Svendsen C, Williams RJ, Spurgeon DJ, Lahive E (2017) Large microplastic particles in sediments of tributaries of the river Thames, UK–Abundance, sources and methods for effective quantification. Mar Pollut Bull 114(1):218–226. https://doi.org/10.1016/j.marpolbul.2016.09.004

Hsieh YP, Davis AP (2005) Effects of vegetation on soil microbial community structure in constructed wetlands. Ecol Eng 25(1):73–88

Hsieh CH, Davis AP, Needelman BA (2007) Bioretention column studies of phosphorus removal from urban stormwater runoff. Water Environ Res 79(2):177–184. https://doi.org/10.2175/106143006X111745

Huber M, Welker A, Helmreich B (2016) Critical review of heavy metal pollution of traffic area runoff: occurrence, influencing factors, and partitioning. Sci Total Environ 541:895–919. https://doi.org/10.1016/j.scitotenv.2015.09.033

Ismail AF, Sapari N, Wahab MMA (2010) Treatment of stormwater runoff from construction sites by vegetative swale. In: World Engineering, Science and Technology Congress (ESTCON), Kuala Lumpur, pp 1–5

Jain N, Yadav S, Taneja S, Ray S, Haritash AK, Pipil H (2024) Phosphate removal from urban stormwater runoff using *Canna Lily* and *Cyperus alternifolius*-based bioretention system. Sustainable Water Resour Manage 10(2):1–10. https://doi.org/10.1007/s40899-024-01076-5

Jamwal P, Mittal AK, Mouchel JM (2008) Effects of urbanisation on the quality of the urban runoff for Delhi watershed. Urban Water J 5(3):247–257. https://doi.org/10.1080/15730620701780348

Jeong H, Choi JY, Lee J, Lim J, Ra K (2020) Heavy metal pollution by road-deposited sediments and its contribution to total suspended solids in rainfall runoff from intensive industrial areas. Environ Pollut 265:115028. https://doi.org/10.1016/j.envpol.2020.115028

Karami M, Sarvestan R, Kleidorfer M (2025) Forecasting and Simulation of Urban Floods Based on the Integration of RegCM and Stormwater Management Models. Water Resour Manage. https://doi.org/10.1007/s11269-025-04133-6

Katagi T (2013) Soil column leaching of pesticides. In: Whitacre D (ed) Reviews of environmental contamination and toxicology volume 221. Springer, New York, NY, pp 1–105. https://doi.org/10.1007/978-1-4614-4448-0_1

Keizer-Vlek HE, Verdonschot PFM, Verdonschot RCM, Dekkers D (2014) The contribution of plant uptake to nutrient removal by floating treatment wetlands. Ecol Eng 73:684–690. https://doi.org/10.1016/j.ecoleng.2014.09.081

Kirkok SK, Kibet JK, Kinyanjui TK, Okanga FI (2020) A review of persistent organic pollutants: Dioxins, furans, and their associated nitrogenated analogues. SN Appl Sci 2:1729. https://doi.org/10.1007/s42452-020-03551-y

Kumar A, Bali K, Singh S, Naja M, Mishra AK (2019) Estimates of reactive trace gases (NMVOCs, CO and NOx) and their Ozone forming potentials during forest fire over Southern Himalayan region. Atmos Res 227:41–51. https://doi.org/10.1016/j.atmosres.2019.04.028

Lange K, Magnusson K, Viklander M, Blecken GT (2021) Removal of rubber, bitumen and other microplastic particles from stormwater by a gross pollutant trap-bioretention treatment train. Water Res 202:117457. https://doi.org/10.1016/j.watres.2021.117457

Lange K, Österlund H, Viklander M, Blecken GT (2022a) Occurrence and concentration of 20–100 Mm sized microplastic in highway runoff and its removal in a gross pollutant trap–Bioretention and sand filter stormwater treatment train. Sci Total Environ 809:151151. https://doi.org/10.1016/j.scitotenv.2021.151151

Lange K, Viklander M, Blecken GT (2022b) Investigation of intra-event variations of total, dissolved and truly dissolved metal concentrations in highway runoff and a gross pollutant trap–bioretention stormwater treatment train. Water Res 216:118284. https://doi.org/10.1016/j.watres.2022.118284

Liu A, Hong N, Zhu P, Guan Y (2018) Understanding benzene series (BTEX) pollutant load characteristics in the urban environment. Sci Total Environ 619:938–945. https://doi.org/10.1016/j.scitotenv.2017.11.184

Liu Y, Wang H, Guan X et al (2025) Urban flood depth prediction and visualization based on the XGBoost-SHAP model. Water Resour Manage 39:1353–1375. https://doi.org/10.1007/s11269-024-04020-6

Lynam MM, Dvonch JT, Hall NL, Morishita M, Barres JA (2014) Spatial patterns in wet and dry deposition of atmospheric mercury and trace elements in central Illinois. USA Environ Sci Pollution Res 21(6):4032–4043. https://doi.org/10.1007/s11356-013-2011-4

Madhani JT, Brown RJ (2015) The capture and retention evaluation of a stormwater gross pollutant trap design. Ecol Eng 74:56–59. https://doi.org/10.1016/j.ecoleng.2014.09.074

Markiewicz A, Björklund K, Eriksson E, Kalmykova Y, Strömvall AM, Siopi A (2017) Emissions of organic pollutants from traffic and roads: priority pollutants selection and substance flow analysis. Sci Total Environ 580:1162–1174. https://doi.org/10.1016/j.scitotenv.2016.12.074

Mazer G, Booth D, Ewing K (2001) Limitations to vegetation establishment and growth in biofiltration Swales. Ecol Eng 17(4):429–443. https://doi.org/10.1016/S0925-8574(00)00173-7

Migliavacca D, Teixeira EC, Wiegand F, Machado ACM, Sanchez J (2005) Atmospheric precipitation and chemical composition of an urban site, Guaiba hydrographic basin, Brazil. Atmos Environ 39(10):1829–1844. https://doi.org/10.1016/j.atmosenv.2004.12.005

Ministry of Agriculture & Farmers Welfare (2022) Contribution of Agriculture Sector towards GDP. https://www.pib.gov.in/PressReleasePage.aspx?PRID=1909213. Accessed 30 January 2022

Mishra S, Bharagava RN, More N, Yadav A, Zainith S, Mani S, Chowdhary P (2019) Heavy metal contamination: an alarming threat to environment and human health. In: Sobti RC, Arora NK, Kothari R (eds) Environmental biotechnology: for sustainable future. Springer, Singapore, pp 103–125. https://doi.org/10.1007/978-981-10-7284-0_5

Mohamad N, Muthusamy K, Embong R, Kusbiantoro A, Hashim MH (2021) Environmental impact of cement production and solutions: A review. Mater Today: Proc 48(4):741–746. https://doi.org/10.1016/j.matpr.2021.02.212

Mohamed K, Majid MIA, Leong YH, Li X (2020) Dioxins in peat and its formation: an overview. Cogent Environ Sci 6(1):1864870. https://doi.org/10.1080/23311843.2020.1864870

Müller A, Österlund H, Marsalek J, Viklander M (2020) The pollution conveyed by urban runoff: a review of sources. Sci Total Environ 709:136125. https://doi.org/10.1016/j.scitotenv.2019.136125

Nichols P, Lucke T (2016) Field evaluation of the nutrient removal performance of a gross pollutant trap (GPT) in Australia. Sustainability 8(7):669. https://doi.org/10.3390/su8070669

Pineau B, Brodeur-Doucet C, Corrivault-Gascon J, Arjoon D, Lessard P, Pelletier G, Duchesne S (2021) Performance of green infrastructure for storm water treatment in cold climate (Canada). J Environ Eng Sci 16(4):185–194

Pipil H, Haritash AK, Reddy KR (2021) Seasonal variability and kinetics of phosphate removal in a *Phragmites*-based engineered wetland. Rend Lincei Sci Fis Nat 32:729–735. https://doi.org/10.1007/s12210-021-01017-w

Pipil H, Haritash AK, Reddy KR (2022) Spatio-temporal variations of quality of rainwater and stormwater and treatment of stormwater runoff using sand–gravel filters: case study of Delhi, India. Rend Lincei Scienze Fis E Naturali 33:135–142. https://doi.org/10.1007/s12210-021-01038-5

Rauch S, Hemond HF, Barbante C, Owari M, Morrison GM, Peucker-Ehrenbrink B, Wass U (2005) Importance of automobile exhaust catalyst emissions for the deposition of platinum, palladium, and rhodium in the Northern hemisphere. Environ Sci Technol 39(21):8156–8162. https://doi.org/10.1021/es050784m

Samuel MP, Senthilvel S, Mathew AC (2012) Water quality assessment of various forms of rainwater and statistical studies on physico-chemical characteristics of stormwater in Coimbatore, India. Nat Environ Pollut Technol 11(1):23–28

Sang M, Huang M, Zhang W, Che W, Sun H (2019) A pilot bioretention system with commercial activated carbon and river sediment-derived Biochar for enhanced nutrient removal from stormwater. Water Sci Technol 80(4):707–716. https://doi.org/10.2166/wst.2019.310

Sharma R, Malaviya P (2021) Management of stormwater pollution using green infrastructure: the role of rain gardens. Wiley Interdisciplinary Reviews: Water 8(2):1507. https://doi.org/10.1002/wat2.1507

Shrestha P, Hurley SE, Wemple BC (2018) Effects of different soil media, vegetation, and hydrologic treatments on nutrient and sediment removal in roadside bioretention systems. Ecol Eng 112:116–131. https://doi.org/10.1016/j.ecoleng.2017.12.004

Sidek L, Basri H, Lee LK, Foo KY (2016) The performance of gross pollutant trap for water quality preservation: a real practical application at the Klang Valley, Malaysia. Desalination Water Treat 57(52):24733–24741. https://doi.org/10.1080/19443994.2016.1145599

Spangler JT, Sample DJ, Fox LJ, Albano JP, White SA (2019) Assessing nitrogen and phosphorus removal potential of five plant species in floating treatment wetlands receiving simulated nursery runoff. Environ Sci Pollut Res 26:5751–5768. https://doi.org/10.1007/s11356-018-3964-0

Stagge JH (2006) Field evaluation of hydrologic and water quality benefits of grass swales for managing highway runoff. Dissertation, University of Maryland

Stagge JH, Davis AP (2006) Water Quality Benefits of Grass Swales in Managing Highway Runoff. In: Water Environment Federation, 2006(6):5518–5527. https://doi.org/10.2175/193864706783775702

Stagge JH, Davis AP, Jamil E, Kim H (2012) Performance of grass Swales for improving water quality from highway runoff. Water Res 46(20):6731–6742. https://doi.org/10.1016/j.watres.2012.02.037

Trowsdale SA, Simcock R (2011) Urban stormwater treatment using bioretention. J Hydrol 397(3–4):167–174. https://doi.org/10.1016/j.jhydrol.2010.11.023

United States Environmental Protection Agency (1983) Results of the nationwide urban runoff program: volume I. Water Planning Division, United States Environmental Protection Agency, Washington, DC

Vijayaraghavan K, Praveen RS (2016) *Dracaena marginata* biofilter: design of growth substrate and treatment of stormwater runoff. Environ Technol 37(9):1101–1109. https://doi.org/10.1080/09593330.2015.1102330

Wadzuk B, DelVecchio T, Sample-Lord K, Ahmed M, Welker A (2021) Nutrient removal in rain garden lysimeters with different soil types. J Sustainable Water Built Environ 7(1):04020018. https://doi.org/10.1061/JSWBAY.0000924

Yang H, McCoy EL, Grewal PS, Dick WA (2010) Dissolved nutrients and atrazine removal by column-scale monophasic and biphasic rain garden model systems. Chemosphere 80(8):929–934. https://doi.org/10.1016/j.chemosphere.2010.05.021

Yang H, Dick WA, McCoy EL, Phelan PL, Grewal PS (2013) Field evaluation of a new biphasic rain garden for stormwater flow management and pollutant removal. Ecol Eng 54:22–31. https://doi.org/10.1016/j.ecoleng.2013.01.005

Yu B, Yuan Z, Yu Z, Xue-song F (2022) BTEX in the environment: an update on sources, fate, distribution, pretreatment, analysis, and removal techniques. Chem Eng J 435:134825. https://doi.org/10.1016/j.cej.2022.134825

Zhang Z, Rengel Z, Liaghati T, Antoniette T, Meney K (2011) Influence of plant species and submerged zone with carbon addition on nutrient removal in stormwater biofilter. Ecol Eng 37(11):1833–1841. https://doi.org/10.1016/j.ecoleng.2011.06.016

Zhao J, Zhao Y, Zhao X, Jiang C (2016) Agricultural runoff pollution control by a grassed Swales coupled with wetland detention ponds system: a case study in Taihu basin, China. Environ Sci Pollut Res 23:9093–9104. https://doi.org/10.1007/s11356-016-6150-2

**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

**Shivani Yadav[1] · Saurav Ambastha[1] · Harsh Pipil[1] · Anil Kumar Haritash[1] · Krishna R. Reddy[2]**

✉ Harsh Pipil
harsh_phd2k19@dtu.ac.in

Shivani Yadav
shivaniyadav_phd2k19@dtu.ac.in

Saurav Ambastha
saurav.ambastha@gmail.com

Anil Kumar Haritash
akharitash@dce.ac.in

Krishna R. Reddy
kreddy@uic.edu

[1] Department of Environmental Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India

[2] Department of Civil, Materials, and Environmental Engineering, University of Illinois, Chicago, IL 60607, USA

2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3315-2169-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/IATMSI64286.2025.10985249

# Design of DTMOS Based Preamplifier for Low Power and High Speed Comparator

Akshay Mann
*Department of Electronics and Communication Engineering*
*Delhi Technological University (DTU)*
Delhi 110042, India
akshay.mann@dtu.ac.in

Neeta Pandey
*Department of Electronics and Communication Engineering*
*Delhi Technological University (DTU)*
Delhi 110042, India
n66pandey@rediffmail.com

Maneesha Gupta
*Department of Electronics and Communication Engineering*
*Netaji Subhas University of Technology (NSUT)*
Delhi 110078, India
maneeshapub@gmail.com

***Abstract*- In this work, an improved Dynamic Threshold MOS (DTMOS) based preamplifier circuit is proposed for low power and high speed comparator. By implementing this technique, the preamplifier circuit could be operated at low supply voltage of 0.6 V, which reduces the overall power consumption of comparator circuit. Additionally, DTMOS transistor enhances the gain (25%) of preamplifier, helping in reducing delay (50%) and increasing the overall speed of the comparator. In this paper, the DTMOS based preamplifier is first implemented with static latch resulting in better performance of comparator in terms of delay, power, PDP and energy as 109 psec, 31.8 µW, 3.5 fJ and 33 fJ/conv respectively. Secondly, the same preamplifier is implemented with dynamic latch, further improving the comparator's performance (delay, power, PDP and energy as 64 psec, 20 µW, 1.3 fJ and 22 fJ/conv respectively). To validate the robustness of circuit to device mismatches, Monte Carlo simulations are also performed. The proposed DTMOS based preamplifier is more suitable for low voltage and high speed comparator applications.**

***Keywords- Dynamic threshold MOS (DTMOS), preamplifier, comparator, static latch, dynamic latch.***

## I. INTRODUCTION

The lifetime of battery operated portable devices can be extended by using low supply voltage and power efficient analog circuits. The comparator is an important building block in Analog to Digital Converters (ADCs) and plays a crucial role in interacting with data from analog to digital domain through different transducers. A formidable work among analog designers is to design a comparator and enhance its performance, mainly gain and speed, at low supply voltages. Different cascading methods for comparator's preamplifier [3] are implemented which enhances the gain by improving the output resistance but this method reduces the output voltage swing. This paper presents comparator circuit where the preamplifier stage is implemented by DTMOS transistor and which works at 0.6 V supply voltage without compromising on power and speed of the comparator circuit.

This technique is a variant of forward body biasing (FBB) where a signal dependent bias is applied to the substrate of MOS transistor. This lowers its threshold voltage and the on-state current by minimizing the leakage currents when transistor is OFF. Therefore, without increasing static power consumption, this technique helps in reducing the delay [4].

The rest of the paper is organized as follows: Section II describes the methodology and techniques of DTMOS technique over the state of art work; Section III discusses the simulation results of the proposed circuit and Section IV concludes the paper.

## II. DTMOS METHODOLOGY AND TECHNIQUE

A two stage traditional comparator is shown in Fig. 1. The first stage comprises of preamplifier block which apprehends the input, while second stage uses static latch.

The comparator works in pre-charge phase (Clk = 0), charging the output nodes of pre-amplifier (Vop and Von) to Vdd. During evaluation phase (Clk = 1), either of the nodes (Vop and Von) discharges from Vdd to ground, based on applied inputs (Vip and Vin). Subsequently, this regenerates the output nodes of static latch (Outp and Outn) to attain logic 1 or 0 [5]. This circuit is suitable for energy-efficient operations, but its performance degrades at low supply voltages and higher clock frequencies. Additionally, the conventional technique found no use in IoT and analog/mixed signal design applications.

Due to advancement of technology scaling and shrinking device size, it also becomes necessary to reduce the power supply voltage [6]. Simultaneously, threshold voltage (Vt) cannot be scaled down in dynamic circuits due to their failure at low power voltages [7]. One possible solution is to use Dynamic Threshold voltage MOSFET (DTMOS) device by connecting the gate and body (substrate) of MOSFET together, in order to obtain a variable Vt. Fig. 2 and 3 presents the proposed comparators where pre-amplifier stage is implemented using DTMOS device. The gain of pre-amplifier stage is given by [8]:

$$Av(Preamp) = \frac{\Delta Vo}{\Delta Vi} \approx \frac{gm_{1s,2s}*(Vdd-Voen)}{Icm} \quad (1)$$

where $g_{m1s,2s}$ is the trans-conductance of transistor M1s or M2s, Voen = |Vop – Von|, is the output level to enable cross coupled

979-8-3315-2169-1/25/$31.00 ©2025 IEEE
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:00:22 UTC from IEEE Xplore. Restrictions apply.

latch and Icm is the common mode current drawn from the capacitances if placed on nodes P and Q.

Also, for MOSFET to work in saturation region, the drain current is expressed as [9]:

$$I_D = \frac{\mu_n C_{ox} W (V_{GS} - V_t)^2 (1 + \lambda V_{DS})}{2 * L} \quad (2)$$

For DTMOS device, $gm_{1s,2s}$ can be expressed by differentiating $I_D$ w.r.t. $V_{GS}$ as:

$$I_D = \frac{\partial I_D}{\partial V_{GS}} = \frac{\mu_n C_{ox} W (V_{GS} - V_t)(1 + \lambda V_{DS})}{L} * \frac{1}{1 + \frac{\gamma}{2 * \sqrt{V_{SB} + \emptyset_f}}} \quad (3)$$

From above equation, it can be inferred for DTMOS that as $V_t$ is reduced radically, $gm_{1s,2s}$ is higher than that of conventional pre-amplifier. Additionally, the higher gain helps to reduce the common mode current $I_{CM}$ and stabilizes it especially for low voltage applications.

Therefore, the pre-amplifier delay can be written as:

$$t_{pre} \approx \frac{C_{eff} |V_{op} - V_{on}|}{gm_{1s,2s} |V_{ip} - V_{in}|} \quad (4)$$

Higher $gm_{1s,2s}$ provides lesser delay.

For first proposed comparator, delay for second stage is expressed as :

$$t_{slatch} = \frac{\tau_{PLH} + \tau_{PHL}}{2} \quad (5)$$

where $\tau_{PLH}$ and $\tau_{PHL}$ are low to high and high to low propagation delays associated with either output nodes Outp / Outn respectively.

Similarly, for second proposed comparator, delay for second stage is written as:

$$t_{dlatch} \approx \frac{C_{xy} | Out_p - Out_n |}{gm_{eqd} | V_{op} - V_{on} |} \quad (6)$$

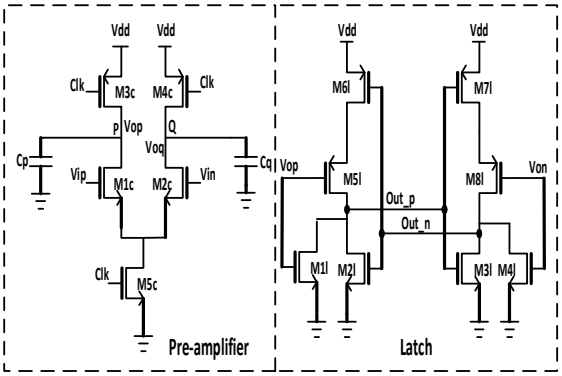



Fig.1. Schematic of Conventional Comparator

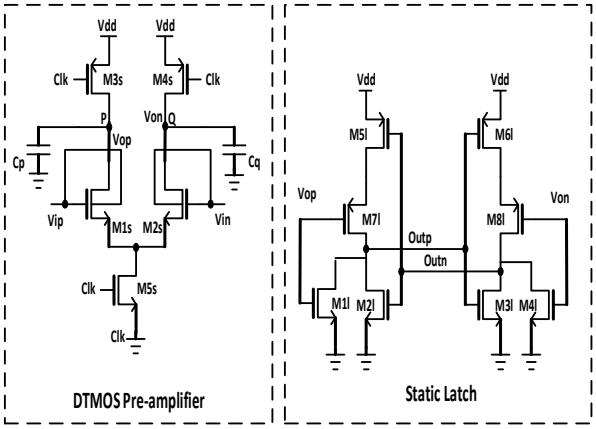



Fig.2. Schematic of First Proposed Comparator

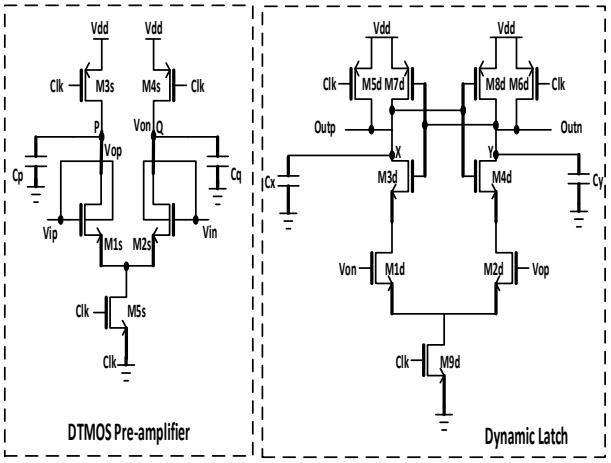
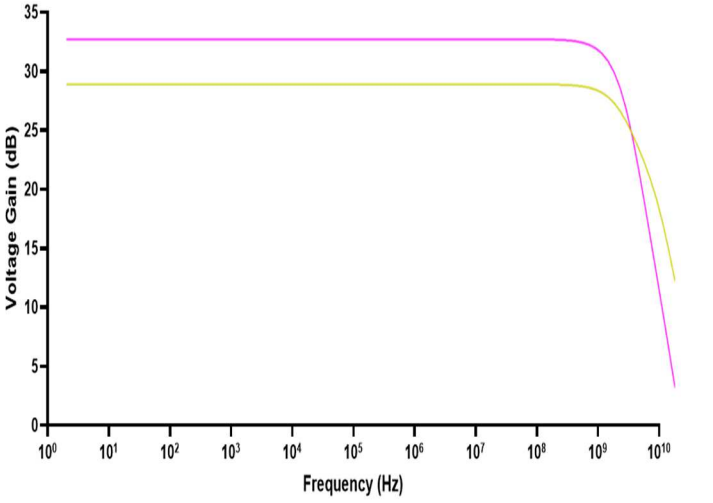



Fig.3. Schematic of Second Proposed Comparator

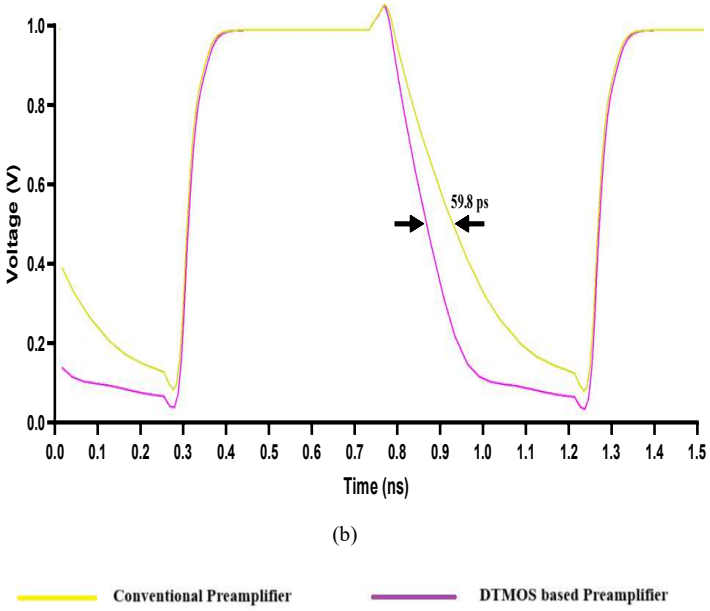
## III. SIMULATION RESULTS

This section describes the experimental analysis along with summary of results for the first and second proposed comparators. The above topologies are designed in 90 nm CMOS process at 0.6 V power supply, 1 GHz clock frequency and characterised on the basis of pre-amplifiers gain and delay as well as overall power consumption, delay, PDP and energy of both comparators. To measure the gain and delay of DTMOS based pre-amplifiers, AC analysis is performed where 0.6 $V_{PP}$ voltage swing is applied to one input while other one is fixed at DC level. From Fig. 4 (a) and (b), a significant improvement in gain (25%) and delay (50%) is observed. Secondly, both comparator circuits are evaluated for delay, power consumption, PDP and energy by applying one of the inputs as ramp signal of 0.6 V voltage swing and other maintained at DC voltage of 300 mV.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:00:22 UTC from IEEE Xplore. Restrictions apply.



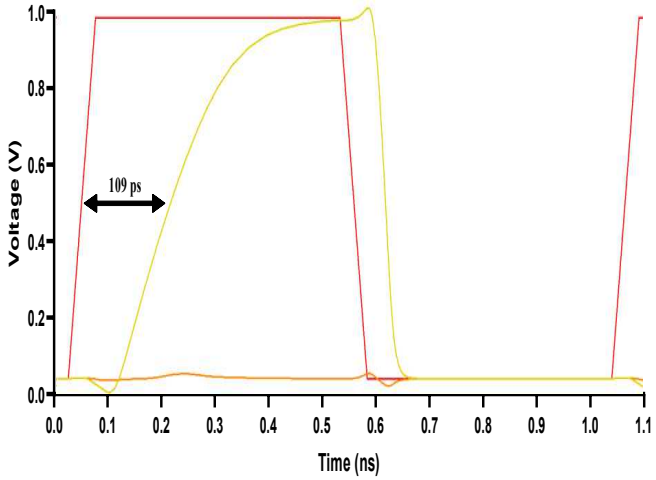

(a)



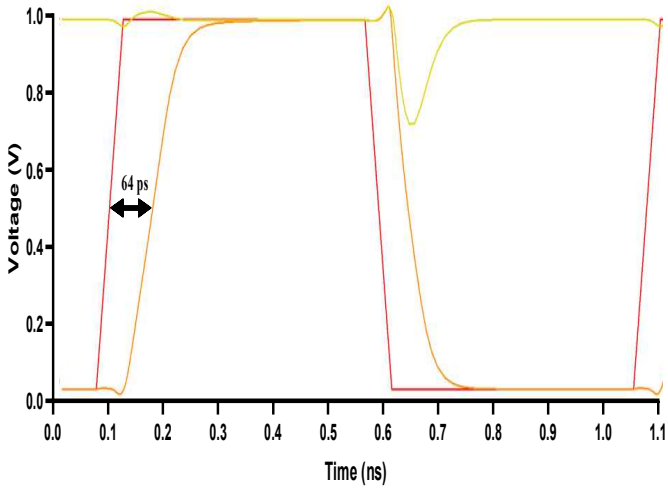

(b)




Fig. 4. Comparison of (a) Gain and (b) Delay of conventional and DTMOS based Pre-amplifiers




(a)




(b)

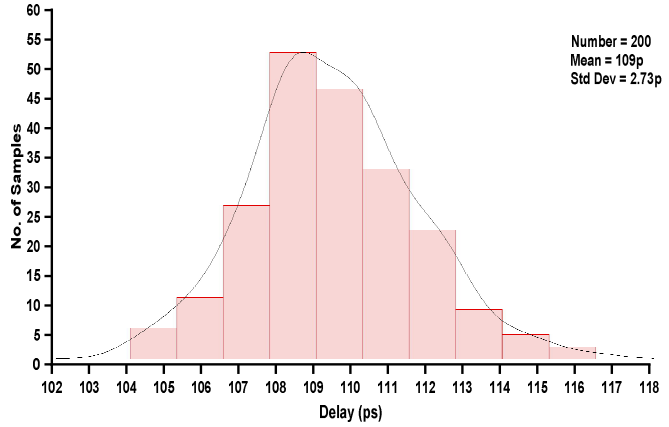



Fig. 5. Transient analysis of (a) First Proposed Comparator and (b) Second Proposed Comparator

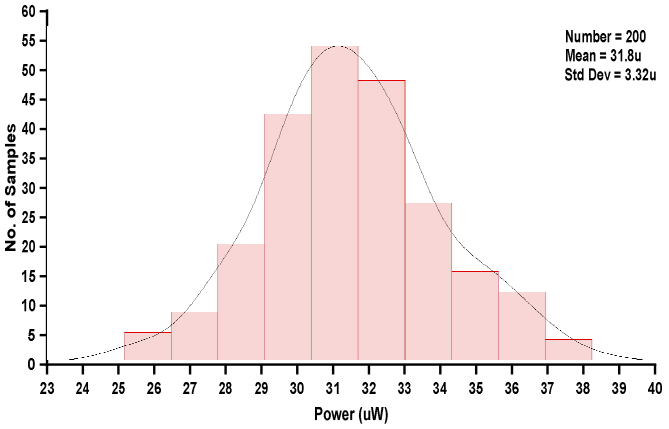
The transient analysis depicting in Fig. 5 (a) and (b) shows the delay of 109 psec and 64 psec of first proposed comparator and second proposed comparator respectively. Also, power consumed, PDP and energy of first proposed comparator are 31.8 µW, 3.5 fJ and 33 fJ/conv respectively while the ame values for second proposed comparator are 20 µW, 1.3 fJ and 22 fJ/conv respectively.

Fig. 6 and 7 shows the Monte Carlo analysis of delay, power and energy using 200 samples for first proposed and second proposed comparator.

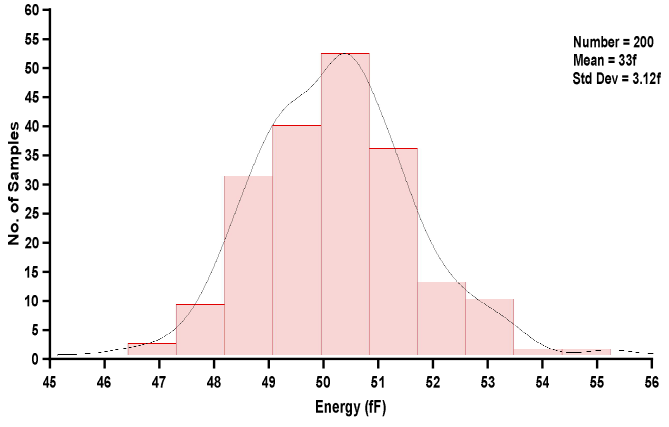
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:00:22 UTC from IEEE Xplore. Restrictions apply.




(a)




(b)




(c)

Fig.6. Monte Carlo simulations for (a) Delay (b) Power and (c) Energy for First Proposed Comparator

Table 1 compares the proposed comparators with recent works. From the table, it can be inferred that with 1 GHz clock frequency, the proposed comparators provide reduced power consumption as compared to [4], and [5]; better delay performance in comparison to [4], [5], [6] and [8]; and are more energy efficient as compared to [4], [5] and [8].

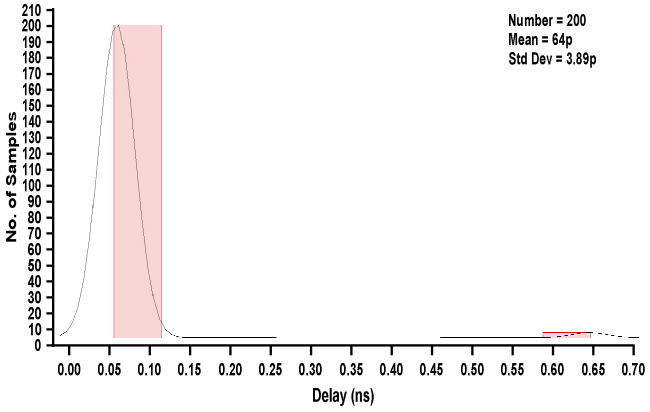
TABLE I

COMPARISON OF PROPOSED METHODOLOGY WITH STATE-OF-ART ALGORITHMS

| Parameters | This Work | | [4] | [5] | [6] | [8] |
|---|---|---|---|---|---|---|
| | P1# | P2* | | | | |
| **Technology (nm)** | 90 | 90 | 55 | 90 | 90 | 90 |
| **Supply Voltage (V)** | 0.6 | 0.6 | 1 | 1 | 0.8 | 1 |
| **Clock Frequency (GHz)** | 1 | 1 | 2 | 1 | 1 | 0.5 |
| **Power Consumption (µW)** | 31.8 | 20 | 109.1 | 57 | 0.33 | 20.4 |
| **Delay (psec)** | 109 | 64 | 82.68 | 72 | 240 | 167.4 |
| **PDP (fJ)** | 3.5 | 1.3 | 9.024 | 4.1 | 0.08 | 3.4 |
| **Energy (fJ/conv)** | 33 | 22 | 121 | 52 | - | 34 |

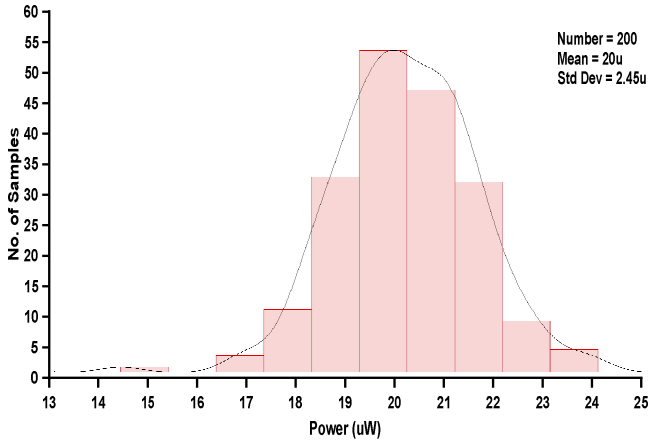
# First Proposed Comparator

* Second Proposed Comparator

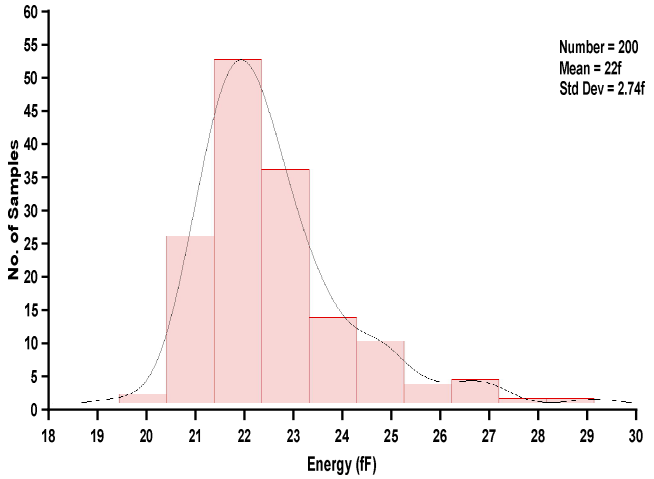



(a)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:00:22 UTC from IEEE Xplore. Restrictions apply.




(b)




(c)

Fig.7. Monte Carlo simulations for (a) Delay (b) Power and (c) Energy for Second Proposed Comparator

## IV. CONCLUSION

We have presented two low power and high speed comparator circuits. Both proposed topologies introduces DTMOS based pre-amplifier circuit operating at 0.6 V supply voltage at 1 GHz clock frequency. The DTMOS based pre-amplifier improves its gain and delay. Also, integrating the novel pre-amplifier with second stage latches helps in reducing overall delay and power as compared to state of art works. Additionally, the proposed comparators achieves less PDP and energy, thereby, making them energy efficient.

## REFERENCES

[1] X. Tang et al., "Low-Power SAR ADC Design: Overview and Survey of State-of-the-Art Techniques," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 69, no. 6, pp. 2249–2262, Jun. 2022.

[2] V. Varshney and R. K. Nagaria, "Design and analysis of ultra high-speed low-power double tail dynamic comparator using charge sharing scheme," AEU - International Journal of Electronics and Communications, vol. 111, p. 152916, 2020.

[3] A. S. Khade, S. Musale, R. Suryawanshi, and S. K. Paul, "A DTMOS-based power efficient recycling folded cascode operational transconductance amplifier," Analog Integrated Circuits and Signal Processing, vol. 107, pp. 227–238, 2021.

[4] V. Spinogatti, R. Della Sala, C. Bocciarelli, F. Centurelli, and A. Trifiletti, "Body Biasing Techniques for Dynamic Comparators: A Systematic Survey," Electronics, vol. 13, no. 4, p. 711, 2024.

[5] A. Mann, N. Pandey, and M. Gupta, "Novel high speed low power comparators imbibing self-cascode preamplifier technique," AEU - International Journal of Electronics and Communications, vol. 185, p. 155429, 2024.

[6] N. Sahana and S. Dwivedi, "Efficient design of a double tail dynamic comparator for high speed and low power applications in 90nm technology," in Proceedings of the 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), pp. 1–4, 2024.

[7] F. Assaderaghi, D. Sinitsky, S. A. Parke, J. Bokor, P. K. Ko and Chenming Hu, "Dynamic threshold-voltage MOSFET (DTMOS) for ultra-low voltage VLSI," in IEEE Transactions on Electron Devices, vol. 44, no. 3, pp. 414-422, March 1997.

[8] V. Varshney, A. Kumar, A. K. Dubey, P. Singh, and R. K. Nagaria, "A high-speed energy-efficient CMOS dynamic latch comparator for low-voltage applications," in Proceedings of the 2019 International Conference on Electrical, Electronics and Computer Engineering (UPCON), pp. 1–6, 2019.

[9] Madhulika, S. Kumari and M. Gupta, "Design and Analysis of DTMOS based Low Voltage OTA and its Filter Application," 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, pp. 685-690, 2021.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:00:22 UTC from IEEE Xplore. Restrictions apply.

Optical and Quantum Electronics (2025) 57:284
https://doi.org/10.1007/s11082-025-08202-3

# Dispersion-engineered segmented core rib waveguide for coherent and broadband supercontinuum generation in both all-normal and anomalous regimes

**Deepak Garg**[1] · **Ajeet Kumar**[1]

Received: 24 February 2025 / Accepted: 31 March 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

**Abstract**
We present a novel, dispersion engineered, segmented core rib waveguide for broadband, highly coherent and flat supercontinuum generation in both all-normal and anomalous dispersion regime. The waveguide uses Aluminum Gallium Arsenide, known as "silicon of non-linear optics", as the core material. To achieve a flat and low dispersion profile, we employed dispersion engineering by optimizing the geometrical parameters of the waveguide. The optimized design features an all normal, flat top dispersion profile with values below—4.43 ps/nm/km across the spectral range of 4.8–7 μm achieving a dispersion value of − 1.71 ps/nm/km at the pump wavelength of 5 μm. Our simulation study, employing the split step Fourier transform method results in coherent supercontinuum spectrum broadening from 2 to 12 μm (40 dB level) and 2.1 to 9.8 μm (20 dB level) in 3 mm long waveguide when pumped with 5 μm secant hyperbolic pulses having pulse width 50 fs and peak power 6 kW. To the best of our knowledge, such broadband supercontinuum spectrum in mid-infrared region at the 20 dB level using waveguide geometry has been reported for the first time. Additionally, when pumped in anomalous dispersion region at 6 μm, our simulations result in supercontinuum broadening from 2.6 to 16 μm (40 dB level) in mid-infrared region under identical laser operating conditions. The proposed waveguide can serve as a promising candidate for on-chip supercontinuum sources and for diverse applications like frequency comb generation, early cancer diagnostic, chemical sensing and food-quality control.

**Keywords** Supercontinuum generation · Dispersion · Non-linear effects · Waveguide

✉ Ajeet Kumar
ajeetdph@dtu.ac.in

Deepak Garg
deepak_24phdap08@dtu.ac.in

[1] Advanced Photonics Simulation Research Laboratory, Department of Applied Physics, Delhi Technological University, Bawana Road, Shahabad Daulatpur, Delhi 110042, India

Published online: 30 April 2025

## 1 Introduction

Supercontinuum Generation (SCG), results from the broadening of ultrashort high intensity laser pulses during the propagation through optically transparent nonlinear medium such as optical fibers, photonic crystal fibers (PCFs) and waveguides. SCG has been a subject of active research for the past two decades due to its diverse applications in multiple fields. These applications ranges from telecommunications and optical coherence tomography (Rao et al. 2021) to spectroscopy (Grassani et al. 2019) and optical frequency combs (Hickstein et al. 2017). The potential of SCG for applications in metrology and imaging has been demonstrated in recent studies (Singh et al. 2018; Poudel et al. 2019). SCG has been explored in PCFs, waveguides and integrated platforms. SCG has been studied in PCFs using various methods, including dispersion engineering, leveraging the design flexibility and effective light confinement offered by these fibers. Moreover, on chip SC sources and integrated optical platforms offers additional advantages over conventional fibers and PCFs in terms of their exceptionally high non-linear coefficients and superior dispersion characteristics than their PCF counterparts, power efficiency, robustness, cost- effectiveness and scalability of compact optical setups. In 2019, a group reported the comparison of SC generation performance in a PCF and a rib waveguide. While, the PCF demonstrated an impressive ultra-broadband and coherent SC spectrum having bandwidth of 18.75 μm, the rib waveguide exhibited an even wider spectrum ranging from 1 to 20 μm under the same pumping conditions (Diouf et al. 2019). Recently, various SCG studies have been reported for on-chip and integrated platforms by using a range of geometries and materials (Fang et al. 2023). Silica has been a widely used background material in the past years due to its well-known electrical and optical properties. However, due to high phonon energy (1150 $cm^{-1}$) of silica, the transmission losses (Wang et al. 2025) beyond 2.5 μm increase, constraining its overall utility. Although it is possible to achieve the desired dispersion characteristic with silica but its low non-linear coefficient and high power requirements for efficient supercontinuum generation have prompted investigations into alternative materials such as chalcogenide glasses (Yu et al. 2013), silica nitride (Johnson et al. 2015), ZBLAN (Agger et al. 2012), silica germanium alloys (Sinobad et al. 2020), Hydex glass (Moss et al. 2013), $LiNbO_3$ (Yu et al. 2019) and group III–V materials (Vyas et al. 2022) etc.

AlGaAs, often referred to as "silicon of non-linear optics", is a group III–V semiconductor compound that has been widely used in the realization of wide range of nonlinear optical components due to its superior nonlinear properties. Its natural suitability for monolithic integration of passive and active non-linear devices (Mobini et al. 2022), wide optical transparency window ranging from 0.9 to 17 μm (Gehrsitz et al. 2000; Savanier et al. 2011), and low linear and nonlinear losses makes it an exceptional material. By tuning the aluminum composition in $Al_xGa_{1-x}As$, its bandgap energy and, consequently, optical and electrical properties can be precisely engineered. Another key advantage of AlGaAs is its suitability for even-order nonlinear optical interactions. Its Zinc blende crystalline structure lacks a center of inversion symmetry, making it an efficient $\chi^{(2)}$ material. For non-linear devices the desirable material properties include large Kerr nonlinearity and low linear and non-linear losses. The chosen core composition (for, $x = 0.18$) i.e. $Al_{0.18}Ga_{0.82}As$, achieves a Kerr coefficient of $1.0 \times 10^{-17}$ $m^2/W$, nearly 100 times larger than that of silica. In our case, here for the cladding material the taken composition of $Al_xGa_{1-x}As$ is (for, $x = 0.80$) i.e. $Al_{0.80}Ga_{0.20}As$. Furthermore, for over entire compositional range, AlGaAs is lattice matched and very few defects are associated with the epitaxy process, which makes the fabrication of heterogeneous AlGaAs waveguides possible (Mobini et al. 2022). Moreover,

due to recent advancements in the fabrication processes, it is now possible to minimize the propagation losses in AlGaAs waveguides (Mobini et al. 2022). Thus, the rationale for choosing these specific material compositions of AlGaAs is strongly supported by both its optical performance and ease of fabrication.

Dispersion engineering and pump wavelength plays a critical role in shaping the underlying dynamics of SC generation, making it essential to efficiently control group velocity dispersion (GVD), quantified by second order dispersion coefficient($\beta_2$). Depending on the sign of GVD ($\beta_2$), two distinct dispersion regimes emerges. In anomalous dispersion regime($\beta_2 < 0$) soliton related propagation effects dominate, leveraging the interplay of self-phase modulation (SPM) and dispersion for spectral broadening. Conversely, in the normal dispersion regime ($\beta_2 > 0$), spectral broadening occurs primarily through SPM. Waveguide dispersion engineering enables shifting the dispersion from normal regime to anomalous regime (which occur at transition wavelength, often referred to as zero dispersion wavelength or ZDW), and allows precise control of dispersion values to achieve a desired and flattened dispersion profile/characteristics. Compared to PCFs, waveguides here also offer significant advantages due to high refractive index contrast between core/cladding interfaces, making their dispersion highly sensitive to small dimension changes.

In recent years, researchers have demonstrated SC generation by pumping in both anomalous and normal dispersion regime. Medjouri et al. numerically demonstrated SC generation spanning from 0.7 to 5.2 μm in a chalcogenide channel waveguide achieved using 100 fs pulses at a wavelength of 2.5 μm with a high peak power of 25 kW (Medjouri et al. 2018). Similarly, Saini et al. reported SC generation in a rib waveguide structure with $As_2Se_3$ as background material. The reported structure produced a broadband SC spectrum from 1 to 10.9 μm by pumping in anomalous dispersion regime using 497 ps hyperbolic pulses at 2.8 μm wavelength with a 6.8 kW of peak power (Saini et al. 2018). In 2019, Ahmad et al. demonstrated the SC generation process using $Si_3N_4$ and $MgF_2$ waveguide structures. Pumping in anomalous dispersion regime resulted in broadband SC spectra spanning 0.8–6.5 μm, while pumping in normal dispersion regime yielded a reduced bandwidth of 0.95–2.1 μm under the same laser operating conditions (50 fs pulse width and 5 kW peak power) (Ahmed et al. 2019). Zhang et al. proposed a polarization- insensitive reverse ridge AlGaAs waveguide for SC generation. This waveguide structure is capable of producing a broadband spectra spanning 2.17–8.53 μm when pumped with 4.2 μm wavelength having pulse width of 90 fs and peak power of 4.8 kW (Zhang et al. 2022). More recently, Hitaishi and Nandam reported SC generation in reverse rib waveguide design pumped at 6 μm wavelength, capable of producing broadband SC spectra ranging from 3.45 to 13.1 μm with 5 kW pulses and extended further to 3.45–16 μm with 10 kW pulses (Hitaishi and Nandam 2023). Furthermore, Sheikhmolaee et al. introduced a tapered rib waveguide design utilizing hybrid chalcogenide GAP-Se for SC generation. This design achieved a spectral range of 2.2–5 μm by pumping at 3.1 μm in 4 mm long waveguide (Sheikhmolaee et al. 2024).

In this study, we numerically investigate and demonstrate the proposed segmented core waveguide incorporating different compositions of AlGaAs in core and cladding region, for coherent and efficient supercontinuum generation in mid-infrared region. The motivation and rationale for choosing the segmented-core design is to leverage additional degrees of freedom for tailoring and flattening the dispersion profile, enabling better control over nonlinear interactions. Dispersion engineering was performed by tailoring the geometrical parameters of the waveguide to achieve an optimized flat and tunable dispersion profile. Based on these optimizations, we finally propose a waveguide structure with both normal and anomalous dispersion characteristics. Additionally, we

analyze the effects of various laser operating conditions on the SC spectra. Using split step Fourier transform method, our simulations reveal coherent SC spectral broadening from 2 to 12 µm at 40 dB level and 2.1 to 9.8 µm at 20 dB level in 3 mm long waveguide. These results are obtained by pumping with 5 µm secant hyperbolic pulses having pulse width 50 fs and peak power 6 kW. Furthermore, when the waveguide is pumped at 6 µm wavelength in anomalous dispersion regime under identical laser operating conditions the SC spectra broadens significantly from 2.6 to 16 µm at 40 dB level.

## 2 Waveguide design methodology

The lateral cross section view of the proposed waveguide design is shown in Fig. 1. The core and cladding region of the proposed waveguide is composed of compositionally varied AlGaAs. The core of the waveguide consists of $Al_xGa_{1-x}As$ with $x = 0.18$, corresponding to 18% aluminum, providing a higher refractive index and cladding layers are made of $Al_xGa_{1-x}As$ with $x = 0.80$, corresponding to 80% aluminum, resulting in lower refractive index. This variation in composition ensures the efficient optical confinement in the core area due to refractive index contrast. The effective mode index contrast versus wavelength is shown in Fig. 2. The core region is divided into a central rib and two adjacent slab sections, forming a segmented structure. The rib and slab, both regions have the same height denoted by '*h*'. The total width of the core is 3*w*, where the width of the rib is 'w', and each slab region on either side of the rib also has a width of '*w*'. This uniformity in the dimensions ensures the symmetry, facilitates the mode confinement and makes fabrication process relatively easier. Dispersion engineering was carried out by altering the geometrical parameters of the waveguide to control the wavelength dependency of effective indices of the propagating modes. And, finally an optimized structure having a desired and flat dispersion profile is proposed. The dispersion profile has been simulated using full-vectorial finite element method based COMSOL Multiphysics Software. The core region of the simulated waveguide contains several modes and in order to prevent intermodal coupling, mode suppressing conditions are met by the effective indices of the modes.

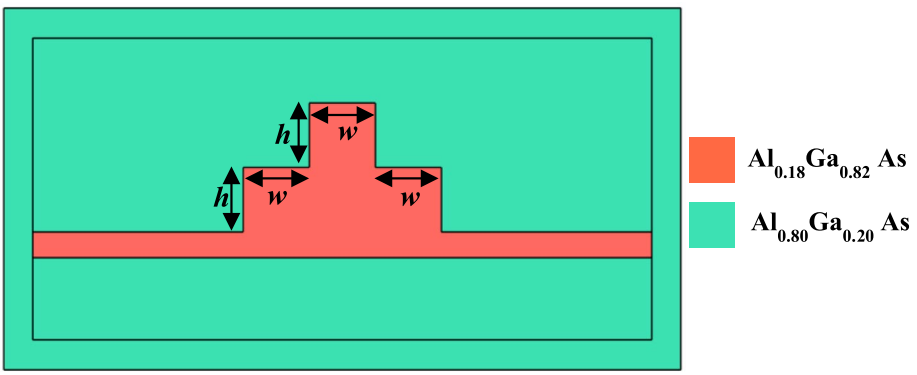



**Fig. 1** Transverse cross-section view of proposed waveguide

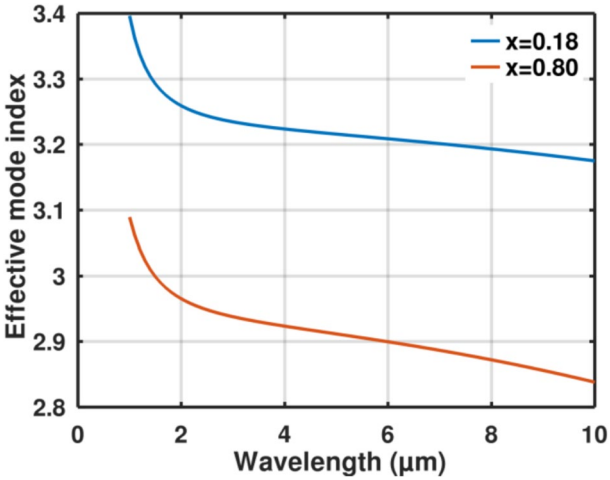



**Fig. 2** Effective mode index v/s wavelength for different composition of $Al_xGa_{1-x}As$

## 3 Numerical analysis

### 3.1 Linear characteristics of SC generation

The modal characteristics of the waveguide geometry, with a focus on effective refractive index ($n_{eff}$) were analyzed by simulating the propagating fundamental TE modes using COMSOL Multiphysics, a commercially available computational software based on full-vectorial finite element method. The wavelength dependent linear refractive index of various compositions of $Al_xGa_{1-x}As$ has been calculated using the equation given below (Kiroriwal and Singal 2021):

$$n^2(\lambda) = 10.906 - 2.92x + \frac{0.97501}{\lambda^2 - C} - 0.002467(1.41x + 1)\lambda^2 \tag{1}$$

Here, in the above equation $x$ is the fractional composition of Aluminum in $Al_xGa_{1-x}As$ and $C = (0.52886 - 0.735x)^2$.

As mentioned above, understanding and managing dispersion is important. The dispersion coefficient and dispersion profile can be calculated using the wavelength dependency of effective refractive indices ($n_{eff}$), given in the equation given below (Kanika et al. 2024):

$$D(\lambda) = \frac{-\lambda}{c}\frac{d^2\mathrm{Re}(n_{eff})}{d\lambda^2} \tag{2}$$

where $Re(n_{eff})$ is the real part of effective refractive indices and c is the speed of light.

Another important parameter to analyze is effective area of the waveguide's fundamental mode that underlies pulse propagation. Effective mode area ($A_{eff}$) quantifies the confinement of optical field where the non-linear interactions takes place. It can be given as (Singh et al. 2024):

$$A_{eff} = \frac{\left(\int\int_{-\infty}^{\infty} |E|^2 dxdy\right)^2}{\int\int_{-\infty}^{\infty} |E|^4 dxdy} \tag{3}$$

where E represents the transverse electric field intensity obtained by solving the eigenvalue problem derived from Maxwell's equations.

### 3.2 Non-linear characteristics of SC generation

Another crucial aspect of SC generation is to analyze the non-linearity of the waveguide structure. The non-linear coefficient ($\gamma$) depends on both effective mode area ($A_{eff}$) of the fundamental mode and non-linear refractive index ($n_2$) of the waveguide core material. It can be calculated using the formula mentioned below (Garg and Kumar 2025):

$$\gamma = \frac{2\pi n_2}{\lambda A_{eff}(\lambda)} \tag{4}$$

To achieve broadband and efficient SC generation, the waveguide's non-linearity should be maximized, which requires low effective mode area. This can be accomplished by selecting a material with high non-linear refractive index or optimizing the waveguide geometry to reduce the effective mode area.

To model the propagation dynamics of an ultra-short, high intensity pulse in a waveguide structure, we employed a pseudo spectral split-step Fourier transform method (SSFM) to solve the generalized non-linear Schrödinger wave equation (GNLSE). The following equation represents the GNLSE (Garg and Kumar 2025):

$$\frac{\partial A}{\partial Z} + \frac{\alpha}{2}A - \left(\sum_{n\geq 2}\frac{i^{n+1}}{n!}\beta_n\frac{\partial^n A}{\partial T^n}\right) = i\gamma\left(1+\frac{i}{\omega_0}\frac{\partial}{\partial T}\right)\left(A(z,t)\int_{-\infty}^{\infty}R(T)\left|A\left(z,T-T'\right)\right|^2 dT + i\hbar_R(z,T)\right) \tag{5}$$

In the above equation right hand side accounts for all non-linear effects, while the left hand side represents linear effects. Here, $\alpha$ denotes the waveguide transmission loss and the slowly varying optical wave envelope is denoted by $\prime A\prime$, while by $z$ and $t$ represent the spatial co-ordinate along the length of the waveguide and the time variable, respectively. Since, we are working with very short waveguide length, the Raman term $\hbar_R$ can be approximated out to zero. Here, $\beta_n$ represents the $n^{th}$ order dispersion parameter, which is obtained from the Taylor series expansion of the propagation constant. Using Taylor series expansion at a central angular frequency $\omega_0$, the propagation constant ($\beta$) can be expanded as follows (Sharma et al. 2020):

$$\beta(\omega) = \beta(\omega_0) + \beta_1(\omega-\omega_0) + \frac{1}{2}\beta_2(\omega_0)(\omega-\omega_0)^2 + \frac{1}{3!}\beta_3(\omega-\omega_0)^3 \ldots\ldots \tag{6}$$

The non-linear response function, $R(T)$ which incorporates both instantaneous electronic Kerr $\delta(t)$ and delayed Raman response $h_R(t)$ is given as (Cheng et al. 2021)

$$R(t) = (1-f_R)\delta(t) + f_R h_R(t) \tag{7}$$

where the value of fractional Raman response $f_R$ is 0.047, representing the Raman contribution (Sharma et al. 2019). The Raman response function $h_R(t)$ takes the following form:

$$h_R(t) = \frac{\tau_1^2+\tau_2^2}{\tau_1\tau_2^2}\exp\left[\frac{-t}{\tau_2}\right]\sin\left(\frac{t}{\tau_1}\right) \tag{8}$$

where the inverse of photon oscillation frequency $\tau_1$ and inverse Raman gain spectrum bandwidth $\tau_2$ are found to be 18.72 fs, and 750 fs respectively (Kim et al. 1979).

A tunable, hyperbolic secant pulse is used for simulating SC which is given as (Sharma et al. 2020):

$$A(z=0,t)=\sqrt{P_0}\mathrm{sech}\left(\frac{T}{T_0}\right) \tag{9}$$

where the secant hyperbolic nature of the input pulse is represented by 'sech', peak power is represented by $P_0$, and $T_0 = T_{FWHM}/1.7627$ and $T$ represents the co-propagating frame at group velocity of the optical wave envelope.

Another important parameter while analyzing an SC source is coherence property. The coherence property of supercontinuum light depends on various factors including the waveguide geometrical parameters, laser operating conditions (such as pulse width, peak power and pump wavelength) and the SC broadening mechanism (Saini and Sinha 2021). To evaluate SC coherence, a semi classical model known as one-photon-per-mode noise model is used. This analysis assess coherence by calculating the absolute value of the complex degree of the first order coherence denoted as $g_{12}^{(1)}$ (Saini and Sinha 2021). The value of $g_{12}^{(1)}$ varies from 0 to 1, indicating the degree of coherence.

$$\left|g_{12}^{(1)}(\lambda,t_1-t_2)\right|=\left|\frac{\left\langle E_1^*(\lambda,t_1)E_2(\lambda,t_1)\right\rangle}{\sqrt{\left\langle\left|E_1(\lambda,t_1)\right|^2\right\rangle\left\langle\left|E_2(\lambda,t_2)\right|^2\right\rangle}}\right| \tag{10}$$

Angle brackets in the above equation denote the ensemble average over complex spectral envelopes.

## 4 Results and discussion

### 4.1 Dispersion engineering

As mentioned above, controlling and managing dispersion plays a crucial role in shaping the underlying dynamics of SC generation. Therefore, tuning dispersion to achieve a desirable dispersion profile is essential. To accomplish this, we have altered the geometric parameters of the waveguide and analyzed the resulting shift in the dispersion values. The dispersion variations corresponding to changes in the waveguide's rib height '$h$' and rib width '$w$' are presented in Figs. 3 and 4, respectively. In Fig. 4, height '$h$' is varied while keeping the width fixed at $w = 1.6$ µm. Here too, dispersion profile moves towards the anomalous region with the increase in '$h$'. It can be seen that with the variation in '$h$' from 1.35 to 1.65 µm with a step size of 0.15 µm (i.e. 150 nm), an all-normal dispersion region with the flattest dispersion near the zero dispersion line is observed at $h = 1.5$ µm. Thus, we chose $h = 1.5$ µm as the optimized height parameter. Similarly, in Fig. 4, width '$w$' is varied while keeping the height at $h = 1.5$ µm. The results shows that on increasing '$w$' from 1.3 to 1.7 µm causes the dispersion profile to shift towards anomalous dispersion region. The increase in '$w$' increases the effective mode index

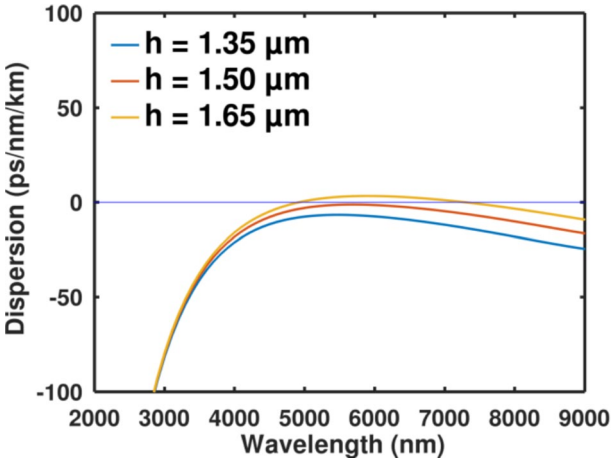



**Fig. 3** Dispersion profile with variation in height '$h$' while keeping $w = 1.6$ µm

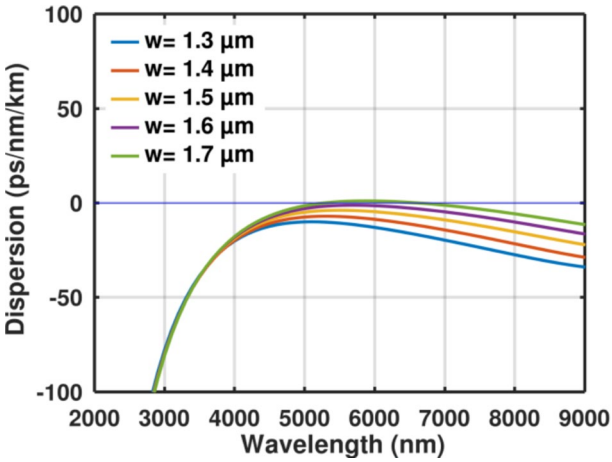



**Fig. 4** Dispersion profile with variation in width '$w$' while keeping $h = 1.0$ µm

**Table 1** Optimized geometrical parameters for all normal dispersion profile

| Geometrical Parameters | Width ($w$) | Height ($h$) |
|---|---|---|
| Values | 1.6 µm | 1.5 µm |

of fundamental mode and hence tailors the dispersion towards anomalous dispersion. Notably, at $w = 1.6$ µm, the dispersion profile falls under an all normal dispersion region having lowest dispersion values approaching zero dispersion line. Since, a flat and all normal dispersion profile is desirable for supercontinuum generation (Saini et al. 2017), we have selected 1.6 µm as the optimized width.

The optimized parameters for the all normal and flattened dispersion profile are tabulated in Table 1. The dispersion profile by taking the optimized parameters in account is illustrated in Fig. 5. As shown, the waveguide achieves a flat dispersion curve in the wavelength range of 4.3–7.0 µm within the dispersion value of $-4.43$ ps/nm/km in the normal regime. The minimum dispersion occurs near the pump wavelength of 5 µm,

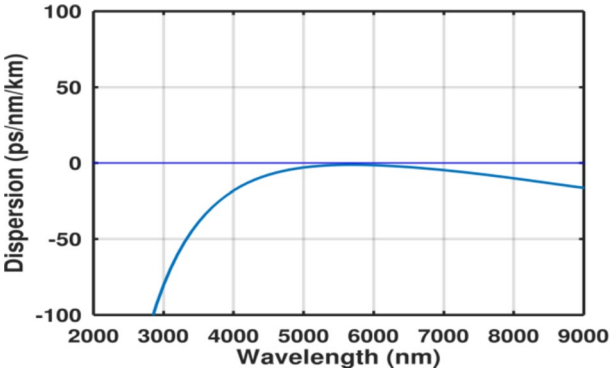



**Fig. 5** All normal dispersion profile at optimized structure parameters

with the value of $-1.71$ ps/nm/km. In addition, robustness of the waveguide is evaluated by considering practical fabrication errors, introducing $\pm 2\%$ uncertainty in optimized parameters listed in Table 1.

The spectral dependence of Effective mode area and Non-linear coefficient at the optimized parameters is illustrated in Fig. 6. At a pump wavelength of 5 μm, within the normal dispersion regime, the effective mode area and non-linear coefficient is 13.87 $\mu m^2$ and 1358.28 $W^{-1}$ $m^{-1}$, respectively. Furthermore, with the fabrication tolerance of $\pm 2\%$ in the optimized parameters, we have observed only a 1% variation in effective mode area and 0.8% variation in non-linear coefficient. These minimal variations highlight the stability and robustness of the proposed design against minor fabrication deviations.

Additionally, as seen in Figs. 3 and 4, at $h = 1.65$ μm and $w = 1.7$ μm, the dispersion profile transitions from normal to anomalous regime. Under these conditions, the waveguide exhibits a flattened anomalous dispersion profile having low dispersion values and two zero dispersion wavelengths (ZDWs) near 5 μm and 7 μm. The corresponding dispersion curve is shown in Fig. 7. Therefore, with the same waveguide design by tuning the

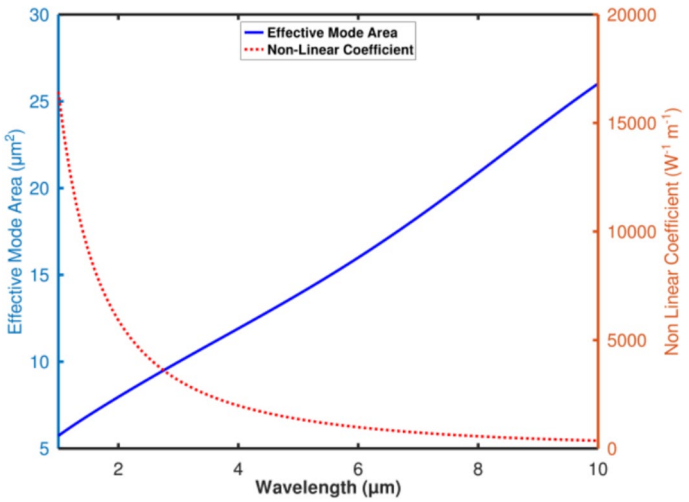



**Fig. 6** Spectral variation of effective mode area and non-linear coefficient

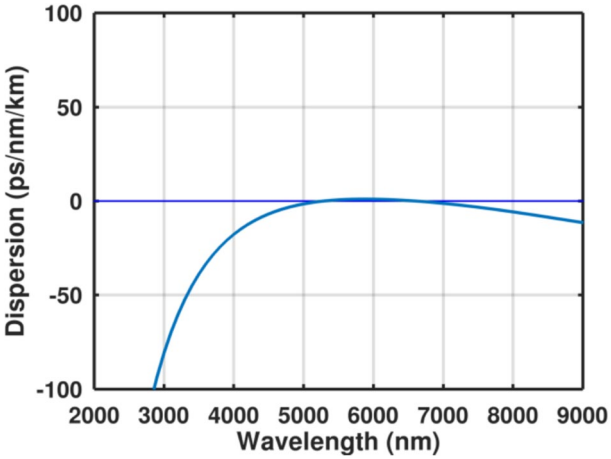



**Fig. 7** Dispersion profile at $w = 1.7$ μm and $h = 1.65$ μm

geometrical parameters and by operating at $h = 1.65$ μm and $w = 1.7$ μm provides the flexibility to explore SC generation in normal as well as anomalous dispersion regime. The value of non-linear coefficient at 6 μm pump wavelength comes out to be 958.09 $W^{-1}$ $km^{-1}$ and simulated value of $\beta_2$ at pump wavelength is $-4.5567 \times 10^{-11}$ $ps^2$/nm.

### 4.2 Supercontinuum generation in all-normal dispersion regime

Supercontinuum generation is studied by solving the GNLSE i.e. Equation (5), using a pseudo spectral numerical method called split step Fourier transform method. This method solves GNLSE efficiently by handling the non-linear effects in time domain and linear dispersion effect in frequency domain. By considering the optimized parameters from Table 1, we can exercise the condition of pumping the waveguide in an all normal dispersion region. In this regime, the main mechanism of spectral broadening is SPM. Since, the chirp imposed by normal dispersion and SPM has the same sign, the spectral broadening in this regime is relatively limited (Bres et al. 2023). However, due to the interplay between dispersion and SPM, a new phenomenon known as optical wave breaking (OWB) occurs which enhances the spectral broadening. It is important to note that during the initial phase of propagation, SPM alone is responsible for spectral broadening. However, as the pulse propagated and reaches a specific distance known as optical wave breaking distance ($Z_{WB}$), OWB begins to contribute in spectral broadening (Bres et al. 2023).

The SC spectrum is simulated for different propagation lengths of waveguide and various laser operating conditions such as varying pulse width and peak power of input ultrashort high intensity pulse. Figure 8(a) illustrates the effect of waveguide propagation length on SC spectrum. From the figure, it can be observed that the maximum broadening occurs when the length of the waveguide is 3 mm (or 0.3 cm). The value of non-linear length $L_{NL}(= 1/\gamma P_0)$, also known as characteristic length of SPM i.e. non-linearity, comes out to be ~ 0.12 mm. Another characteristic length, the dispersive length $L_D(=T_0^2/|\beta_2|)$ comes out to be ~ 23.07 mm, where our simulated results shows the value of $\beta_2 = 3.48 \times 10^{-11}$ $ps^2$/nm and $T_0 = 28.36$ fs. Now, the soliton order $N(\sqrt{L_D/L_{NL}})$ comes out to be ~ 13.86. These calculations will help determine whether the optical wave breaking plays a role in observed SC broadening. The calculated value for $Z_{WB}(=L_D/\sqrt{4N^2e^{-3/2}-1})$ comes out to be

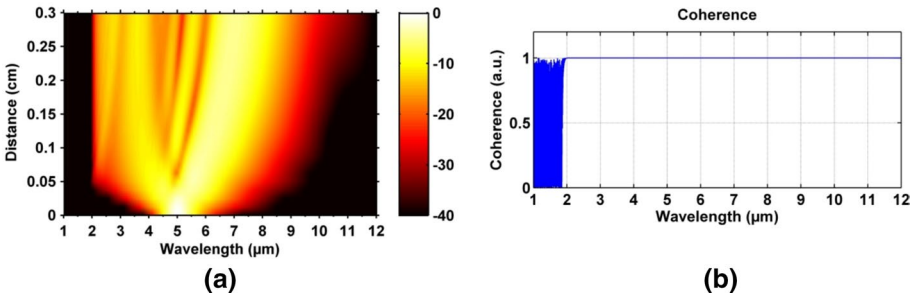



**Fig. 8** **a** Spectral evolution with length of waveguide. **b** Coherence at 3 mm length

1.76 mm. This confirms that here in this case both SPM and OWB plays crucial role in SC broadening. This conclusion is further supported by Fig. 8(b), which shows that the high degree of coherence in the generated SC spectrum is attributed to the dynamics of SPM and OWB. Since, the dynamics of SPM and OWB generation are less susceptible to noise amplification, the resulting SC spectrum remains stable and coherent. Hence, when pumped with 5 μm pulses in all normal dispersion regime with pulse width of 50 fs and peak power of 6 kW, the proposed waveguide is predicted to generate a highly coherent broadband SC spectrum ranging from 2 to 12 μm at 40 dB level.

Figure 9 illustrates the effect of pulse width on SC spectral evolution while keeping the waveguide length and peak power constant at 3 mm and 6 kW, respectively. The figure shows that as the pulse width increases, the SC broadening decreases i.e. while decreasing the pulse width from 150 to 50 fs, the non-linear effect (SPM) becomes more

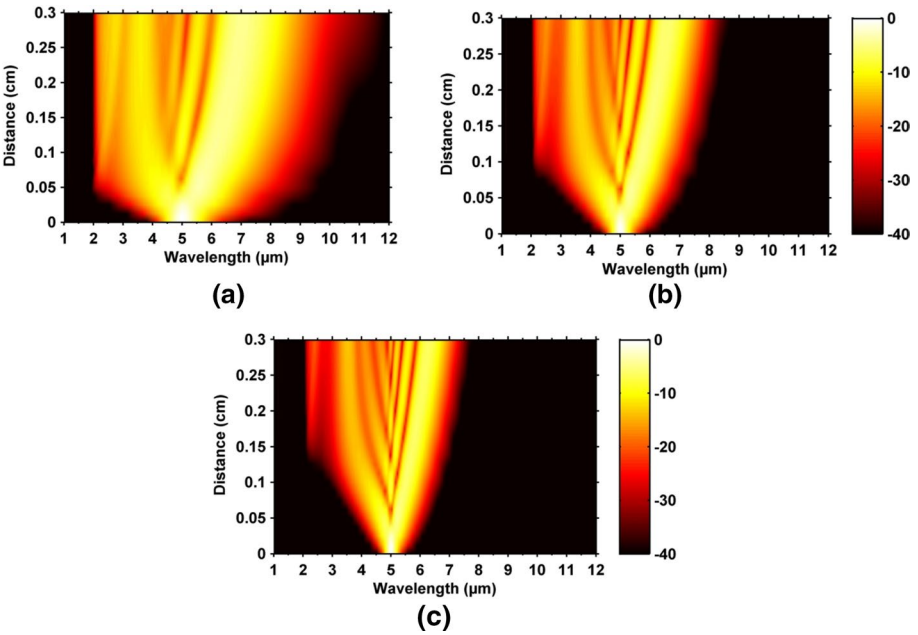



**Fig. 9** Spectral evolution plots of generated SC at **a** 50 fs, **b** 100 fs, and **c** 150 fs

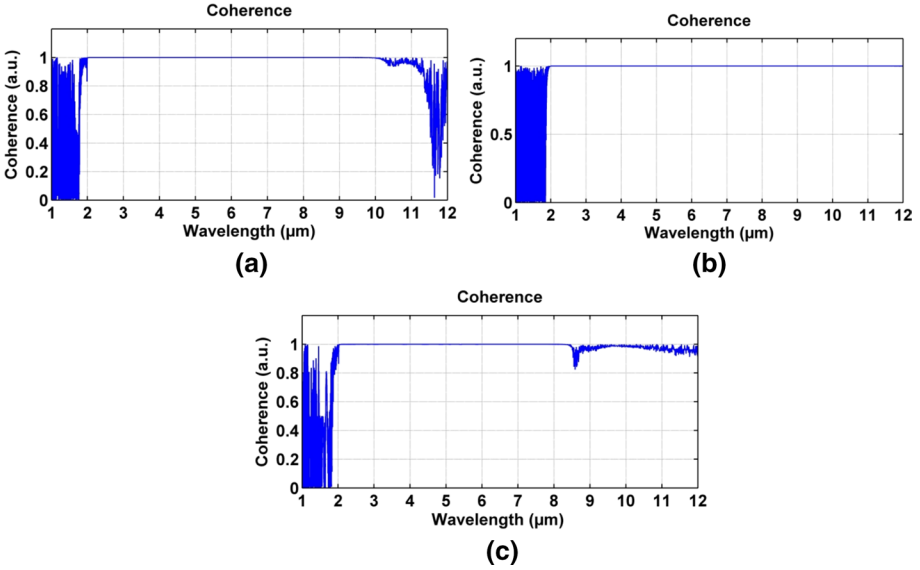



**Fig. 10** Coherence plots of generated SC at **a** 4 kW, **b** 6 kW, and **c** 8 kW

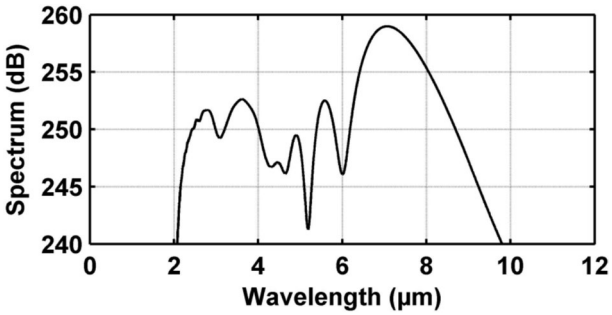



**Fig. 11** SC line plot at 20 dB level in all normal dispersion regime

dominant thereby leading to greater spectral broadening. This suggests that shorter pulse widths favors the spectral broadening, especially for short propagation lengths, such as 3 mm waveguide used in this study.

Figure 10 illustrates the effect of peak power on the coherence property of the generated spectrum. Here, we have simulated SC spectrum to calculate coherence at 4 kW, 6 kW and 8 kW. It can be observed from the figure that on increasing the power level from 4 to 6 kW, there is an increase in the coherence level over a broader range of wavelengths i.e. at 6 kW the output spectra is coherent over the entire SC range from 2 to 12 μm. However, beyond a certain threshold (here, 8 kW) coherence slightly decreases in the long wavelength region, likely due to introduction of noise at higher peak powers (Genty et al. 2016).

Figure 11 shows the SC line plot of the generated spectrum at 20 dB level, where the spectral broadening extends from 2.1 to 9.8 μm. When pumped at 5 μm wavelength in the all normal dispersion regime using 50 fs hyperbolic secant pulses with a peak power 6 kW, the proposed waveguide is predicted to generate a stable and coherent broadband

supercontinuum spectrum covering 2–12 μm at 40 dB level and from 2.1 to 9.8 μm at 20 dB level.

### 4.3 Supercontinuum generation in anomalous dispersion regime

By tuning the waveguide geometric parameters i.e. through dispersion engineering we achieved a flat and low valued anomalous dispersion profile with two ZDWs. This allowed us to study the SC generation by pumping the waveguide in anomalous dispersion regime. SC in anomalous dispersion regime is generated by leveraging the opposite effects of SPM and dispersion. Pumping in anomalous dispersion regime enables soliton dynamics and soliton fission, where ultrashort pulses maintain their shape while undergoing non-linear interactions, enabling significant spectral broadening. The non-linear characteristic length ($L_{NL}$) is approximately ~ 0.17 mm which is shorter than the waveguide length used in our simulations. Dispersive length comes out to be ($L_D$) comes out to be ~ 17.65 mm satisfying the condition $L_D > L_{NL}$ for soliton regime pumping. The corresponding soliton number ($N$) based on non-linear length and dispersive length turns out to be ~ 10. Now, while working in anomalous dispersion regime, two additional characteristics lengths namely soliton fission length ($L_f = L_D/N$) and modulation instability length ($L_{MI} \approx 16L_{NL}$) are crucial for understanding the dynamics of SC. Soliton fission length comes out to be 1.76 mm and modulation instability length comes out to be 2.72 mm indicating that soliton fission occurs before the modulation instability. Additionally, the degree of coherence in this regime is related to the number of solitons. The soliton number $N < 16$ ensures the high degree of coherence (Bres et al. 2023) which confirms the highly coherent nature of our output broadband spectrum.

Figure 12 shows the SC spectral evolution and coherence of the output spectrum. By pumping in anomalous dispersion regime at 6 μm under the identical laser operating conditions i.e. with 50 fs hyperbolic secant pulses with peak power 6 kW, we achieved a highly coherent and broadband SC spectrum ranging from 2.6 to 16 μm at 40 dB level is observed in just 3 mm long waveguide.

Through dispersion engineering the proposed waveguide geometry enables both an all normal dispersion regime and an anomalous dispersion regime providing the flexiblity to operate in either regime. When pumped in all-normal dispersion regime, our simulations demonstrates highly coherent and broadband SC generation from 2 to 12 μm (at 40 dB level) and 2.1 to 9.8 μm (20 dB level) in 3 mm long waveguide. This is achieved using 5 μm secant hyperbolic pulses having pulse width 50 fs and peak power 6 kW. After examining

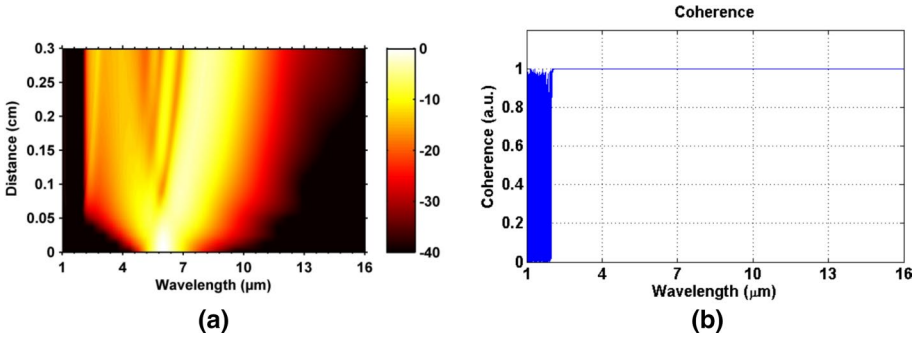



**Fig. 12** **a** Spectral evolution and **b** coherence while pumping in anomalous dispersion regime

**Table 2** Comparison of our proposed structure with previously reported results

| Material | Geometry | Regime | Pump wavelength | Peak power | Spectral range (μm) | References |
|---|---|---|---|---|---|---|
| $As_2S_5$ | Channel waveguide | All-normal | 2.5 μm | 25 kW | 0.7–5.2 μm | Medjouri et al. (2018) |
| $As_2Se_3$ | Rib waveguide | Anomalous | 2.8 μm | 6.8 kW | 1–10.9 μm | Saini et al. (2018) |
| $Si_3N_4$ | Rectangular waveguide | Anomalous | 1.55 μm | 5 kW | 0.8–6.5 μm | Ahmad et al. (2019) |
| | | All-normal | 1.55 μm | 5 kW | 0.95–2.1 μm | |
| AlGaAs | Reverse ridge waveguide | Anomalous | 4.2 μm | 4.8 kW | 2.17–8.53 μm | Zhang et al. (2022) |
| $As_2Se_3$ | Reverse rib waveguide | Anomalous | 6 μm | 10 kW | 3.45–16 μm | Hitaishi and Nandam (2023) |
| $As_2S_3$ | Photonic crystal fiber | All-normal | 4.5 μm | 20 kW | 1.5–8 μm | Sheikhmolaee et al. (2024) |
| | | Anomalous | 6 μm | 6 kW | 2–15 μm | |
| $Al_xGa_{1-x}As$ | Segmented core waveguide | All-normal | 5 μm | 6 kW | 2.1–12 μm | This work |
| | | Anomalous | 6 μm | 6 kW | 2.6–16 μm | |

the effect of all laser operating conditions, we confirm our decision to propose the segmented core waveguide geometry for efficient SC generation in mid-infrared region. Furthermore, under identical laser operating conditions but in anomalous dispersion regime, by pumping at 6 μm, our simulation results in coherent SC broadening spanning 2.6–16 μm in the mid-infrared region. Table 2 shows the comparison of our work with previously reported structures for SC generation in mid-infrared region.

## 5 Conclusion

A new design of segmented core rib waveguide for broadband, highly coherent and flat SC generation in mid-infrared region by pumping in both all-normal and anomalous dispersion regime is reported in this paper. After the analysis of the dispersion characterstics and by performing dispersion engineering we simulated the proposed waveguide for SC genertion. In an all-normal dispersion regime by pumping with 5 μm hyperbolic secant pulses having 50 fs pulse width and 6 kW peak power, our simulations result in coherent SC broadening from 2 to 12 μm (40 dB level) or 2.1–9.8 μm (20 dB level) in 3 mm long waveguide. Under the identical laser operating conditions, in anomalous dispersion regime by pumping at 6 μm, our simulation results in coherent SC broadening spanning 2.6–16 μm in mid-infrared regime covering both atmospheric transparent window i.e. 3–5 μm and 8–13 μm of spectral region and molecular fingerprint region i.e. 2–15 μm. Finally, the high coherence, flat and broadband SC spectrum in mid-infrared region from both dispersion regimes positions our proposed waveguide design as a promising candidate for future photonics integrated circuits, on-chip SC sources and for diverse range of non-linear applications including spectroscopy, environmental monitoring and bio-sensing.

**Acknowledgements** Deepak Garg highly acknowledges the support from the Delhi Technological University, Delhi, for providing the research fellowship and necessary support to conduct this research work.

**Author contributions** D Garg has done writing—original draft, validation, methodology, investigation and A Kumar has done writing—review and editing, supervision, software, and formal analysis.

**Funding** The authors have not disclosed any funding.

**Data availability** No datasets were generated or analysed during the current study.

## Declarations

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

## References

Agger, C., Petersen, C., Dupont, S., Steffensen, H., Lyngsø, J.K., Thomsen, C.L., Thøgersen, J., Keiding, S.R., Bang, O.: Supercontinuum generation in ZBLAN fibers—detailed comparison between measurement and simulation. JOSA B **29**(4), 635–645 (2012)

Ahmad, H., Karim, M.R., Rahman, B.M.A.: Dispersion-engineered silicon nitride waveguides for mid-infrared supercontinuum generation covering the wavelength range 0.8–6.5 μm. Laser Phys. **29**(2), 025301 (2019). https://doi.org/10.1088/1555-6611/aaf63d

Brès, C.S., Torre, A.D., Grassani, D., Brasch, V., Grillet, C., Monat, C.: Supercontinuum in integrated photonics: generation, applications, challenges, and perspectives. Nanophotonics **12**(7), 1199–1244 (2023)

Cheng, Y., Lai, J., Yuan, J., Mei, C., Zhou, X., Wu, Q., Liu, B., et al.: Highly coherent and multi-octave mid-infrared supercontinuum generations in a reverse-strip AlGaAs waveguide with three zero-dispersion wavelengths. Appl. Opt. **60**(31), 9994–10001 (2021)

Diouf, M., Mandeng, L.M., Tchawoua, C., Zghal, M.: Numerical investigation of supercontinuum generation through $AsSe_2/As_2S_5$ chalcogenide photonic crystal fibres and rib structures. J. Lightw. Technol. **37**(22), 5692–5698 (2019)

Fang, Y., Bao, C., Li, S., Wang, Z., Geng, W., Wang, Y., Han, X., et al.: Recent progress of supercontinuum generation in nanophotonic waveguides. Laser Photonics Rev. **17**(1), 2200205 (2023)

Garg, D., Kumar, A.: Highly coherent on-chip mid-infrared supercontinuum generation from 1.5 to 25 μm in CdTe-based Rib waveguide. Opt. Quant. Electron. (2025). https://doi.org/10.1007/s11082-024-08007-w

Gehrsitz, S., Reinhart, F.K., Gourgon, C., Herres, N., Vonlanthen, A., Sigg, H.: The refractive index of $Al_xGa_{1-x}As$ below the band gap: accurate determination and empirical modeling. J. Appl. Phys. **87**(11), 7825–7837 (2000)

Genty, G., Friberg, A.T., Turunen, J.: Coherence of supercontinuum light. Prog. Opt.. Opt. **61**, 71–112 (2016)

Grassani, D., Tagkoudi, E., Guo, H., Herkommer, C., Yang, F., Kippenberg, T.J., Brès, C.-S.: Mid infrared gas spectroscopy using efficient fiber laser driven photonic chip-based supercontinuum. Nat. Commun.Commun. (2019). https://doi.org/10.1038/s41467-019-09590-3s

Hickstein, D.D., Jung, H., Carlson, D.R., Lind, A., Coddington, I., Srinivasan, K., Ycas, G.G., Cole, D.C., Kowligy, A., Fredrick, C., Droste, S., Lamb, E.S., Newbury, N.R., Tang, H.X., Diddams, S.A., Papp, S.B.: Ultrabroadband supercontinuum generation and frequency-comb stabilization using on-chip waveguides with both cubic and quadratic nonlinearities. Phys. Rev. Appl. (2017). https://doi.org/10.1103/PhysRevApplied.8.014025

Hitaishi, V., Nandam, A.: Broadband supercontinuum generation using dispersion engineered $As_2Se_3$-GeAsSe-GeAsS waveguide at 6 μm. IEEE Access **11**, 12294–12302 (2023)

Johnson, A.R., Mayer, A.S., Klenner, A., Luke, K., Lamb, E.S., Lamont, M.R.E., Joshi, C., Okawachi, Y., Wise, F.W., Lipson, M., Keller, U., Gaeta, A.L.: Octave-spanning coherent supercontinuum generation in a silicon nitride waveguide. Opt. Lett. **40**(21), 5117 (2015). https://doi.org/10.1364/OL.40.005117

Kanika, J.N., Saini, T.S.: Dispersion-engineered SiN-coated $TeO_2$ hybrid waveguide for nonlinear applications. J. Electron. Mater. **53**, 1–10 (2024)

Kim, O.K., Spitzer, W.G.: Infrared reflectivity spectra and Raman spectra of $Ga_{1-x}Al_xAs$ mixed crystals. J. Appl. Phys. **50**(6), 4362–4370 (1979)

Kiroriwal, M., Singal, P.: Numerical analysis on the supercontinuum generation through Al0.24Ga0.76As based photonic crystal fiber. Sādhanā (2021). https://doi.org/10.1007/s12046-021-01695-0

Medjouri, A., Abed, D., Ziane, O., Simohamed, L.M.: Design and optimization of $As_2S_5$ chalcogenide channel waveguide for coherent mid-infrared supercontinuum generation. Optik **154**, 811–820 (2018)
Mobini, E., Espinosa, D.H.G., Vyas, K., Dolgaleva, K.: AlGaAs nonlinear integrated photonics. Micromachines **13**(7), 991 (2022)
Moss, D.J., Morandotti, R., Gaeta, A.L., Lipson, M.: New CMOS-compatible platforms based on silicon nitride and hydex for nonlinear optics. Nat. Photonics **7**(8), 597–607 (2013)
Poudel, C., Kaminski, C.F.: Supercontinuum radiation in fluorescence microscopy and biomedical imaging applications. JOSA B **36**(2), A139–A153 (2019)
Rao, D.S.S., Jensen, M., Nielsen, L.G., Olsen, J.T., et al.: Shot-noise limited, supercontinuum-based optical coherence tomography. Light Sci. Appl. **10**(1), 133 (2021)
Saini, T.S., Sinha, R.K.: Mid-infrared supercontinuum generation in soft-glass specialty optical fibers: a review. Prog. Quantum Electron.. Quantum Electron **78**, 100342 (2021)
Saini, T.S., Kumar, A., Sinha, R.K.: Design and modelling of dispersion-engineered rib waveguide for ultra-broadband mid-infrared supercontinuum generation. J. Mod. Opt. **64**(2), 143–149 (2017)
Saini, T.S., Tiwari, U.K., Sinha, R.K.: Design and analysis of dispersion engineered rib waveguides for on-chip mid-infrared supercontinuum. J. Lightwave Technol. **36**(10), 1993–1999 (2018)
Savanier, M., Andronico, A., Lemaître, A., Galopin, E., Manquest, C., Favero, I., Ducci, S., Leo, G.: Large second-harmonic generation at 155 μm in oxidized AlGaAs waveguides. Opt. Lett. **36**(15), 2955 (2011). https://doi.org/10.1364/OL.36.002955
Sharma, M., Dhasarathan, V., Skibina, J.S., Rajan, M.S.M., Konar, S., Hoang, T.T., Ngo, Q.M.: Giant nonlinear AlGaAs-doped glass photonic crystal fibers for efficient soliton generation at femtojoule energy. IEEE Photonics J. **11**(4), 1–11 (2019). https://doi.org/10.1109/JPHOT.2019.2927492
Sharma, R., Kaur, S., Chauhan, P., Kumar, A.: Computational design & analysis of $GeSe_2$-$As_2Se_3$-PbSe based rib waveguide for mid-infrared supercontinuum generation. Optik **220**, 165032 (2020)
Sheikhmolaee, M., Alizadeh, M.R., Olyaee, S., Seifouri, M.: Supercontinuum generation in tapered planar rib waveguide based on GAP-Se hybrid chalcogenide. Opt. Quant. Electron. (2024). https://doi.org/10.1007/s11082-023-05666-z
Singh, N., Xin, M., Vermeulen, D., Shtyrkova, K., Li, N., Callahan, P.T., Magden, E.S., et al.: Octave-spanning coherent supercontinuum generation in silicon on insulator from 1.06 μm to beyond 2.4 μm. Light Sci. Appl. **7**(1), 17131 (2018)
Singh, G., Sharma, S., Kumar, A.: Design and simulation of solid-core octagonal photonic crystal fiber for terahertz wave propagation. Microw. Opt. Technol. Lett.. Opt. Technol. Lett. (2024). https://doi.org/10.1002/mop.34173
Sinobad, M., Torre, A.D., Armand, R., Davies, B.L., Ma, P., Madden, S., Mitchell, A., et al.: Mid-infrared supercontinuum generation in silicon-germanium all-normal dispersion waveguides. Opt. Lett. **45**(18), 5008–5011 (2020)
Vyas, K., Espinosa, D.H.G., Hutama, D., Jain, S.K., Mahjoub, R., Mobini, E., Awan, K.M., Lundeen, J., Dolgaleva, K.: Group III–V semiconductors as promising nonlinear integrated photonic platforms. Adv. Phys. X (2022). https://doi.org/10.1080/23746149.2022.2097020
Wang, X., Pan, J., Yao, C., Yang, X., Yang, Y., Liu, Y., Li, P.: A new method of dispersion-regulated soliton group pulses for efficient mid-infrared fiber laser generation. J. Lightwave Technol. **43**(8), 4007–4012 (2025). https://doi.org/10.1109/JLT.2025.3530505
Yu, Y., Gai, X., Wang, T., Ma, P., Wang, R., Yang, Z., Choi, D.Y., Madden, S., Davies, B.L.: Mid-infrared supercontinuum generation in chalcogenides. Opt. Mater. Express **3**(8), 1075–1086 (2013)
Yu, M., Desiatov, B., Okawachi, Y., Gaeta, A.L., Lončar, M.: Coherent two-octave-spanning supercontinuum generation in lithium-niobate waveguides. Opt. Lett. **44**(5), 1222–1225 (2019)
Zhang, L., Yuan, J., Cheng, Y., Mei, C., Lai, J., Zhou, X., Wu, Q., et al.: Polarization-insensitive reverse-ridge AlGaAs waveguide for the mid-infrared supercontinuum generation. Opt. Commun.commun. **502**, 127407 (2022)

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

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

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



# EAGLE: Enhanced Activity Recognition in Occluded Environments With Improved GAIN and Bi-LSTM

Diksha Kurchaniya and Sanjay Kumar

***Abstract*—Activity recognition (AR) using surveillance cameras remains challenging because of the variation between the data required for deep learning and the capabilities of CCTV-based frameworks for data recording. Most research on skeleton-based activity focuses on improving recognition using complete data, often overlooking performance on incomplete skeleton data due to occlusion or noise. However, in practical situations, it is inevitable to obtain a noisy or occluded human skeleton. When some informative keyjoints are unavailable or disturbed, it can compromise the efficiency and effectiveness of existing methods. This article proposes a skeleton-based model for activity recognition to deal with the occlusion problem. The main approach of the proposed model is to address occlusion by treating it as a missing value imputation problem within the feature matrix. First, it extracts features such as normalized joints, relative joint coordinates, angles, and velocity using a CNN-based technique and encodes the skeleton data into a feature matrix. The innovative use of generative adversarial imputation networks (GAIN) with bidirectional long short-term memory (Bi-LSTM), enhanced by spatial and temporal attention mechanisms, effectively handled missing data due to occlusions, ensuring a complete feature matrix for subsequent classification. Additionally, this article introduces autoencoder-based Bi-LSTM with attention modules that improve activity classification by capturing and leveraging spatial and temporal dependencies. A comprehensive experiment was conducted on both occluded and unoccluded datasets to validate the effectiveness and improvements of the proposed model. Comparative results show that the proposed model outperforms the existing state-of-the-art methods, achieving superior recognition accuracy in an occluded environment.**

***Index Terms*—Activity recognition, bidirectional long short-term memory (Bi-LSTM), generative adversarial imputation network (GAIN), OpenPose, partial occlusion.**

## I. Introduction

ACTIVITY recognition (AR) using surveillance cameras is a prominent research problem due to its wide range of real-life applications based on visual content analysis, including surveillance systems [1], video retrieval, annotation, and human–computer interaction. The input data of the AR system can vary significantly rely on the application, including RGB videos, infrared data, and depth data obtained from time-of-flight or structured-light-based sensors data [2]. The video-based AR framework is mainly based on RGB and Skeleton data-based activity recognition. The RGB-based method performs better at capturing the appearance information of objects but can be easily affected by background changes [3]. In contrast, skeleton data-based approaches are less sensitive to the background and lighting while effectively capturing changes in position and motion velocity [4]. On the other hand, skeleton data-based techniques significantly reduce the need for computational power.

Although AR system has significantly advanced in recent years, identifying human movements remains challenging due to changing viewpoints, varying velocities, interactions with multiple objects, and, most critically, occlusion. However, occlusion can occur in real-life situations like crowded areas and interactions between various individuals. When objects or other people fully or partially obscure a subject, the data becomes unclear or hidden. Consequently, traditional AR systems need help interpreting and classifying activities accurately, often leading to unsatisfactory performance in real-world scenarios. Therefore, improving AR systems to handle these challenges is essential for enhancing their performance and reliability in practical applications. Most research articles have shown that occlusion remains a prominent cause of shortcomings in pose estimation [5]. Numerous existing methods face difficulties in recovering an accurate pose of the object if the object is partially or completely occluded from another object. Because of this, most of the methods cannot handle occlusion explicitly [6]. Some researchers attempt to determine various 3-D postures, penalize occluded keyjoints, or estimate the risk that keyjoints or body parts would be obscured from an image. Temporal information is used to eliminate the issue of visual input.

Many activity recognition methods do not use any specific approach to deal with the occlusion. They only analyzed the unoccluded area of the skeleton to perform the experiments. This approach drastically limits the extent of information that can be obtained by generating a feature matrix with a high percentage of missing values. If excessive missing data exists, a subsequent classifier may not have enough properties to identify activities appropriately. The motivation behind this article comes from the major challenges that AR systems have in real-world scenarios, especially when occlusion environments are present

Received 2 October 2024; revised 7 February 2025 and 19 March 2025; accepted 1 April 2025. *(Corresponding author: Sanjay Kumar.)*

The authors are with the Department of Computer Science and Engineering, Delhi Technological University, New Delhi 110042, India (e-mail: dikshakurchaniya@gmail.com; sanjay.kumkar@dtu.ac.in).

Digital Object Identifier 10.1109/TCSS.2025.3558970

2329-924X © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. 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 May 13,2025 at 07:07:41 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.



and parts of the human skeleton are hidden by objects or other people. In this article, occlusion environment is defined as the situation where some of the 13 key skeleton keyjoints are missing due to obstacles, human interactions, or self-occlusion. These occlusions occur when body parts are obstructed by objects, other individuals, or camera angle limitations, leading to missing keyjoints in the extracted skeleton data. Traditional approaches frequently overlook occluded regions, leading to incomplete feature matrices and unsatisfactory classification results. This work aims to overcome these constraints by developing an effective AR system that takes occlusion into account directly. This work provides an activity recognition method that integrates effective feature extraction and transformation techniques to tackle this significant limitation. Our approach utilizes an enhanced activity classification method and advanced information imputation techniques specifically designed to deal with occlusion issues. First, we use the OpenPose algorithm to extract the skeleton from video sequences and apply some preprocessing. A convolutional neural network (CNN) algorithm extracts discriminative features from the human skeleton. Inspired by the Li and Li [7], this framework utilizes GAIN with a Bi-LSTM to deal with the missing values. Integrating spatial and temporal attention ensures that it efficiently predicts and imputes the missing features and provides a complete feature matrix. At last, the proposed model uses an autoencoder-based Bi-LSTM with an attention module (ABSA) followed by a softMax activation function to classify human activities. This comprehensive framework ensures the availability of sufficient and accurate features for precise activity recognition and classification. The key contribution of the article is summarized as follows:

1) We propose an occlusion-based framework consisting of feature extraction, feature imputation, and activity classification to handle occlusion in different scenarios.
2) A Data imputation model based on the generative adversarial imputation network is introduced, utilizing Bi-LSTM with a spatial and temporal attention module as the generator to predict and imputes the missing feature values. It ensures the reconstruction of a complete feature matrix, even in occlusion.
3) We proposed an autoencoder-based Bi-LSTM with an attention module to classify activities. It captures the spatial dependencies of each skeleton and temporal dynamics between frames.
4) This performs a comprehensive analysis on the NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD datasets with occluded skeleton parts and introduces a synthesized dataset based on the different occlusion scenarios to better deal with occlusion problems.

The remaining content of this article is further systemized as follows. Section II discussed about the related work. The proposed work is elaborated in Section III. The related datasets and performance of our system are described in Section IV. The conclusion of this work is discussed in Section V.

## II. Related Work

Activity recognition is a prominent area in computer vision and pattern recognition. Researchers have explored a wide range of methods for this task, encompassing both traditional machine-learning algorithms and deep-learning models. Though it is generally acknowledged that occlusion is one of the key factors impairing the effectiveness of AR approaches [8], a few research investigations investigated the effects of occlusion on recognition performance. Charles et al. [9] proposed a method to detect and identify the occluded keyjoints for temporal propagation. The multistreamfusion graph convolutional network (MSFGCN) effectively incorporates multiple streams to handle different occlusion scenarios. This approach allowed for specialized processing paths that can address varying degrees and types of occlusion. Chen et al. [10] introduced a multiview graph convolution fusion approach to tackle challenges related to occlusion and viewpoint variations. This method integrates long short-term memory networks (LSTM) and graph convolutional networks (GCN). The gait identification method is used by Nangtin et al. [11] to deal with partial occlusion. The model utilized six modules, in which gait energy image (GEI) is divided into four modules, and the other three are grouped into two modules, respectively. Charles et al. [9] introduced a method to detect and identify the occluded keyjoints for temporal propagation. They have used temporal neural network (TCN) other than the optical flow to achieve a broader temporal perceptive field. Theodorakopoulos et al. [12] introduces an approach for activity recognition that begins by refining skeletal data to create stable and invariant pose representations. Following this, dissimilarity vectors are calculated by comparing these poses to a predefined set of prototype activities. Pavllo et al. [6] developed a temporal convolution-based approach to generate 3-D poses from sequences of 2-D key joints. However, this method relies on accurately estimating 2-D key joints for each frame and assumes that prediction errors are independent and do not persist over time.

A fully connected CNN with two residual modules was developed by Martinez et al. [13] in order to recognize 3-D poses from 2-D joint locations. Fang et al. [14] used a pose grammar network to improve 3-D pose estimation. Cheng et al. [15] implemented a GrGCN based on directed joints and bones for more precise 3-D posture estimation in multi-person scenarios. To ensure temporal smoothness, Hossain et al. [16] utilized LSTM seq-to-seq architecture. However, its fixed-size input and output remain a limitation. Khan et al. [17] used LSTM model that is able to classify scarce labeled activity data. Zhou et al. [18] proposed a pose estimation method that deals with performance degeneration due to occlusion. This method used an attention mechanism to calculate the attention maps explicitly supervised and clean the feature map. Li et al. [7] proposed a method for identifying building activity in locations that are obstructed. The missing data in the feature matrix is imputed using GAIN. The skeleton of the identified worker was first converted into a feature matrix. To fill in any missing values in the feature matrix, an attention-based model was then incorporated into the data imputation technique based on generative adversarial networks (GAN). Ultimately, the residual network is used to classify the imputed feature matrix.

Simonyan and Zisserman [19] developed a dual-stream convolutional network that divides video into spatial and temporal streams for activity recognition based on RGB video sequences.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



The temporal stream represents the combined optical flow displacement between multiple consecutive frames, whereas the single frame was captured by the spatial stream for visual details. Although the optical flow-based techniques often have a high processing complexity, they directly investigate the motion information between video frames. The approaches for 3-D convolution fall within the second category. 3-D CNN is more suited for processing video data since it can directly model temporal information compared to 2-D CNN. Ji et al. [20] originally introduced 3-D convolution in 2013 and used it for activity recognition challenges.

Ansari and Singh [21] introduced a framework that utilizes postural data captured by a depth camera to identify human activities and detect shoplifting events. The architecture described the preprocessing of the postural data, extracted features from the skeleton, and classified the activities using machine learning algorithms. Le and Kim [22] proposed a model for detecting anomalies in video sequences by combining residual learning, attention mechanism, and autoencoder architecture. The results indicate the model's efficiency and potential applications in diverse fields such as surveillance, industrial monitoring, and healthcare.

The existing AR methods highlight the challenges associated with occlusions. While RGB-based methods effectively capture appearance information, they are sensitive to background variations, whereas skeleton-based approaches offer robustness to lighting and background changes. However, many existing methods lack explicit mechanisms to handle occlusion, relying on incomplete or noisy skeletal data, which leads to limited performance in real-world scenarios. As compared with existing works, our method simplifies occlusion handling by reducing the number of skeleton keyjoints from 21 to 13, improving computational efficiency while maintaining accuracy. Approaches such as GCNs, LSTMs, and attention mechanisms have shown promise in improving robustness but often involve high computational overhead or limited adaptability. These gaps motivate the proposed framework, which focuses on leveraging GAIN for effective imputation, Bi-LSTM with spatiotemporal attention for robust feature learning, and a reduced set of keyjoints to enhance computational efficiency.

The methods discussed above for AR under occlusion employ various approaches such as feature propagation, multi-view fusion, graph-based modeling, and deep learning techniques. Still, they have key limitations in handling missing skeletal data effectively. Charles et al. [9] (MSFGCN) use a multi-stream fusion graph convolutional network to propagate keypoint information across multiple streams, but they do not explicitly impute missing joints. Similarly, Chen et al. [10] use a multi-view graph convolution fusion approach that integrates LSTM and GCN for skeletal reconstruction but primarily focuses on multiview learning rather than direct imputation of missing key joints. In contrast, our proposed method explicitly reconstructs missing keyjoints using GAIN with Bi-LSTM and spatiotemporal attention, ensuring accurate spatial and temporal recovery of occluded skeleton joints. Our method improves upon this by leveraging Bi-LSTM, which preserves sequential consistency in missing joint recovery. Li et al. [7] use GAIN for data imputation but rely on ResNet for classification, which extracts only spatial features. In contrast, our autoencoder-based Bi-LSTM classifier effectively captures both spatial and temporal motion dependencies, ensuring robust classification of occluded actions.

## III. Proposed Methodology

The proposed work aims to identify activities in the presence of occlusion. Fig. 1 depicted the working of the proposed framework. The proposed framework utilizes video sequences captured by surveillance cameras to identify human activities and consists of four steps. First, it extracts the skeleton information from the input videos and detects the spatial position of the body keyjoints. After that, all the irrelevant frames and keyjoints are discarded, select keyjoints that are relevant to identify the human activity, construct the feature matrix, and identify the missing keyjoints. In the next step, it calculates necessary features like normalized joints, relative joint coordinates, the angle $\lambda_{ij}^t$ of each joint $i$ to the horizontal axis, the angle $\lambda_i^t$ of each relative joint to the reference center point and velocity from the N-encoded skeletons and constructs a feature matrix. The feature matrix has some missing values due to occlusion problems. The proposed system utilizes the imputation method to impute the missing data and complete the skeleton with estimated keyjoints. GAIN is used to fill the missing values by incorporating BI-LSTM with spatial and temporal Attention methods and produce the predicted feature matrix with complete data. After that, this feature matrix is passed to the Autoencoder-based Bi-LSTM with Spatiotemporal Attention module (ABSA), followed by the softmax activation function to classify human activities. The forthcoming subsections describe the aforementioned technicalities.

### A. Skeleton Generation and Keyjoint Detection

An OpenPose [23] algorithm is utilized to extract the skeleton information from a human. The OpenPose algorithm extracts a total of 21 skeleton keyjoints that represent each tracked joint. To minimize improper training and unreliable results, the actual skeleton data comprising 21 body key joints is not utilized directly for extracting significant key features. Only 13 skeleton keyjoints are selected to be used for this work, as these keyjoints are enough to give the relevant information. The selection principle focuses on retaining the most informative joints that contribute to activity recognition, particularly under occlusion while discarding joints with minimal relevance. First, it discards the frames that do not have any skeleton information, then we remove four joints (two eyes and two ears) from the head, two joints from hands (left hand, right hand) and two joints from legs (left foot and right foot), as illustrated in Fig. 2. These discarded keyjoints contribute less to high-level activity recognition and are more prone to detection errors, ensuring a more robust and computationally efficient representation for occlusion-aware activity recognition. The selected 13 skeleton keyjoints are the Head, Neck, LeftShoulder, RightShoulder, LeftElbow, RightElbow, LeftWrist, RightWrist, Hip, LeftKnee, RightKnee, LeftAnkle, and RightAnkle. These are crucial for capturing human motion dynamics and distinguishing between different activities. These keyjoints provide spatial and temporal cues related to posture, limb movements, and transitions between actions.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 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.



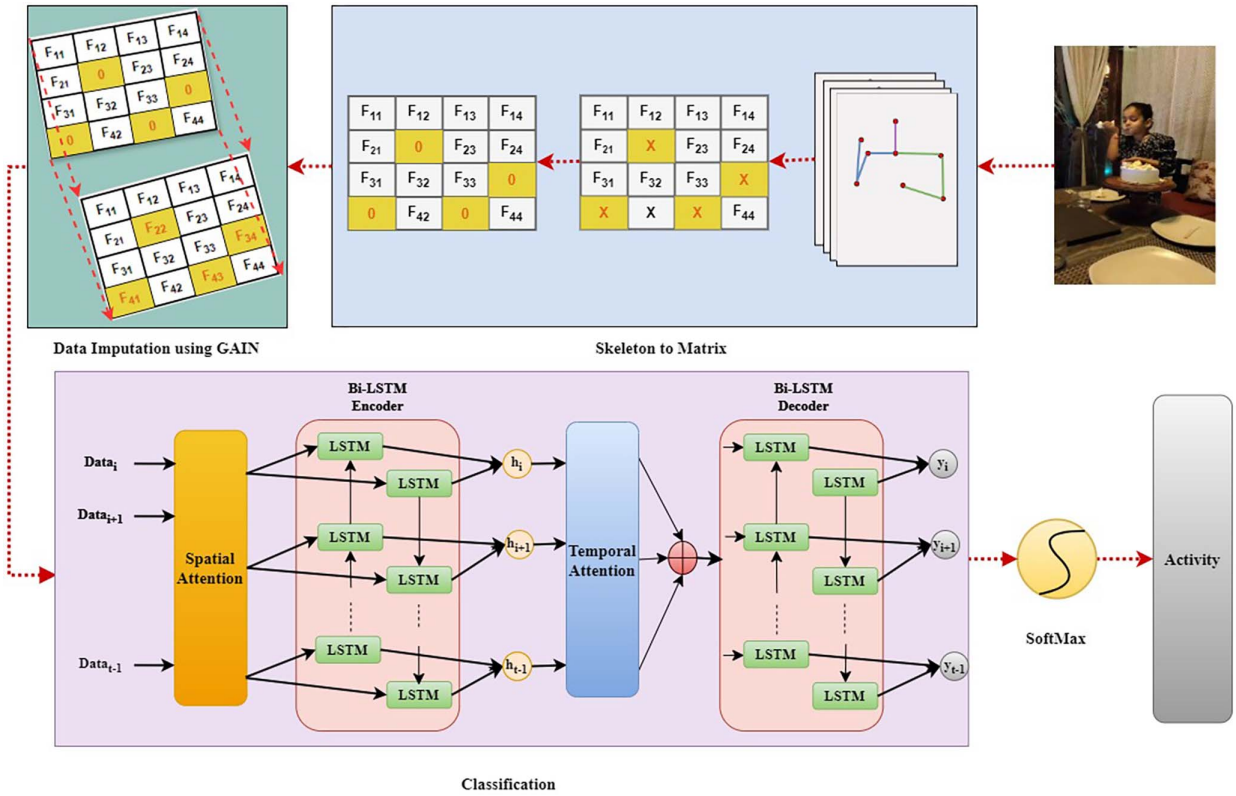



Fig. 1. Pipeline of proposed EAGLE model for activity recognition and classification consisting of three main components: skeleton to matrix encoding, data imputation using GAIN, and autoencoder-based Bi-LSTM with attention module for activity classification.

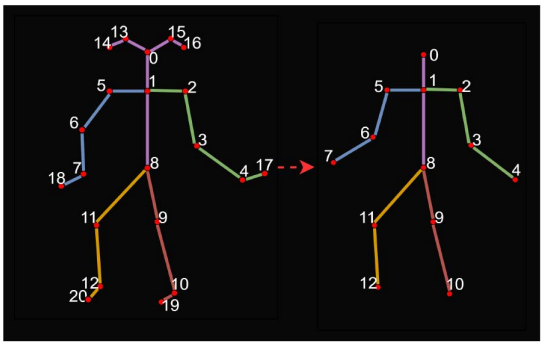


Fig. 2. Illustration of human skeleton representation with 21 keyjoint for preprocessing and selected relevant 13 keyjoints after preprocessing.

TABLE I
DETAIL OF SELECTED RELEVANT SKELETON KEYJOINTS COORDINATES AND DERIVED FEATURES

| Lable | Description |
|---|---|
| Skeleton joints | Head, Neck, LeftShoulder, RightShoulder, LeftElbow, RightElbow, LeftWrist, RightWrist, Hip, LeftKnee, RightKnee, LeftAnkle, and RightAnkle. |
| Features | Xnorm, relative joint coordinates, angle $\lambda_{ij}^{t}$, angle $\theta_{i}^{t}$, Velocity. |

Table I shows the details of selected keyjoints and the extracted features from the skeleton.

The scale of the skeleton coordinates extracted by the OpenPose algorithm is based on the location and size of the human body [24]. Thus scaling is required to transform these coordinates to make them invariant to location. Algorithm 1 illustrates the process for coordinates scaling, where it takes the input as a skeleton with a set of keyjoints and generates output as scaled keyjoints. The scaling process involves positioning the hip joint to the origin (0, 0). This is achieved by subtracting the hip joint's coordinates from the coordinates of all other body joints.

Furthermore, the scaled body offset skeleton needs to be normalized to make the coordinates invariant to size. In this process, one skeleton from a sequence is taken as a reference, and all other skeletons in that sequence are normalized so that their part lengths match the corresponding lengths of the reference skeleton. The normalized skeleton keyjoints are calculated by

**Algorithm 1:** Coordinate Scaling via Hip Positioning.

**Input:** Skeleton $S$ with selected $i$ body joints
$S = \{x_0, y_0, x_1, y_1, x_2, y_2, ..., x_i, y_i\}$
**Output:** Scaled coordinate of skeleton $S$

```
function SCALE_COORDINATES(S)
    S ← S.copy()
    x_h, y_h ← evaluate_joint(S, HIP)
    for j in [0, i] do
        S[2j] ← S[2j] − x_h        ▷ Scale x-coordinates
        S[2j+1] ← S[2j+1] − y_h    ▷ Scale y-coordinates
    end for
    return S
end function
```

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 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.



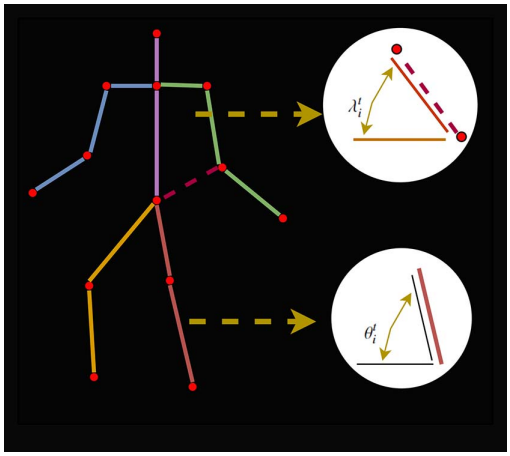



Fig. 3. Systematic diagram for the angle between each joint and the angle between each joint to relative joint.

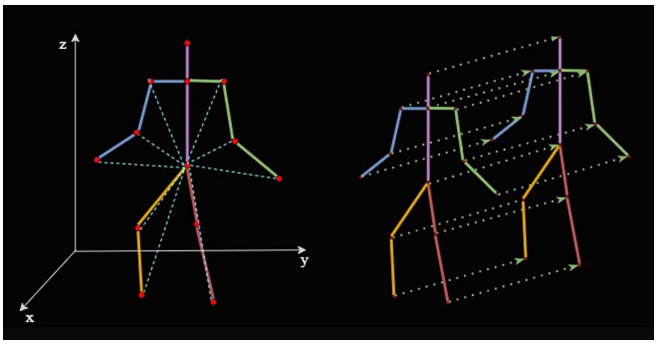



Fig. 4. Illustration of relative joint coordinates and velocity.

normalizing each skeleton keyjoints for $N$ skeletons. It is calculated by dividing the body offset of each skeleton by the average height of the $N$ skeletons, where $N$ is the number of preprocessed skeleton frames that encode an entire human activity. The Euclidean distance between the head and ankle is calculated to determine the height of each skeleton.

### B. Feature Evaluation and Matrix Encoding

This part uses pose-encoded skeleton data with 13 keyjoints each to extract the characteristics. In this case, $N$ preprocessed skeleton frames that encode an entire human act are represented by a window of size $N$. Given $X = \{S_0, S_1, S_2, S_3, ..., S_N\}$, the set of skeletons for $N$ frames is represented by $X$. The pair of coordinates $(x_i, y_i)$ are used to represent each keyjoint in the skeleton.

The proposed model first determines the missing entries in the feature matrix $F$ to handle these missing values and replaces the missing coordinates with 0. Then, extract four different features as shown in Figs. 3 and 4 and construct the feature vector in the following steps.

1) Calculate relative joint coordinates $R_i$ represents a spatial relationship with joint $i$ relative to a reference joint.
2) Compute geometric features to capture orientation and distance $\gamma$ between joints and then angles $\lambda_{ij}^t$ and $\theta_i^t$.
3) Compute the motion feature for capturing the velocity of each joint $i$ within the consecutive frames i.e., $t$ and $t+1$.
4) Combine all the feature vectors into a single feature matrix $F$. The process of calculating features using $N$ frames to obtain unique features and their body keyjoints from the skeletons data is discussed in Eqs. (1)–(10).

The relative joints $\mathbf{r}_i^t$ represent the relationship between each joint to reference joint (i.e. hip joint) as shown in

$$\mathbf{r}_i^t = (x_i^t - x_{\text{hip}}^t, y_i^t - y_{\text{hip}}^t). \tag{1}$$

The feature vector for relative coordinates

$$R^t = \begin{bmatrix} r_1^t \\ r_2^t \\ \vdots \\ r_n^t \end{bmatrix}. \tag{2}$$

Distance $\gamma_i^t$ between each pair of joints$(i,j)$ is

$$\gamma_i^t = \|(x_i^t, y_i^t) - (x_j^t, y_j^t)\| = \sqrt{(x_i^t - x_j^t)^2 + (y_i^t - y_j^t)^2}. \tag{3}$$

The angle $\lambda_{ij}^t$ of each joint $i$ to horizontal axis

$$\lambda_{ij}^t = \arccos\left(\frac{x_j^t - x_i^t}{\sqrt{(x_i^t - x_j^t)^2 + (y_i^t - y_j^t)^2}}\right). \tag{4}$$

The feature vector for angle

$$A_\lambda^t = \begin{bmatrix} \lambda_1^t \\ \lambda_2^t \\ \vdots \\ \lambda_n^t \end{bmatrix}. \tag{5}$$

The angle $\theta_i^t$ of each relative joint to the reference center point

$$\theta_i^t = \arccos\left(\frac{x_i^t - x_{\text{hip}}^t}{\sqrt{(x_i^t - x_{\text{hip}}^t)^2 + (y_i^t - y_{\text{hip}}^t)^2}}\right). \tag{6}$$

The feature vector for angle

$$A_\theta^t = \begin{bmatrix} \theta_1^t \\ \theta_2^t \\ \vdots \\ \theta_n^t \end{bmatrix}. \tag{7}$$

The velocity $\mathbf{v}_i^t$ of joint $i$ between frames $t$ and $t+1$

$$\mathbf{v}_i^t = \left(\frac{x_i^{t+1} - x_i^t}{\Delta t}, \frac{y_i^{t+1} - y_i^t}{\Delta t}\right). \tag{8}$$

The feature vector for velocity

$$V^t = \begin{bmatrix} v_1^t \\ v_2^t \\ \vdots \\ v_n^t \end{bmatrix}. \tag{9}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 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.



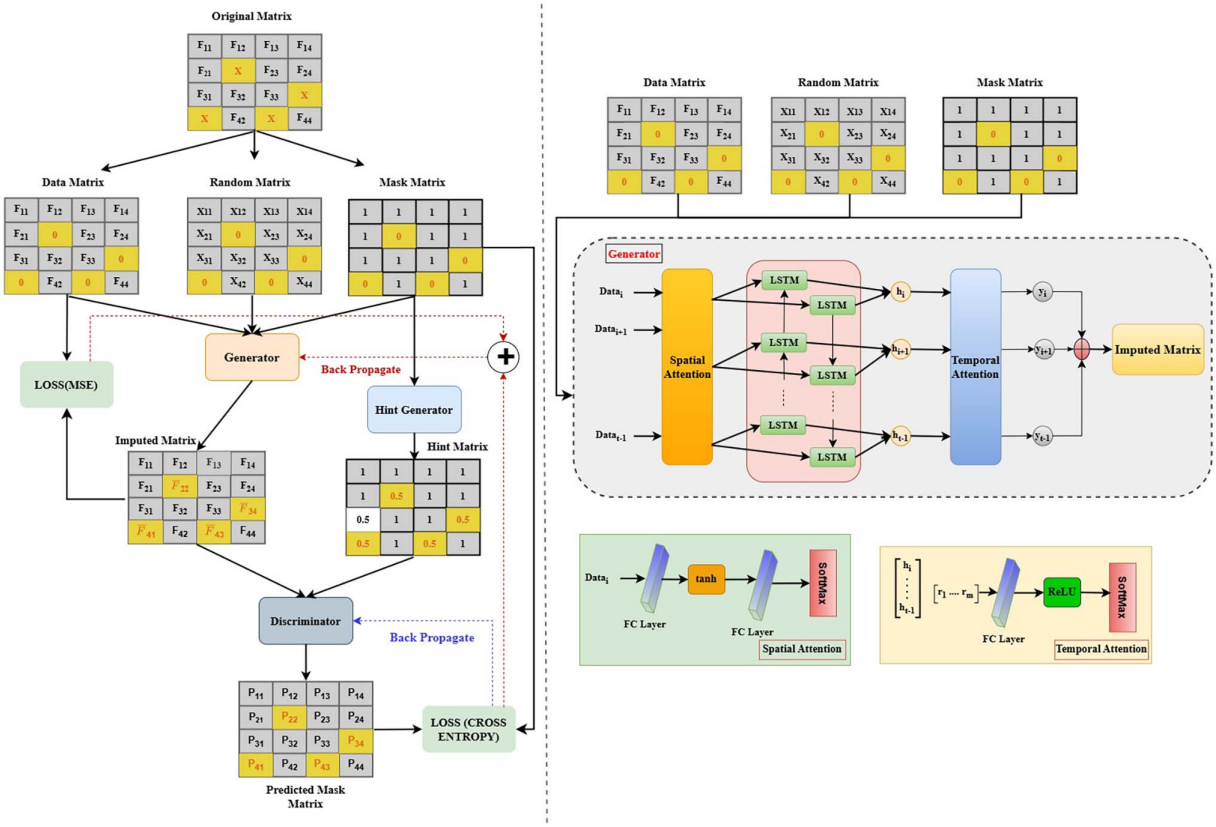



Fig. 5. Architecture of the proposed data imputation model with Bi-LSTM with spatial and temporal attention modules. It consists of three main parts: (a) generator; (b) discriminator; and (c) hint matrix. The modified generator comprises a Bi-LSTM network with spatial and temporal attention modules.

Combine all the feature vectors into a single feature matrix $F$ for each frame $t$

$$F = [R^t, \quad A^t_\lambda, \quad A^t_\theta, \quad V^t]. \tag{10}$$

### C. Missing Data Imputation Using GAIN

Due to occlusion, some features are missing in the feature matrix $F$. To deal with these missing values, there are some existing methods, like simple imputation methods, that use mean, median, and mode to fill in the missing values. Another technique uses machine learning-based models to handle the missing data, such as XGBoost [25], and lightGBM [26]. The next technique is advanced statistical methods that use the most similar features to fill the missing values, such as KNN [27]. These methods do not fill in the missing values or recover the original features very well when it comes to complex classification. To deal with this problem, it requires some advanced deep learning-based methods that can calculate the true values and fill the space for missing values.

There are some deep learning models, like GAN, used for data imputation in many problems. Goodfellow originally proposed GAN in 2014 [28] to design realistic images which consist of a discriminator and a generator. The generator generates data that resembles the source data, and the discriminator distinguishes among generated and real data. Through adversarial training, the discriminator learns to accurately classify samples from the generator, pushing it to improve the output to deceive the discriminator. This adversarial process continues until the generator creates data that closely aligns with the real data distribution. The GAIN [29] model is a variant of GAN that is used for missing data imputation.

The proposed model is based on the temporal GAIN and includes a number of improvements and modifications to increase its efficacy. The proposed framework consists of three main components as depicted in Fig. 5: (1) generator; (2) discriminator; and (3) hint generator. The generator component takes the original data vector with missing values, observes each input data component, and imputes the missing data based on observations, then gives the complete data vector. In order to focus on the missing data, the generator consists of a Bi-LSTM network with spatial and temporal attention modules. The spatial attention processes the input data to generate spatial attention weights and emphasizes the important components. The temporal attention module processes the hidden states of the BiLSTM network to generate temporal attention weights and emphasizes the important time steps. The discriminator evaluates the imputed matrix and differentiates between the actual observed data and the data generated by the generator. It utilizes a hint matrix containing additional information to determine the appropriate distribution in the generator component, providing insight into the missing section of original data that helps the discriminator focus more on it and forcing the generator to produce a more accurate imputation aligned with the actual data. The Hint generator generates the hint matrix that depends on the mask matrix, where 1 represents the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 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.



**Algorithm 2:** Spatio-Temporal GAIN-Based Data Imputation.

**Input:** Incomplete feature matrix $F$, Mask matrix $M$
$F = \{f_1, f_2, ..., f_t\}$, where $f_t \in \mathbb{R}^n$ and $n$ is the number of features at time step $t$.
$M = \{m_1, m_2, ..., m_t\}$, where $m_t \in \{0,1\}^n$ indicates observed (1) or missing (0) features.
**Output:** Imputed feature matrix $\hat{F}$

**Initialize:** Random noise vector $Z \sim \mathcal{N}(0,1)$
**Generator:**
**for** each time step $t$ **do**
  $f'_t \leftarrow M \odot f_t + (1-M) \odot G(f_{1:t-1}, Z)$ ▷ Impute missing features using Bi-LSTM with spatial and temporal attention
  $\hat{F} \leftarrow \hat{F} \cup f'_t$
**end for**
**Hint Generator:**
**for** each element $m_{ij} \in M$ **do**
  **if** $m_{ij} == 1$ **then**
    $h_{ij} \leftarrow 1$ ▷ Observed features
  **else**
    $h_{ij} \leftarrow 0.5$ ▷ Missing features
  **end if**
**end for**
**Discriminator:**
**for** each time step $t$ **do**
  $D(h_t, \hat{f}_t) \leftarrow$classify real and imputed data using the hint matrix
**end for**
**Adversarial Training:**
**while** not converged **do**
  Minimize $\mathcal{L}_G = \text{MSE}(F, \hat{F}) + \lambda \cdot$ Adversarial loss
  Maximize $\mathcal{L}_D =$ Binary cross-entropy loss
**end while**
**Output:** Imputed data matrix $\hat{F}$

retaining data, and 0 represents the missing data in the original matrix. When the hint matrix represents 1, it reveals the accuracy of the given data to the discriminator, and when $h_i$ represents 0.5, it indicates the missing data. The generator is trained by reducing the mean square error (MSE) of imputed features of the $t$th frame and the missing features. It uses features from $i$th frame to $t-1$ frame as input to obtain the imputed features of $t$th frame. $[r_1 \cdot r_m]$ is an $m$-dimensional random vector used in the temporal attention module, where m is the dimension of data and the $r_m$, and the vector $r_m$ acquired during training represents the weight of the input data. Algorithm 2 illustrates the processing of GAIN with Bi-LSTM and spatiotemporal attention module.

The proposed algorithm demonstrates strong performance in imputing missing data by integrating Bi-LSTM and spatiotemporal attention mechanisms. This enables effective learning of both temporal dependencies and spatial relationships, ensuring robust handling of occlusions. The adversarial training framework (using GAIN) further enhances the imputation quality by leveraging the discriminator to guide the generator, resulting in an accurate reconstruction of the missing data. The algorithm's ability to work with sequential data makes it well-suited for activity recognition tasks, especially under occlusion scenarios.

The main components to determine the computational complexity of the proposed algorithm are: the generator, hint generator, discriminator, and adversarial training. The generator uses Bi-LSTM with spatial and temporal attention and iterates across $t$ time steps for $n$ number of features. Because of sequential dependencies, the complexity is $O(t^2.n)$. The hint generator gives $O(t.n)$ by processing the mask matrix $M$ element-by-element, whereas the discriminator distinguishes real and imputed data at each time step with an identical complexity of $O(t.n)$. The complexity of adversarial training is $O(k(t^2.n + t.n))$, for $k$ iterations and loss computations for both the discriminator and generator. Therefore, the overall complexity of the algorithm is $O(t^2.n)$.

### D. Activity Calculation

The imputed feature vector obtained by the GAIN is processed by the ABSA module to classify the activities followed by the softMax activation function. The visual representation of this module is shown in the classification part of the proposed framework Fig. 1. The spatial attention processes each input feature to aim at the most relevant features for each time step and feeds it to the Bi-LSTM encoder. Bi-LSTM encoder integrates the contextual information in forward as well as backward direction. It consists of two LSTMs for forward & backward directions. The forward one processes the input in the same sequence ($D_1$ to $D_N$), and the backward one processes the input in the reverse sequence ($D_N$ to $D_1$). The temporal attention layer applies to the encoder's hidden states to generate a context vector highlighting important time steps. Bi-LSTM decoder takes the context vector and the hidden states, processes them using bidirectional LSTMs, and generates the predicted outputs.

## IV. Experimental Result and Analysis

### A. Performance Evaluation

The experimental setup is configured using Python version 3.7 with CUDA 10.2 and CuDNN 7.6.5. The experimentation is performed on a system with a Ryzen 7 octa-core CPU, 16 GB of RAM, 1 TB of SSD, and a GeForce RTX 4050 GPU with 6 GB of memory. The operating environment is Windows 11. During the training of the model, a total of 200 epochs are executed. The learning rate is set at 0.001. The datasets are split into two sets: training (70%) and testing (30%) sets.

Performance metrics such as precision, recall, accuracy, and loss evaluate a classification model's efficiency. Precision measures how accurate positive predictions are by dividing the number of accurately identified positive instances by the total number of positive instances. Recall measures how well a model detects the number of accurately classified positive instances by the total number of real positive instances. Accuracy metrics consider all the correct classified instances divided by the total number of instances. Finally, loss quantitatively measures the difference between predicted and actual values, guiding model optimization during training. In this work, accuracy and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 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.



loss are used to evaluate the performance of the proposed framework.

### B. Datasets

*1) NTU RGB+D 60:* NTU RGB+D 60 [30] is a publicly available large-scale dataset of 56,880 activity video samples for human activity recognition. This dataset consists of 60 activity classes with three different viewpoints performed by 40 subjects. Each activity video includes skeleton data, RGB video, depth map sequences, and infrared video captured by Kinect v2 cameras. In this work, skeleton data is used to perform the experiments. Each skeleton has a total of 25 joints. The cross-subject (x-sub) and cross-view (x-view) benchmarks provided by NTU-RGB-D contain distinct classifications for the test and train split of the total clips. 20 out of 40 participants were assigned to the training dataset and another 20 to the testing sets.

*2) NTU RGB+D 120:* The NTU RGB+D 120 is a publicly available large dataset and an upgraded version of the NTU RGB+D 60 dataset. This dataset added an extra 57,600 video clips of 60 activities in the NTU RGB+D 60 dataset and has a total of 114 480 video sequences of 120 different classes. Two test benchmarks for the NTU-RGB+D 120 dataset, cross-setting (CSET) and cross-subject (CSUB) recommended by Shahroudy et al. [31]. A total of 59 477 samples were selected for testing and 54 468 video sequences for training.

*3) PKU-MMD:* The PKU-MMD [32] is a publicly available dataset used to detect and recognize human activity. This dataset consists of 1076 video sequences and 51 activity classes performed by 66 subjects. It includes RGB, depth, infrared radiation, and skeleton video captured by Kinect v2 cameras. There are two subsets of the PKU-MMD dataset. Part 1 has 706 videos with 51 classes, whereas Part 2 consists of 370 videos of 50 classes.

*4) Synthesized Dataset:* The partial occlusion dataset was constructed to simulate real-world scenarios where objects are occluded. To better address the complexities of real scenes, we synthesized various large-scale occluded skeleton data. The dataset was developed in our laboratory, on our university campus, and in various indoor and outdoor locations using a 12-megapixel camera. Videos were captured at a frame rate of 30 fps with $640 \times 480$ resolution. The dataset comprises two classes, including both normal and abnormal activities. The normal class contains daily living activities (running, sitting, eating, walking), and the abnormal class contains some unusual activities (fighting, vandalism, shoplifting, threats). Each class contains 70 videos, each lasting 8 s. The sample instances of the synthesized dataset are shown in Fig. 6.




Fig. 6. Sample images of synthesized dataset.

### C. Experimental Results on Unoccluded Datasets

This article compares the proposed methodology and existing state-of-the-art (SOTA) activity recognition methods on the NTU-RGB+D 60 datasets, NTU-RGB+D 120 datasets, and PKU-MMD dataset. Recent activity recognition algorithms use a variety of strategies to encode temporal and spatial streams. A sequential 3-DCNN network is used by method [33] to capture action dynamics. MobileNetV2 and GRU are combined by method [3] to extract both spatial and deep features. CNN+LSTM approach is used by method [34] to recognize abnormal activities, among other methods. These techniques are used to represent activities in various contexts. On the other hand, our proposed method extracts spatial and motion streams for activity representation by using a CNN and autoencoder-based Bi-LSTM with an attention module. Using three unoccluded datasets, Table II examines the different approaches and demonstrates the effectiveness of the proposed method evaluated against other approaches. More precisely, the proposed approach achieves impressive accuracy rates of 93.53% for the NTU-RGB+D 60 dataset's CV benchmark and 94.82% and 97.7% for the NTU-RGB+D 120 dataset's CSET and CSUB benchmarks, respectively. Tariq et al. [3] achieves an optimal accuracy of 95.38% for the CS benchmark of the NTU-RGB+D 60 dataset, while on the PKU-MMD dataset, Sharma and Singh [35] obtains 93.46% accuracy. After extensive analysis, we determined that the proposed method demonstrates significant

TABLE II
ANALYSIS OF PROPOSED EAGLE FRAMEWORK WITH EXISTING TECHNIQUES ON THE UNOCCLUDED DATASETS (NTU RGB+D 60, NTU RGB+D 120, AND PKU-MMD)

| Methods/ Datasets | NTU RGB+D 60 | | NTU RGB+D 120 | | PKU-MMD |
|---|---|---|---|---|---|
| | CS | CV | CSET | CSUB | CV |
| Sharma and Singh [35] | 91.72 | 90.5 | 92.56 | 91.37 | **93.46** |
| Ansari et al. [33] | 90.95 | 91.4 | 94.1 | 93.21 | 89.63 |
| Tariq et al. [3] | **95.38** | 93.17 | 92.85 | 96.3 | 91.26 |
| Jayaswal et al. [34] | 89.89 | 91.53 | 90.5 | 91.1 | 93.38 |
| **Proposed method** | 95.29 | **93.53** | **94.82** | **97.7** | 93.39 |

Note: Bold values indicate the highest validation accuracy for each respective comparison.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 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.



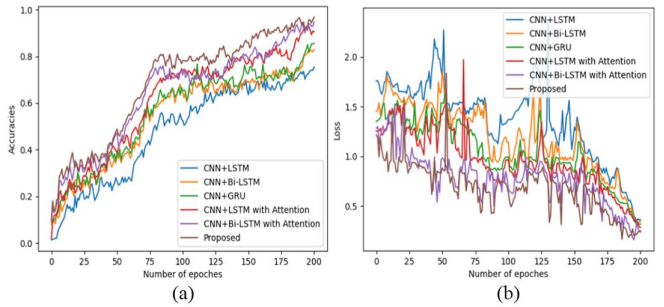



Fig. 7. Tradeoff between: (a) validation accuracy with baseline methods and (b) validation loss with baseline methods for NTU RGB+D 60 dataset.

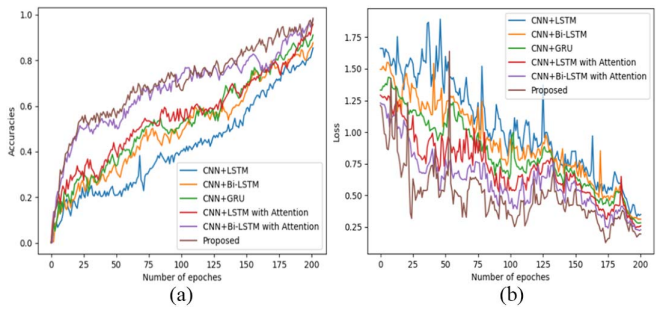



Fig. 8. Tradeoff between: (a) validation accuracy with baseline methods; and (b) validation loss with baseline methods for NTU RGB+D 120 dataset.

improvements in accuracy and a significant level of efficiency over the SOTA methods.

### D. Experimental Results on Occluded Datasets

There is frequent variation and unpredictability in occlusion during the activity sequences. There are some unreliable factors of occlusion: the part where occlusion occurs, the moment at which it occurs, and the duration of occlusion. To our knowledge, no publicly accessible dataset for human skeleton activity recognition has been developed especially to evaluate the effects of partial occlusion. Therefore, to conduct experiments and assess the effectiveness of the proposed model against other approaches, we created a synthetic occluded data set that includes the CS from NTU RGB+D 60, CSET from NTU RGB + D 120 benchmarks, and the PKU-MMD data set. This article evaluates various baseline methods on these datasets along with the synthesized dataset. Meanwhile, it also compares the proposed model with baseline methods i.e., CNN+LSTM, CNN+Bi-LSTM, CNN+ GRU, CNN+LSTM with attention, and CNN+Bi-LSTM with attention.

*1) NTU RGB+D 60:* The tradeoff graph in Fig. 7 represents the validation accuracy and loss for each baseline method compared to the proposed method on the occluded NTU RGB+D 60 dataset. According to the trade-off graphs, the proposed approach performs more accurately than the other baseline approaches.

*2) NTU RGB+D 120:* The tradeoff graph in Fig. 8 represents the validation accuracy and loss for each baseline method compared to the proposed method on the occluded NTU RGB+ D 120 dataset. The graphical representation of validation accuracy and loss demonstrates that the proposed method is more accurate than other baseline methods.

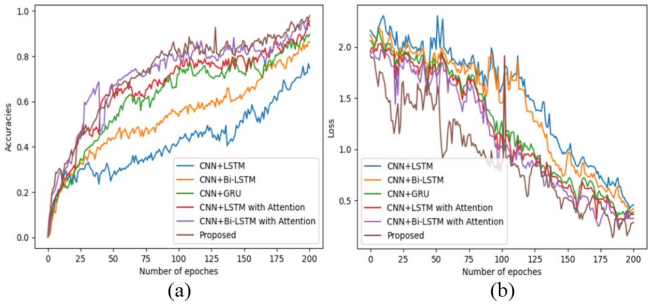



Fig. 9. Tradeoff between: (a) validation accuracy with baseline methods; and (b) validation loss with baseline methods for PKU-MMD dataset.

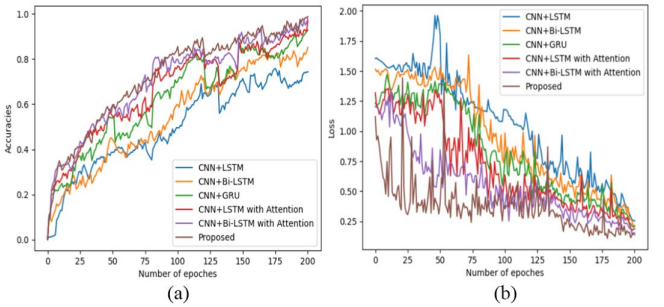



Fig. 10. Tradeoff between: (a) validation accuracy with baseline methods; and (b) validation loss with baseline methods for synthesized dataset.

*3) PKU-MMD:* The tradeoff graph in Fig. 9 represents the validation accuracy and loss for each baseline method compared to the proposed method on the occluded PKU-MMD dataset. The graphical representation of the validation accuracy and loss illustrates that the proposed method is more accurate than other baseline methods.

*4) Synthesized Dataset:* The tradeoff graph in Fig. 10 represents the validation accuracy and loss for each baseline method compared to the proposed method on the occluded synthesized dataset. The graphical representation of the validation accuracy and loss demonstrates that the proposed method is more accurate than other baseline methods.

Table III compares the validation accuracy of the proposed method with all baseline methods for the occluded NTU RGB+D 60 dataset, NTU RGB+D 120 dataset, PKU-MMD dataset, and the synthesized dataset, clearly showing that the proposed method outperforms the others.

### E. Comparison With the Existing SOTA Methods

We performed a quantitative comparison with multiple SOTA methodologies to illustrate the efficacy of the proposed method. The comparative analysis for the NTU RGB+D 60, NTU RGB+D 120, PKU-MMD, and synthetic datasets is displayed in Table IV. The proposed method produced the best possible level of recognition accuracy. With rates of 96.71% for the NTU RGB+D 60, 98.3% for the NTU RGB+D 120, 97.9% for the PKU-MMD, and 98.7% for the synthesized dataset, the proposed method demonstrated remarkable accuracy.

### F. Experimental Result for Cross Datasets

To evaluate the performance of the proposed approach in various data sets, this article performed cross-data set tests in which

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



TABLE III
COMPARISON OF VALIDATION ACCURACIES BETWEEN SOTA WITH BASELINE METHODS FOR NTU RGB+D 60 DATASET, NTU RGB+D 120 DATASET, PKU-MMD DATASET AND SYNTHESIZED DATASET

| Datasets | Methods | Accuracy |
|---|---|---|
| NTU RGB+D 60 | CNN+LSTM | 0.754 |
| | CNN+Bi-LSTM | 0.829 |
| | CNN+GRU | 0.855 |
| | CNN+LSTM with attention | 0.905 |
| | CNN+Bi-LSTM with attention | 0.946 |
| | **CNN+autoencoder Bi-LSTM with attention** | **0.967** |
| NTU RGB+D 120 | CNN+LSTM | 0.853 |
| | CNN+Bi-LSTM | 0.876 |
| | CNN+GRU | 0.911 |
| | CNN+LSTM with attention | 0.957 |
| | CNN+Bi-LSTM with attention | 0.976 |
| | **CNN+autoencoder Bi-LSTM with attention** | **0.983** |
| PKU-MMD | CNN+LSTM | 0.749 |
| | CNN+Bi-LSTM | 0.864 |
| | CNN+GRU | 0.897 |
| | CNN+LSTM with attention | 0.934 |
| | CNN+Bi-LSTM with attention | 0.953 |
| | **CNN+autoencoder Bi-LSTM with attention** | **0.979** |
| Synthesized | CNN+LSTM | 0.742 |
| | CNN+Bi-LSTM | 0.850 |
| | CNN+GRU | 0.927 |
| | CNN+LSTM with attention | 0.968 |
| | CNN+Bi-LSTM with attention | 0.979 |
| | **CNN+autoencoder Bi-LSTM with attention** | **0.987** |

Note: Bold values indicate the highest validation accuracy for each respective comparison.

the model was trained in the NTU RGB+D 60 dataset and tested in the PKU-MMD dataset. The accuracy graph in Fig. 11 shows the trade-off between the training accuracy in the NTU RGB+D 60 dataset and the testing accuracy in the PKU-MMD dataset, and the loss graph shows the training loss in the NTU RGB+D 60 dataset and the testing loss in the PKU-MMD dataset. The analysis demonstrates that the proposed approach is still successful in identifying activities across datasets, even if domain changes cause a slight decrease in performance.

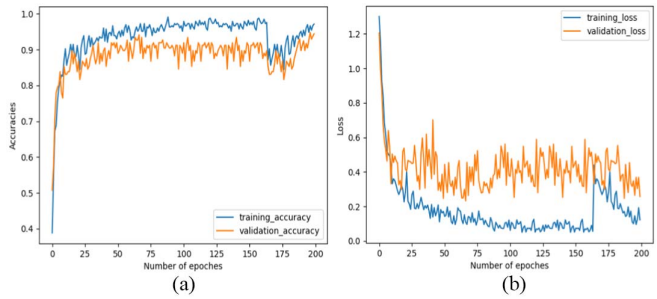
### G. Analysis of Computation Time

The comparison between computation time for NTU RGB+D 60, NTU RGB+D 120, PKU-MMD, and Synthesized datasets for various batch sizes is depicted in Fig. 12. It is clearly observed that larger batches are more effective in reducing iterations since batch size 2 consistently takes the highest computational time. PKU-MMD takes higher computational time than others, and the synthesized dataset processes faster than others.

TABLE IV
QUANTITATIVE ANALYSIS OF PROPOSED EAGLE FRAMEWORK WITH SOTA TECHNIQUES ON THE OCCLUDED NTU RGB+D 60 DATASET, NTU RGB+D 120 DATASET, PKU-MMD DATASET AND SYNTHESIZED DATASET

| Dataset | Authors | Methods | Accuracy |
|---|---|---|---|
| NTU RGB+D 60 | Chen et al. [36] | GCN+LSTM | 95.29 |
| | Li et al. [7] | CNN+GAIN+Resnet50 | 88.45 |
| | Ansari and Singh [21] | FNNN+PCA | 89.40 |
| | Shi et al. [37] | GCN+cross entropy | 91.90 |
| | **Proposed model** | **CNN+GAIN+autoencoder Bi-LSTM with attention** | **96.71** |
| NTU RGB+D 120 | Chen et al. [36] | GCN+LSTM | 97.50 |
| | Li et al. [7] | CNN+GAIN+Resnet50 | 90.17 |
| | Ansari and Singh [21] | FNNN+PCA | 91.59 |
| | Shi et al. [37] | GCN+Cross entropy | 92.47 |
| | **Proposed model** | **CNN+GAIN+autoencoder Bi-LSTM with attention** | **98.3** |
| PKU-MMD | Chen et al. [36] | GCN+LSTM | 95.10 |
| | Li et al. [7] | CNN+GAIN+Resnet50 | 90.29 |
| | Ansari and Singh [21] | FNNN+PCA | 94.36 |
| | Shi et al. [37] | GCN+Cross entropy | 92.27 |
| | **Proposed model** | **CNN+GAIN+autoencoder Bi-LSTM with attention** | **97.9** |
| Synthesized | Chen et al. [36] | GCN+LSTM | 96.45 |
| | Li et al. [7] | CNN+GAIN+Resnet50 | 92.06 |
| | Ansari and Singh [21] | FNNN+PCA | 94.70 |
| | Shi et al. [37] | GCN+Cross entropy | 93.59 |
| | **Proposed model** | **CNN+GAIN+autoencoder Bi-LSTM with attention** | **98.7** |

Note: Bold values indicate the highest validation accuracy for each respective comparison.




(a) (b)

Fig. 11. Tradeoff between: (a) training accuracy for NTU RGB+D 60 and validation accuracy for PKU-MMD dataset; and (b) training loss for NTU RGB+D 60 and validation loss for PKU-MMD dataset.

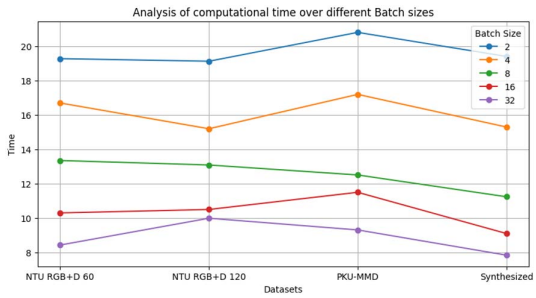



Fig. 12. Analysis of computation time over different batch sizes.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 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.



## V. Conclusion

This article proposed an efficient framework to address the problem of activity recognition under different occlusion situations. We provide a method that integrates advanced approaches to feature extraction, data imputation, and activity classification to improve the performance and robustness of activity recognition systems in occluded environments. The model processes skeleton keyjoints using a CNN-based algorithm to extract features such as relative joint coordinates, angles, and velocity from the N-encoded skeletons and construct a feature matrix. The innovative utilization of GAIN with Bi-LSTM augmented by spatial and temporal attention efficiently handles the missing feature data and imputes the missing values. The proposed ABSA learns the spatial and temporal dependencies to refine the activity classification further. We performed quantitative analysis on three benchmark datasets and our synthesized dataset in the unclouded and occluded scenarios. This article also compares the proposed model with existing SOTA approaches. It demonstrates that the proposed method outperforms the existing SOTA techniques and achieves the best possible recognition accuracy in different occluded environments. It has been observed throughout the experimentation that the occluded upper body part of the skeleton significantly affects the accuracy of the model as compared to the occluded lower part. We can use it in the future to further enhance our work and increase the accuracy of our recognition system.

## References


[1] N. Nigam and T. Dutta, "Emotion and gesture guided action recognition in videos using supervised deep networks," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 5, pp. 2546–2556, 2023.

[2] K. Gościewska and D. Frejlichowski, "Action classification for partially occluded silhouettes by means of shape and action descriptors," *Appl. Sci.*, vol. 11, no. 18, 2021, Art. no. 8633.

[3] T. Ahmad and J. Wu, "Sdigru: Spatial and deep features integration using multilayer gated recurrent unit for human activity recognition," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 1, pp. 973–985, 2024.

[4] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[5] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer Int. Publishing, 2016, pp. 483–499.

[6] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,* 2019, pp. 7753–7762.

[7] Z. Li and D. Li, "Action recognition of construction workers under occlusion," *J. Building Eng.*, vol. 45, 2022, Art. no. 103352.

[8] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view human action recognition under occlusion based on fuzzy distances and neural networks," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO).* Piscataway, NJ, USA: IEEE Press, 2012, pp. 1129–1133.

[9] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing human video pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2016, pp. 3063–3072.

[10] Y. Chen, L. Wang, D. Hu, and H. Cheng, "Multi-view graph convolution network for the recognition of human action with spatial and temporal occlusion problems," *J. Vis. Commun. Image Represent.*, vol. 97, 2023, Art. no. 103957.

[11] P. Nangtin, P. Kumhom, and K. Chamnongthai, "Gait identification with partial occlusion using six modules and consideration of occluded module exclusion," *J. Vis. Commun. Image Represent.*, vol. 36, pp. 107–121, 2016.

[12] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 12–23, 2014.

[13] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.,* 2017, pp. 2640–2649.

[14] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3d pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[15] Y. Cheng, B. Wang, B. Yang, and R. T. Tan, "Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, pp. 1157–1165, 2021.

[16] M. R. I. Hossain and J. J. Little, "Exploiting temporal information for 3d human pose estimation," in *Proc. Eur. Conf. Comput. Vision (ECCV),* 2018, pp. 68–84.

[17] P. Khan, Y. Kumar, and S. Kumar, "Capslstm-based human activity recognition for smart healthcare with scarce labeled data," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 1, pp. 707–716, Jan. 2024.

[18] L. Zhou, Y. Chen, Y. Gao, J. Wang, and H. Lu, "Occlusion-aware Siamese network for human pose estimation," in *Proc. Computer Vision–(ECCV): 16th Eur. Conf.*, Glasgow, UK. New York: Springer, 2020, pp. 396–412.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances Neural Inf. Process. Syst.*, vol. 27, 2014.

[20] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2012.

[21] M. A. Ansari and D. K. Singh, "Identifying human activities in megastores through postural data to monitor shoplifting events," *Neural Comput. Appl.*, vol. 35, no. 9, pp. 6515–6528, 2023.

[22] V.-T. Le and Y.-G. Kim, "Attention-based residual autoencoder for video anomaly detection," *Appl. Intell.*, vol. 52, no. 5, pp. 1–15, 2022.

[23] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[24] A. Taha, H. H. Zayed, M. Khalifa, and E.-S. M. El-Horbaty, "Skeleton-based human activity recognition for video surveillance," *Int. J. Sci. Eng. Res.*, vol. 6, no. 1, pp. 993–1004, 2015.

[25] T. Chen et al., "Xgboost: Extreme gradient boosting," *R Package Version 0.4-2*, vol. 1, pp. 1–4, 2015.

[26] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neur. Inf. Process. Syst.*, vol. 30, 2017.

[27] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[28] I. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[29] J. Yoon, J. Jordon, and M. Schaar, "Gain: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. on Mach. Learn.* PMLR, 2018, pp. 5689–5698.

[30] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3d human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2016, pp. 1010–1019.

[31] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+ D 120: A large-scale benchmark for 3d human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, 2019.

[32] L. Chunhui, H. Yueyu, L. Yanghao, S. Sijie, and L. Jiaying, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," 2017, *arXiv:170307475*.

[33] M. Ansari, D. Singh, and V. Singh, "Detecting abnormal behavior in megastore for crime prevention using a deep neural architecture," *Int. J. Multimedia Inf. Retrieval*, vol. 12, 2023, Art. no. 8.

[34] R. Jayaswal and M. Dixit, "Human Behavior Dataset (hbd21)," 2021.

[35] A. Sharma and R. Singh, "ConvST-LSTM-Net: Convolutional spatiotemporal LSTM networks for skeleton-based human action recognition," *Int. J. Multimed. Inf. Retr.*, vol. 12, no. 2, 2023.

[36] Y. Chen, L. Wang, D. Hu, and H. Cheng, "Multi-view graph convolution network for the recognition of human action with spatial and temporal occlusion problems," *J. Vis. Commun. Image Represent.*, vol. 97, 2023, Art. no. 103957.

[37] W. Shi, D. Li, Y. Wen, and W. Yang, "Occlusion-aware graph neural networks for skeleton action recognition," *IEEE Trans. Ind. Informat.*, vol. 19, no. 10, pp. 10 288–10 298, 2023.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:07:41 UTC from IEEE Xplore. Restrictions apply.

Wireless Networks
https://doi.org/10.1007/s11276-025-03970-y

ORIGINAL PAPER

# EECRPOA: energy efficient clustered routing using pelican optimization algorithm in WSNs

**Yogendra Mohan**[1,2] · **Rajesh Kumar Yadav**[1] · **Manisha Manjul**[3]

Accepted: 5 May 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

**Abstract**
Energy efficiency and network life are crucial concerns in Wireless Sensor Networks (WSNs) due to the limited battery of sensor nodes. Cluster Head (CH)-based routing is a prominent strategy that aims to mitigate and tackle these overheads by dividing the network into clusters. The CHs gather data from the cluster members (CMs) and send data to the base station (BS) using the optimal path. This helps to minimize the amount of data transmission required, and the model is energy-efficient. In this paper, for the cluster head selection (CHS) and routing, a Pelican Optimization Algorithm (POA) is modelled with suitable fitness functions. The CHS algorithm is modelled using five fitness functions (FFs) named node centrality, the residual energy (RE) of the CH, the distance between cluster members (CMs) and CH, and the distance between CHs and BS. For each CH, a potential function is developed to allocate CMs. The proposed POA-based routing scheme is modelled using three fitness functions to send data to the BS. This work presented a comprehensive examination of LEACH, DEEC, BOAACO, and Airproofed. The CHS and routing algorithms are investigating their influence on network performance, encompassing energy consumption and network longevity. The simulation findings confirm that the proposed scheme significantly improves the overall efficiency and robustness of WSNs when compared to LEACH, DEEC, and BOAACO.

**Keywords** Cluster head selection · Routing · Meta-heuristic optimization algorithm (MHOA) · Energy-efficient · WSNs

## 1 Introduction

WSNs are made up of small, low-power, distributed sensors that connect wirelessly to collect data from their surroundings. Sensors, microcontrollers, and wireless transceivers are all found in these nodes. This data is collected and processed by a central base station (BS). The most common applications of WSNs are environmental monitoring, healthcare, automation, smart homes, and smart cities. Because battery-powered nodes necessitate data aggregation and efficient communication, energy efficiency is a critical challenge. Energy management, scalability, reliability, security, and maintenance are the challenges. Despite these obstacles, WSNs are critical for real-time data collecting and monitoring in a variety of applications. Clustering [1, 2] is one of the most basic methods for extending the lifespan and increasing the usefulness of wireless sensor networks (WSNs). However, selecting CHs [3, 4] for energy stabilization to increase network life expectancy is still a significant problem in WSNs. Nodes self-organize [5–7] into clusters is called cluster formation [8]. Several techniques, like LEACH [50], DD, PEGASIS, HEED, and GPSR, can be employed for cluster formation. A CH gathers data from CMs and transmits data to the BS. The energy level, distance to the BS, or a combination of both parameters are frequently considered while choosing the CH. Data fusion and transmission are typically handled by CHs by reducing the size of data delivered to the base station, data aggregation helps conserve bandwidth and energy (Table 1).

The scalability is improved by clustering. As the network expands, adding more nodes or clusters becomes relatively simple because CHs administer their clusters, which reduces

✉ Yogendra Mohan
yogendra.mohan@gmail.com

Rajesh Kumar Yadav
rkyadav@dtu.ac.in

Manisha Manjul
bhavikamanisha@gmail.com

[1] Delhi Technological University, Delhi, India

[2] Department of Computer Science and Engineering, NERIST, Nirjuli, Itanagar, India

[3] G.B. Pant DSEU Okhala-1 Campus, New Delhi, India

Published online: 15 May 2025

**Table 1** Abbreviation and description

| Abbreviation | Description | Abbreviation | Description |
|---|---|---|---|
| *ABC* | Artificial bee colony | $ff_R$ | Fitness Function for Routing |
| *ABCOA* | Artificial Bee Colony Optimization Algorithm | *FND* | First node dead |
| *ACO* | Ant-Colony Optimization | *GIT* | Greedy-Incremental-Tree Algorithm |
| *ALOC* | Ant-Lion Optimization for Clustering | *GA* | Genetic Algorithm |
| ALNs | Active/Alive nodes | *GIT* | Greedy Incremental Tree Algorithm |
| *AEC* | Average Energy Consumption | *HGWCSOA-OCHS* | Hybrid GOA and CSOA-based optimal CHS |
| *BOAACO* | Butter fly Optimization ACO | *HABC* | Hybrid Artificial Bee Colony |
| *CH/CHs* | Cluster-Head/Cluster-Heads | *HEED* | Hybrid Energy-Efficient Distributed Clustering |
| *CHS* | Cluster Head Selection | *ICOA* | Improved Coyote Optimization Algorithm |
| *CM/CMs* | Cluster Member/Cluster Members | *IJO-LF* | Improved Jaya Optimization Algorithm with Levy Flight |
| *DEEC* | Distributed Energy Efficient Clustering | ICOA | Improved Coyote Optimization Algorithm (ICOA) |
| *DTN* | Distance to neighbors | IJOA | Improved Jaya Optimization Algorithm |
| $D_{tr}$ | Distance between transmitter and receiver | *LND* | Last Node Dead |
| *DALO* | Discrete Ant Lion Optimization | *LEACH* | Low Energy Adaptive Clustering Hierarchical Routing |
| $E_{afs}$ | Energy consumed when free space model is used | *LEACH-C* | *LEACH-Centralized* |
| $E_{amp}$ | Energy consumed when multipath fading model is used | *ODTD-LI* | *Optimal data-transmission with data-location integration* |
| $E_{trans}$ | Energy lost to operate transmitter | *PEGASIS* | Power-Efficient Gathering in Sensor Information Systems |
| $E_{receiv}$ | Transmission amplifier's energy dissipation | *POA* | Pelican Optimization Algorithm |
| $E_{edtr}$ | Energy dissipated at transmitter receiver | *PEDAP* | Power Efficient Data Gathering and Aggregation Protocol |
| $E_{afs}$ | Energy consumed when free space model is used | *PSO* | Particle Swarm Optimization |
| *ER-SR* | Energy-efficient region source routing | PDR | Packet Delivery Ratio |
| *EEICS* | Energy-Efficient intra-cluster data communication scheme | *RE* | Residual Energy |
| *EEYSGPO* | Energy Efficient Yellow Saddle Goatfish POA | *SA* | Search Agent |
| *FCGWO* | Firefly Cyclic Grey Wolf Optimization | *SN/SNs* | Sensor Node(s) |
| *FFOA* | Firefly Optimization Algorithm | *TEEN* | Threshold sensitive energy efficient Sensor Network |
| *FF* | Fitness Function | *YSGA* | Yellow Saddle Goatfish Algorithm |

the complexity of network management. The load can be distributed uniformly among CHs with the aid of clustering, preventing certain nodes from using up their energy resources more quickly than others. This might prolong the life of the network. Clustering can offer some fault tolerance if a CH fails or a sensor node becomes inaccessible. To preserve network operation, other cluster nodes can choose a new CH or change their responsibilities. Nodes may periodically reorganize themselves during dynamic clustering [9–11] in response to variables like energy levels, network architecture changes, or data traffic patterns.

Energy-efficient (EE) routing [12, 13] algorithms are essential in WSNs to optimize energy consumption and extend the life of the network. Given these nodes' limited energy resources, it is critical to create routing protocols that minimize energy usage while preserving data delivery. Many EE routing techniques have been modeled for WSNs, each with its own strategy for preserving energy and prolonging network longevity. These protocols are following. *LEACH* [14, 50] is one of the most extensively used EE CHS and routing [15, 16] techniques for WSNs. It uses a clustering strategy, in which nodes are arranged into clusters, with one node serving as the cluster leader known as cluster head (CH). The *CH* combines and transmits data from cluster members (*CMs*) [17] to the BS. To balance the consumption of energy [18] across the network, nodes take turns acting as CHs. LEACH works by randomly selecting CHs in each round of iteration, allowing every sensor node an equal opportunity to become a CH, and balancing the energy consumption of the network. However, LEACH and its derivatives have some limitations, such as the random selection of CHs, which can lead to suboptimal CHS and uneven distribution of energy. LEACH does not consider the RE of the nodes, which can result in the premature death of CHs and

reduce the overall network lifetime. Additionally, there is a communication overhead due to periodic advertising of their candidacy. It has static sensor networks, inadequate handling of node failures, and no direct address to data aggregation. Meta-heuristics Optimization algorithm [48, 49] is essential for clustering and routing.

To collect data from the sensors deployed in the sensor field and sending it to the BS is the primary goal of the proposed work. To excel the performance of the proposed scheme, a POA-MHOA [19] scheme is proposed. This work introduces a new clustered based routing protocol incorporating a MHOA for increasing energy efficiency. Energy-efficient CH selection (CHS) and the generation of the best route for data transmission extend the lifespan of the network. In addition, by reducing the energy consumption of nodes when transmitting data packets, the BS receives more packets. The performance parameters [20] of the proposed model are AEC, ALNs, LND, and FND.

### 1.1 Contributions

Extensive research have been cried out by the researchers and they used two meta-heuristic optimizations, one for cluster formation and other for routing to send data to the BS which are not energy-efficient due to computation overheads and communication overheads. This work proposed a model which required single optimization algorithm (POA_proposed) for the CHS and routing. The key contributions of the work are as follows:

- The clustering and routing are NP-hard problems, and five FFs for CHS and three FFs for routing are used with the original POA to solve the problems.
- POA is modelled using the suitable fitness function to select the *cluster head*, and the selection of the optimal CH depends on the five *FFs*. The performance parameters of the FFs are node degree, distance to the neighbor's nodes, distance to the BS, node centricity, and *RE*.
- POA is also modelled using the suitable fitness function to select the optimal path between CHs and BS and the selection of the optimal path between CHs and BS depends on the three FFs. The performance parameters of the FFs are node degree, distance, and RE.
- Total fourteen different scenarios of sensor node deployments are considered for the simulation and proposed POA_Proposed is compared with LEACH, DEEC, and BOAACO.

The organization of the remaining work is as follows**:** Sect. 2 presents related research studies about routing algorithms and clustering. Section 3 presents the many preliminary studies considered for this research work. Section 4 presents a comprehensive description of the Energy-Efficient clustered routing strategy utilizing POA algorithms. Section 5 represent the performance evaluation and comparative analysis of the performance metrics. Section 6 represents the conclusion and future directions.

## 2 Related work

Pitchaimanickam et al. [21] introduced HFAPSO, a hybrid approach of FFOA and PSO to determine the optimal CH for the LEACH-C algorithm. Through integrating PSO, this method enhances the global search capability of Firefly. This methodology is evaluated based on throughput, RE, and active node count. Findings indicate prolonged network longevity, leading to reduced energy usage and augmented active/alive node count (ALNs).

A HGWCSOA-OCHS was proposed by Subramanian et al. [22]. By balancing the living and dead sensors in the network, the AEC of the network has been reduced and the lifespan of the network has been increased.

Sengathir et al. [23] introduced a hybrid modified ABC algorithm aimed at securing the clustering process against issues like slow convergence and premature convergence and avoiding trapping the system at the optimal point. The outcomes of HMABCFA demonstrate significant enhancements compared to standard benchmarks, with network lifespan, energy stability, and network latency showing improvements of 23.21%, 22.88%, and 19.84%, respectively.

Yogarajan et al. [24] proposed an Ant Lion algorithm tailored for CHS, aimed at prolonging network lifespan through optimization techniques. Furthermore, they utilized a discrete antlion optimization approach to identify the shortest path for mobile tanks. Leveraging Ant Lion's selective optimization methods, the mobile receiver efficiently gathers data by visiting chosen CH nodes. Simulation findings demonstrate that the suggested clustering technique enhances both network throughput and lifespan.

Zongshan Wang et al. [25] presented a clustering approach for CHS employing a modified ABC algorithm. This modified ABC algorithm, derived from the original ABC algorithm, addresses the clustering challenges in WSNs by considering factors such as CH energy, density, and location. Initially, all nodes in the network possess identical energy levels, but the modified *ABC* method maximizes the disparity among *CH* groups. The authors introduced an energy-efficient routing strategy based on enhanced ant colony optimization for paths between a central storage center and a BS. To improve residual energy of the sensors and overall network performance, they proposed a mechanism called polling control based on active/inactive nodes during cluster members and

cluster head communication in the continuous transmission phase. The efficacy of the proposed approach was evaluated across various conditions, with simulation results demonstrating its superiority over several comparable methods.

Chuan Xu et al. [26] proposed an *EE* area source routing model named *ER-SR*, drawing inspiration from approaches in the power distribution field. In this protocol, the source routing node strategically selects the optimal path for each shared node, allowing partial node involvement in routing while ensuring balanced energy consumption across sensor nodes. To mitigate data transmission energy consumption, the authors proposed an effective remote ACO algorithm to identify the most efficient transmission paths for each node. Simulation outcomes demonstrate that ER-SR outperforms in WSNs regarding network lifespan, packet delivery rate, and delivery delay, indicating a notable enhancement in energy efficiency.

Wei Xiang et al. [27] proposed a routing technique tailored for EE in Software-Defined WSNs (SDWSN). They proposed a scheme where control nodes are dynamically designated for various tasks. Addressing the NP-hard nature of selecting control nodes based on transmission distance and *RE* nodes, they put forth a strategy grounded in particle swarm optimization. Evaluation results demonstrate the superiority of their method over contemporary comparison techniques across various simulation scenarios.

Amit Sarkar et al. [28] proposed a Firefly contributions and Firefly Cyclic Randomization (FCR) within the context of the *CHS*. This approach employs Gamma, Normal, and Uniform distribution functions to derive randomly generated solutions. The Firefly Cyclic GWO (FCGWO) is then examined by altering random GWO vectors. Both *FCR* and *FCGWO* algorithms are utilized to identify the optimal CHs, aiming to minimize energy stability, delay, and inter-node distances. Evaluation is conducted for each distribution, focusing on energy efficiency, active node count, and network longevity.

Melaku Tamene Mekonnen et al. [29] put forward clustering methodologies employing *PSO* and simulated annealing algorithms to identify the most efficient set of *CHs* in the network and enhance its longevity by optimizing the distance between the sink node, RE and the individual nodes. Toolboxes for both SA and PSO have been created in C++ and seamlessly incorporated into the OMNeT++ simulation environment to implement these clustering strategies. The simulation assesses the performance of the algorithms in terms of energy conservation, load distribution, and overall network lifespan.

Amit Sarkar et al. [30] proposed a firefly-based cyclic randomized method for selecting the most efficient CH, aiming to minimize consumption OF energy, delay, and inter-node distances to increase the network's longevity. The efficiency of CHS using Firefly contributions was compared against ABC, FABC, GA, GSO, and Firefly algorithms which is much better than existing schemes considered for comparative analysis.

Mojgan Rayenizadeh et al. [31] introduced the Swarm Intelligence-based Hesitant Fuzzy (SIHF) system. Utilizing hesitant fuzzy values, the FFOA determines the optimal weight of nodes, while hesitant fuzzy logic aids in CHS. The SIHF system underwent comparison with SIF, LEACH, MODLEACH, and M-GEAR protocols across various scenarios. Results demonstrated that the SIHF protocol exhibited a lower energy depletion rate compared to other protocols, consequently extending the network's lifespan.

Amit Sarkar et al. [32] introduced FCGWO for selecting the optimal CH. This method prioritizes energy stabilization, minimization of delay, and reduction of node distance. The performance of the FCGWO-based CH selection (CHS) was compared against conventional techniques such as GWO-based CHS, FFOA), ABC, Half-Altered Artificial Bee Colony (FABC), Group Search Optimization (GSO), GA, and FFOA with Cyclic Randomization (FCR). The network's life of the proposed CHS scheme is better the existing scheme.

Priyanka et al. [33] introduced a method for enhancing network longevity by implementing circular sectorization in WSN along with dynamic CHS using the WOA. Their proposed model suggests that achieving relatively balanced energy consumption is feasible in circular WSNs through sectorization. The effectiveness of this approach is evaluated by comparing it with the PSO and LEACH protocols.

Deena Sivakumar et al. [34] utilized a Chaotic Gorilla Troops Optimization Algorithm (EACP-CGTOA) model to introduce an energy-conscious clustering technique characterized by fast convergence rate and sensitivity. The EACP-CGTOA model incorporates three input parameters—energy ratio (ER), distance to BS, and distance to neighbor nodes—to formulate a fitness function. Numerous simulations were conducted to validate the improved performance of the EACP-CGTOA approach, with the outcomes analyzed from various perspectives. Experimental results demonstrated that the EACP-CGTOA model outperforms existing methods significantly in terms of energy efficiency and network longevity.

Zongshan Wang et al. [35] introduced EECRAIFA, a cluster-based routing methodology. Initially, network nodes undergo clustering using a Self-Organizing Map neural network to determine optimal CHs. Subsequently, the FFOA is applied to optimize clustering by incorporating CH energy and average distance within clusters. FFOA defines the decision domain to further disperse CHs and create more suitable clusters. For inter-cluster routing, an advanced ACO is employed, incorporating the statistical variation coefficient in pheromone update, and optimizing paths based on distance and energy. Moreover, a polling control strategy

based on busy/idle nodes is integrated to enhance network throughput during intra-cluster communication. Simulation experiments demonstrate that EECRAIFA effectively manages network energy consumption across various application scenarios, leading to increased throughput and extended network lifespan.

Kathiroli Panimalar et al. [36] introduced a solution to address the energy efficiency challenge in WSNs through CHS, combining a Differential Evolution algorithm with a hybrid Sparrow Search method. Node longevity is enhanced by leveraging the effective Sparrow Search Algorithm and the dynamic capability of Differential Evolution. This hybrid approach optimizes performance based on throughput, RE, and the quantity of active and inactive nodes. The proposed modified Sparrow search technique high residual energy and throughput compared to previous algorithms, utilizing the differential evolution model to identify the most suitable CHs.

Within the framework of a WSN, Wang et al. [37] proposed a CHD-based routing protocol and scrutinized the benefits and limitations of the LEACH protocol. The researchers juxtaposed the results with those of the existing LEACH methodology and proposed a refined iteration named LEACH-Impt. Enhanced performance in CHS, specialized node handling, and inter-cluster routing challenges are observed with LEACH-Impt compared to traditional LEACH.

Sariga Arjunan et al. [38] enhanced network longevity through the proposal of an ACO-based routing scheme coupled with fuzzy logic-driven Cluster Head (CH) size adjustment. The endeavor encompasses cluster maintenance, inter-cluster routing, and CH selection. Employing a fuzzy-based technique, the network is partitioned into heterogeneous clusters, selecting CHs based on criteria such as node centrality, node degree, distance to BS, and distance to Cluster Members (CMs). Integration of ACO with suitable Forwarding Factors facilitates data transmission to the BS via inter-cluster routing. This method employs both proactive and reactive routing approaches, utilizing the shortest path to transmit data to the BS while maintaining load balance through threshold-based data transfer.

Madhusudhanan Baskaran et al. [39] introduced a variant of FFOA termed synchronous FFOA to enhance network longevity. Simulation results reveal a 9.63% average reduction in packet loss ratio compared to EEHC and LEACH, indicating an improvement in network lifespan.

Jin Duan et al. [40] introduced an enhanced version of the Population-based Optimization Algorithm (EPOA) for CHS (EPOA-CHS). The population initialization is carried out using the logistic-sine chaotic map, and a unique fitness function is adopted to facilitate suitable Cluster Head Selection (CHS), which is further hybridized with Levy flight for CH selection.

Adumbabu et al. [41] introduced an energy-efficient dynamic CHS and routing strategy employing ICOA and IJO-LF for CHS and routing, respectively. The performance of the proposed model illustrates that the ICOA and IJO-LF required less energy consumption as compared to PEGASIS and TEEN and the life of the proposed network is better than PEGASIS and TEEN.

Jing Li et al. [42] proposed a clustered routing technique named ODTD-LI. Comparative simulation results against EEICS, PEDAP, GIT, and LEACH illustrate superior performance in terms of network lifetime, average delay, and coverage rate with the suggested system.

Uma Meena et al. [43] introduced a CHS method named EEYSGPO. This method considers parameters such as RE, communication quality, load, delay, and distance in CHS decisions. Simulation results indicate that EEYSGPO outperforms SSMOECHS, LEACH-PRO, EECHIGWO, and FGWSTERP in terms of throughput, number of dead nodes, and energy consumption.

Hao Zhang et al. [44] proposed a reliable data transmission scheme named Geodesic Distance-based Routing (GDDR) have an irregular terrain structure to estimate the geodesic distance using sampling and flooding and the time complexity is reduced. A predetermined number of forwarding candidates are selected based on the remaining energy and geodesic distance. The geodesic distance is applied in the data transmission process. The forwarding candidates ensure the correct direction for data transfer, path connectivity and routing hole avoidance.

Bo Zheng et al. [45] presents an innovative protocol called the Airborne Greedy Geographic Routing (AGGR) for Flying Ad hoc Networks (FANETs). This protocol incorporates an adaptive Hello mechanism and a greedy forwarding technique. The AGGR can adjust to the rapid changes in networks, and it selects the next-hop node based on the time it takes to reach the communication range of the destination. This routing criteria leads to significant improvement in the performance of the AGGR.

Le Hoang Son et al. [46] proposed an energy-efficient routing scheme for clustering in wireless underground sensor networks (WUSN). The network life of the WUSN is 1.5 times longer than that of the Fuzzy C-Mean (FCM) and the number of alive nodes is 1.6 times higher than that of the FCM.

Yinhao Liu et al. [47] proposed a node classification and multi-level data transmission (NCMLT) scheme that categorized nodes into two levels (S1 and S2). S1 and S2 level nodes transmit data in a chain structure and clusters respectively. The S1 nodes can connect to the base station directly. The S2 nodes are associated with the LEACH algorithm to select the CHs. To transmit data to the base

station, multilevel multi-hop routes are established. The network life of the proposed scheme is increased by 47.43%.

Akhilesh Panchal et al. [51] proposed a clustered-routing scheme named ECHR-FCM in WSNs, which is energy-efficient. The hierarchical clustering using the Fuzzy C-means (FCH) algorithm is modeled to select the CHs and Grid Heads (GHs). The routing scheme is modeled using a suitable fitness function to increase the network's life. Akhilesh Panchal et al. [52] proposed an energy-efficient CHS and grid head selection model named EOCGS in WSNs to extend the network life and coverage. The data packet collected by the CHs from the CMs is sent to the BS via GHs or directly to the BS to reduce the communication overheads and increase the network's lifetime. Singh, Rajat Kumar et al. [53] proposed a modified Residual Energy-based CHS in WSNs named MARCH. The proposed model is more energy-efficient than RCH-LEACH. The CHS is based on the energy level of the alive nodes in the network.

Guo et al. [54] proposed a reinforcement learning based clustering and routing scheme named the intelligent clustering routing approach (ICRA) for unmanned aerial vehicles. The scheme reduced the end-to-end delay and enhanced the packet delivery rate, topology stability, cluster efficiency, and quality of service.

To enhance the security of power in IoT, Zheng et al. [55] proposed an Adaptive Route Mutation based on Network State Awareness (SARM). The SARM scheme enhances the DDoS and defense against sniffing attacks, which are approximately 35% and 30%, respectively. We proofread the complete manuscript and corrected grammatical mistake in the manuscript.

Conclusion drawn from the literature survey: A comprehensive review of CHS models and routing protocols is carried out, that presenting a detailed summary of their features and functionalities in Table 2. While certain protocols prioritize energy-efficient routing strategies, others emphasize load balancing and enhancing network stability. Through this analysis, the various gaps within the existing literature are identified. These gaps may include areas such as comprehensive integration of EE and load balancing, adaptation to dynamic network conditions, and effective handling of heterogeneous node environments. The proposed protocol seeks to bridge these gaps and modelled a CHS and routing methodologies within WSNs.

## 3 Preliminaries

This section present the network architecture, energy model, and POA.

### 3.1 Network architecture

A high-level diagram of the proposed WSN model is shown in Fig. 1. The following things are considered when developing the network model. The sensor nodes are deployed randomly in sensor field. The distance between CM to CM, and CM to CH is estimated. Node degree, node centrality, and residual energy of the sensor are also estimated. The selection of the CH and optimal route from CH to BS is estimated using these fitness function parameters. The POA is modelled with suitable fitness functions. The simulation is carried out using POA scheme proposed. The simulation setup is based on the following conditions:

- •All SNs are homogenous in their property.
- •All SNs are deployed randomly in the sensor field of area 200*200 square meters except sink node.
- Two scenarios have been considered to deploy the sink's node as shown in the Fig. 2(a) and (b). In the first scenarios, the BS is deployed at the center (100,100) of the sensor filed ($200 \times 200$ square meters), and in the second scenarios, the BS is deployed outside of the sensor deployment field at (100,250).
- All sensors have the same transmission radius.
- The sensors' position is static after the deployment.
- Every sensor is aware of its neighbor sensors.
- Each sensor has 0.5 Jules initial energy.
- The distance between the SNs is estimated using the Euclidian distance measurement method.
- The sensors deployed in the environment send the node RE and position to the BS. The BS executes the CHs selection algorithm to select the CHs, and CMs are assigned to the suitable CH. Afterwards, the POA-based routing is executed to choose the optimal path from CHs to BS.

The popular clustering and routing taxonomy is shown in Fig. 3. The choice of algorithm is determined by the WSN's unique requirements and characteristics, including the application, energy limitations, and network size. POA is chosen and modified using the suitable fitness function to achieve the proposed objectives (Fig. 4).

### 3.2 Energy model

The sensor nodes have limited energy, but they can aggregate data to make efficient use of that energy. The amount of energy needed to transport data depends on the distance between nodes and the number of bits being sent. The amount of energy needed for receiving is also influenced by how many bits are being received. The energy required for the direct transmission between the transmitter and receiver is as follows:

**Table 2** Summary of the related works

| Ref no. | Algorithm(s) | Objective(s) | Performance parameter(s) | Achievement(s) | Limitation(s) |
|---|---|---|---|---|---|
| [21] | FFOA with PSO and LEACH-C | To find the optimal CH using energy-efficient MHOA | ANLs, RE, and throughput | Improvement in network life, ANLs, and reduced energy consumption | LEACH-C is not a MHOA and compare with MHOAs |
| [22] | HGWCSOA-OCHS | to enhance network life and reduce energy consumption | RE, network life, ALNs, dead nodes | The mean RE performs better than the other clustering methods | Under different sink placements, the suggested HGWCSOA-OCHS-mean residual energy remains unaltered. The result is not verified |
| [23] | HABC and FFOA | An energy-efficient approach to find the CH using HABC and FFOA and improve the network life | ALNs, RE, FND, HND, and LND, network life, throughput | Compared to the benchmarked techniques, the HMABCFA improves energy stability by 19.84%, average network life by 23.21%, and reduces network latency by 22.88% | Tables 2 and 3 listed secure nodes but no description about the security is mentioned the paper |
| [24] | ALOA and DALOA | Optimized CHS and routing | Lifetime of the nodes | The network life is improved by 13% compared to existing | Authors stated that the model is suitable for large scale network, but simulation is carried out with maximum 100 nodes |
| [25] | Improved ABC | Design an optimal CHS using ABC and energy EE routing | Network lifetime, average energy consumption, throughput, and network stabilization time | Energy efficient CHS is implemented | Limited scalability, and static environment assumption |
| [26] | ER-SR | To select EE nodes using ER-SR algorithm for | Energy efficiency is higher in ER-SR | According to simulation studies, ER-SR has greater energy efficiency | In comparison to current routing systems, performance gains on the network lifetime, PDR, and delivery delay are moderate |
| [27] | SDWSN | Routing | ALNs, FND | Energy efficient | Area is 100×100 m only and SNs are 300 to 600 which is not cost effective |
| [28] | FCR and FCGWO | CHS | ALNs, AE, and network lifetime | The FCGWO and FCR schemes had attained 59 and 58 ALNs out of 100 SNs respectively | Area of the sensor field is not mentioned which is needed for better analysis |
| [29] | Simulated Annealing and PSO | CHS | ALNs, Energy consumption | Improve Network life and load balancing | All node dead after 800 rounds of operations which is very less |
| [30] | FCR | CHS and Routing | ALNs, Energy consumption, CHs, and network lifetime | The proposed FCR scheme has outperformed the conventional CHS | It is not clear from the simulation result that when last node is dead |
| [31] | Hesitant fuzzy and FFOA | CHS | FND, LND, HND, and packets delivered to the BS | Network life is improved using clustering method | Fitness function is based on only FND, HND, and LND |
| [32] | FFOA and GOA | Optimal CHS | Normalized network energy, alive node | The FCGWO method has better energy stabilization and run till 2000 rounds | The no of ANs in the proposed scheme is like other existing methods chosen |

**Table 2** (continued)

| Ref no. | Algorithm(s) | Objective(s) | Performance parameter(s) | Achievement(s) | Limitation(s) |
|---|---|---|---|---|---|
| [33] | WOA | CHS | RE and delay | Energy efficient, dynamic CHS model is proposed | An initial energy of WOA, PSO and LEACH are 50jules, 75 Jules and 100 Joules which is too high energy for WSNs |
| [34] | Chaotic gorilla troops optimization algorithm | CHS | ALNs, Average throughput, residual energy, and communication overheads | Energy efficient CHS | The senor field is 400*400 square meters the total SNs are 1000 and node density is high and not cost effective |
| [35] | FFOA AND ACO | Cluster based Routing to improve the network energy consumption, lifetime, and throughput | RE, Network stability period, and throughput | Energy efficient CHS and routing | An initial energy of the nodes was considered 50 Jules which is high in case of WSNs |
| [36] | Sparrow Search Algorithm | CHS | ALNs, RE, dead nodes, delay, and mean throughput | CH stability is high | The BS positions are (100,200) and (200,100) which are very similar |
| [37] | LEACH_IMP | Comparative analysis of LEACH and LEACH-IMP in CHS and Routing | Data transmission rate, energy consumption, and dead nodes | LEACH-IMP is found much better than LEACH | Simulated on static network |
| [38] | Fuzzy and ACO | Fuzzy based CHS and ACO based routing | ALNs, FND, LND, AEC, Total packets received by BS | Improved network lifetime | SNs and BS are static |
| [39] | FFOA | CHS | Residual energy, alive node | Improved network lifetime | Only Residual energy, alive node are the performance parameters analyzed |
| [40] | Enhanced POA | CHS | ALNs, Packet delivered to BS, and dead node | Improve network life | The fitness function includes only distance and energy parameters. Other parameters like node degree and centrality needed to be included to make the better FF |
| [41] | ICOA and IJOA with Levy Flight (IJO-LF) | Energy efficient CHS and routing | ALNs, AEC, and network life. PDR, throughput, and routing overhead | The proposed method performs better than TEEN and PEGASIS | Then BS is dynamic but movement is not defined |
| [42] | ODTD-LI | Energy efficient CHS and routing | Average delay, network lifetime, coverage rate, and network load balance | The energy consumption is low in the ODTD-LI algorithm | It is stated that the network contains no attack nodes but the mechanism used for protection is not mention in the proposed work |
| [43] | *YSGA and POA* | CHS, and energy efficient routing using POA | Network stability, dead node, throughput, and AEC | Energy efficient CHS and Optimal route is selection is performed using POA | The *FF* is designed based on the energy and distance between BS and CH. Other dependent factors like Node centrality and density are not considered for *FF* |
| **[C]** | *POA_Proposed* | Energy-efficient clustering and Routing | AEC, LND, ALN, FND are the performance parameters | The POA_Proposed performs better than LEACH, DEEC, AND BOAACO | Grid heads can be added for further improvement in energy consumption in the clustering |

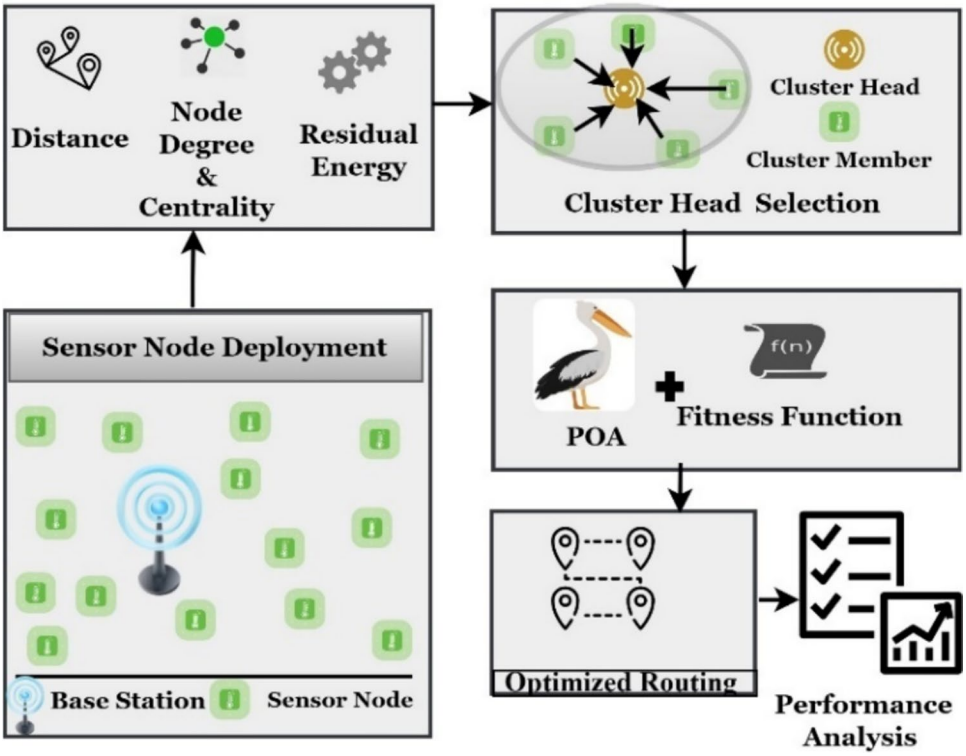



**Fig. 1** Proposed model

$$E_{trans}(l,d) = \begin{cases} E_{edtr} * w + E_{afs} * w * d_{tr}^2, if d_{tr} < d_{th} \\ E_{edtr} * w + E_{amp} * w * d_{tr}^4, if d_{tr} > d_{th} \end{cases}, \quad (1)$$

$$where\ d_{th} = \sqrt{\frac{E_{afs}}{E_{amp}}}$$

$$E_{receiv}(w,d) = E_{edtr} * w \quad (2)$$

The energy lost to operate the transmitter is $E_{trans}$, the transmission amplifier's energy dissipation is $E_{receiv}$, w is the message's length in bits, and $d_{tr}$ is a distance between the transmitter and receiver. $E_{edtr}$ is an energy dissipated at transmitter receiver. $E_{edta}$ is an energy dissipated in amplification. Threshold distance is $d_{th}$.

In the multi-hope scenario, the energy required for the transfer of data from the source to the destination node will be reduced by half compared to direct transmission.

$$E_{trans} = E_{edtr} * w + E_{edta} * w * \left(\frac{d_{tr}}{2}\right)^2 \quad (3)$$

### 3.3 POA

POA [13] is a nature-inspired stochastic optimization algorithms developed by Pavel Trojovský in 2021. In POA, pelicans are the search agents that hunt prey. The Pelican defines a set of SNs from which to choose CHs from a network's sensors. The CHs are used to send data to the BS.

## 4 Proposed scheme

The proposed work aims to transfer data from CHs to BS in such a way that the energy consumed by the nodes are least. The proposed work can be divided into two phases. The phase one is cluster formation, algorithm-1 represents the CH selection scheme using POA, and phase-2 is routing optimization to select the optimal route from sources (CHs) to the BS.

### 4.1 Selection of CHs using POA

1.**Input:** **Parameter initialization:** Pelican population=$N_p$
2. Maximum no of Iteration=$T_{max}$
3. Estimate the fitness value ($ff_{CH}$) using Equ. (8)
4. for $i = 1: T_{max}$
4. Random initialization of the prey using equ. (11)
5. for $j = 1: N_p$
6. **Exploration phase:** Movement of *SA* towards the prey
7. For $p = 1:m$
8. Estimate the new status of the *CH* using Equation (14)
9. End for
10. Update the $i^{th}$ pelican position using Equ. (15)
11. **Exploitation phase:** To attack on prey
12. For $p = 1:m$
14. Estimate the new position of the *CH* using Equation (16)
15. End.
16. Update the $i^{th}$ *SA* using Equ. (17)
17. End.
18. Update the solution for the best candidate
19. End.
20. **Output:** *CH* is the best candidate solution attained by POA-Proposed

### 4.2 Fitness function of POA

A class of optimization methods called MHOA is intended to identify the best solutions to complex problems. In the optimization process, the objective function also referred to as the fitness or cost function acts as a criterion to assess a solution's quality. Fitness evaluation, search guidance, convergence rate, adaption and learning, and problem-specific information are the main tasks of the *FF* in MHOA.

#### 4.2.1 Distance among the CHs and BS

It specifies a distance between CH and BS. The length of the transmission channel determines how much energy the node uses. For data transmission, the node that is closer to the BS is preferred because the BS requires more energy if it is located far away from the CH. So, more energy consumption could be the cause of the abrupt decline in CH. The *FF* of the distance between CHs and BS is expressed in Eq. (4) as follows:

$$ff_1 = \sum_{i=1}^{N_{CH_j}} \left(dist(CH_i, BS)\right) \tag{4}$$

#### 4.2.2 Distance among CMs and CH

When the distance between the CH and the CMs increases, the energy consumed by the nodes also increases. This is because the nodes may need to use more energy to transmit data over longer distances to reach the CH. Hence, The CM located at a shorter distance will be preferred as a CH for data transfer to the BS. It uses less energy. The *FF* for the distance between CH and CM is expressed in Eq. (5) as follows:

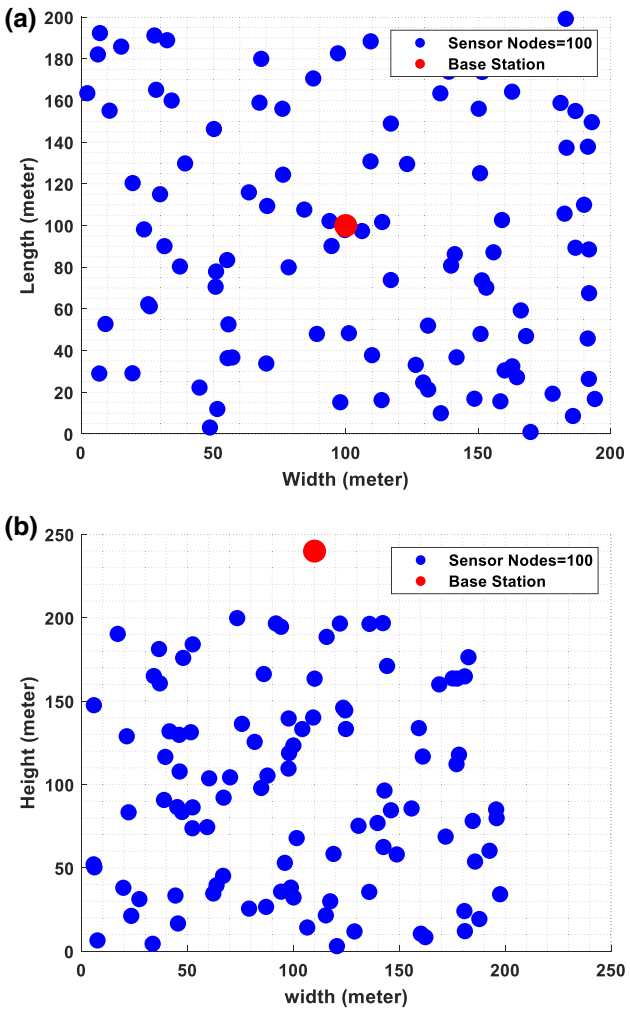



**Fig. 2** SNs and BS **a** (100,100) deployment **b** (100,250) deployment

$$ff_2 = \sum_{j=1}^{M_{CH}} \left( \sum_{i=1}^{N_{CH_j}} \left( dist(CM_i, CH_i) / N_{CH_j} \right) \right) \tag{5}$$

#### 4.2.3 Residual energy of the CH

To gather information from the CMs and forward it to the BS, CH performs several operations. The node with more RE is preferred to be a *CH* since the *CH* needs higher levels of energy to complete the task. The *FF* of the RE is presented in Eq. (6) as follows:

$$ff_3 = \sum_{j=1}^{N_{CH}} \frac{1}{E_{CH_j}} \tag{6}$$

where $N_{CH}$ is the size of the *CH* in the network and $E_{CH_j}$ is the RE of the $j^{th}$ CH.

#### 4.2.4 CH degree (DCH)

The degree of CH specifies how many sensor nodes are a part of the specific CH. The CHs with fewer sensors are chosen because the CHs with more CMs consume energy more quickly. The *FF* of the CH degree is presented in Eq. (7) as follows:

$$ff_4 = \sum_{j=1}^{N_{CM}} CH_j \tag{7}$$

where $N_{CM}$ is the total number of CM associated with $CH_j$.

#### 4.2.5 Node centrality

Node centrality is a measure of how centrally positioned a node is about its neighbors. The *FF* of the node centrality is presented in Eq. (8) as follows:

$$ff_5 = \sum_{i=1}^{N_{CM}} \frac{\sqrt{\sum_{J \in N_{CH}} dist^2(i,j)/N_{CM}}}{network_{dim}} \tag{8}$$

where $N_{CM}$ is total number of *CM* of the $CH_j$. These five objective functions have a weight value assigned to it. The many objectives in this instance are reduced to a single *FF*.

To prevent node failure when acting as a *CH*, priority is given to ω1 to elevate the importance of *RE*. Subsequently, ω2 and ω3 are prioritized next to identify the *CH* from the BS with the shortest distance, thereby reducing energy consumption. The fourth priority, assigned to node degree (ω4), aids in selecting the *CH* with the least node connections. Lastly, node centrality (ω5) is considered the least priority, serving to enhance the connection between *CMs* and the *CH*. The transformed *FF* from five objective functions is shown in the Eq. (9).

$$ff_{CH} = \sum_{i=1}^{5} \omega_i * ff_i \tag{9}$$

where weighted values ($\omega_i$) [6] $\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 = 1$.

### 4.3 Cluster formation

The CHs are chosen by the *POA*. The potential function mentioned in Eq. (10) is the basis of becoming the *CH* of any *CMs*. The potential of the any sensor is more if the sensor node has higher *RE* and are closer to the *CMs*. If the two *CHs* are at the same distance from the *SN*, then the *SN* becomes the *CM* of the *CH* which has higher *RE*.

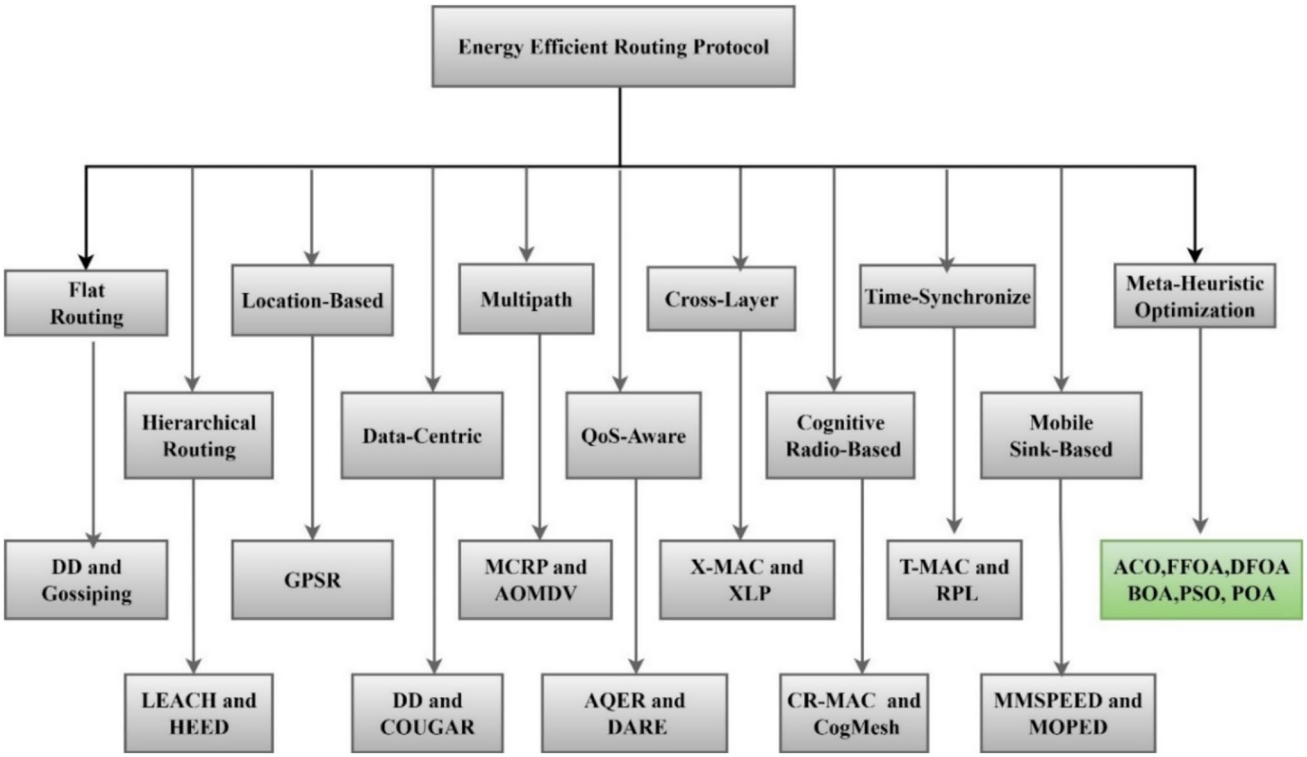



**Fig. 3** Taxonomy of energy-efficient CHS and routing protocols in WSNs

$$P_{SN} = \frac{\kappa * E_{CH_j}}{dist(SN_i, CH_j)} \tag{10}$$

where $\kappa$ is the proportionality constant, $E_{CH_j}$ is the *RE* of the $CH_j$, and $dist(SN_i, CH_j)$ is the distance between the $SN_i$ and $CH_j$.

### 4.4 Routing using POA

Communicating directly between *BS* and *CHs* is not feasible due to the limited communication range and battery of the SNs. That is why multi-hop routing is the requirement of the proposed work. Figure 5 repenting the general routing scenario. POA is chosen for optimized routing. In the routing phase, dimension(m) of each pelican is equal to the size of *CHs*. Each Pelican is a search agent (*SA*) and *SA* is candidate solution of the problem. The SAs are randomly deployed in the search space using Eq. (11). In the search field, the prey's position is generated randomly.

$$P_i{:}P_{i,j} = lowerBound_j - rand * \left(upperbound_j - lowerbound_j\right), i = 1toN, j = 1tom \tag{11}$$

where $P_i$ is the location of $i^{th}$ pelican in search area. The value of the *j*th decision variable is $P_{i,\,j.}$ Where *N* is the search agent. The lower and upper bounds of the *j*th decision variables are $lowerbound_j$ and $upperbound_j$ respectively. The population matrix of a pelican is mathematically represented by matrix P as fallows.

$$P = \begin{bmatrix} P_1 \\ \vdots \\ P_i \\ \vdots \\ P_N \end{bmatrix} = \begin{bmatrix} P_{1,1} & \cdots & P_{1,j} & \cdots & P_{1,m} \\ \vdots & & & & \\ P_{i,1} & & P_{i,j} & & P_{i,m} \\ \vdots & & & & \\ P_{N,1} & & P_{n,j} & & P_{N,m} \end{bmatrix}_{Nxm} \tag{12}$$

The objective function (*ff*) vector, as illustrated in Eq. (13), is used to derive the values obtained for the *FFs*. The goal of routing is to find the optimal route from *CHs* to the *BS*. To do this, *RE*, distance, and node degree are combined to create a fitness function ($ff_R$). The $ff_R$ of the distance, RE, and node degree are expressed in the Eqs. (5), (6), and (7) respectively. The total fitness value for the routing optimization is as follows:

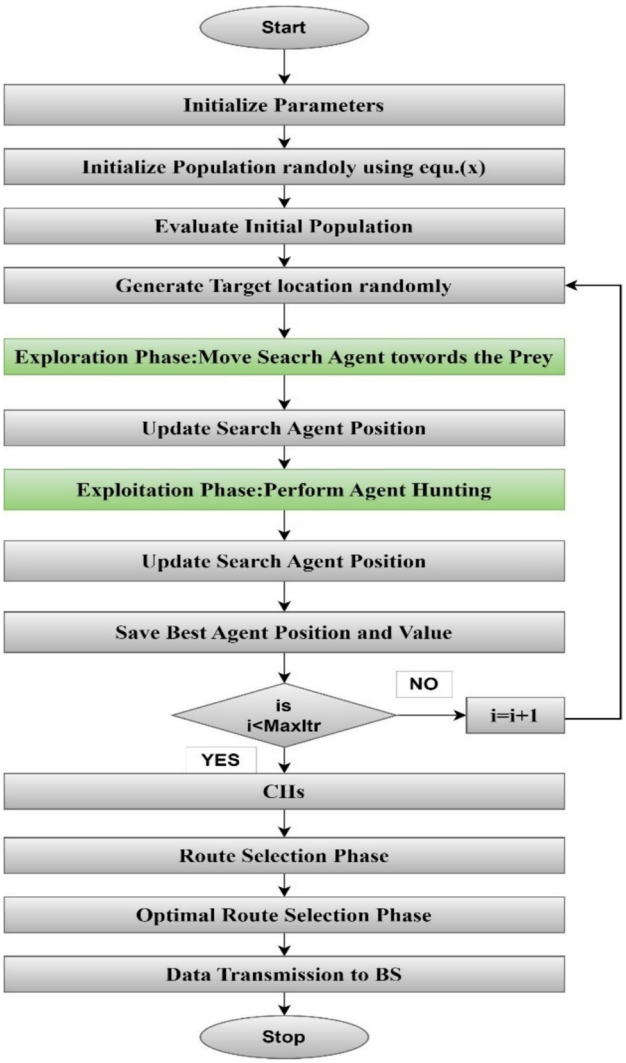



**Fig. 4** Flowchart of the proposed work for the clustering and routing

$$ff_R = \sum_{i=1}^{3} \varphi_i * ff_i \tag{13}$$

where the weighted values $(\varphi_i)$ [5] $\varphi_1 + \varphi_2 + \varphi_3 = 1$ *and* the weighted values $(\varphi_i)$ are 0.5, 0.3 and 0.2 for the $\varphi_1$, $\varphi_2$ and $\varphi_3$ respectively.

To update the candidate position for optimal solutions, the proposed POA mimics the behavior of pelican. The two main phases of the simulation are exploration and exploitation. In the exploration phase, *SA* moves toward the prey shown in Eq. (14).

$$P_i\left(x_{i,j}\right)^{exploration} = \begin{cases} x_{i,j} + rand.\left(position_{prey_j} - I.x_{i,j}\right), & if\ ff_{prey} < ff_{sa} \\ x_{i,j} + rand.\left(x_{i,j} - position_{prey_j}\right), & otherwise \end{cases} \tag{14}$$

where $P_i\left(x_{i,j}\right)$ is the current position of the *i*th SA in the exploration phase. *Random number(I)* has the value 1 or 2 which is chosen randomly. $ff_{prey}$ is the current *FF* value of the prey and $ff_{sa}$ is the current *FF* value of the *SA*. The *SA* position is updated if the current *FF* value of the *SA* ($ff_p$) is better than the *SA* as shown in Eq. (15).

$$x_i = \begin{cases} x_i^{SA}, & ff_{new} < ff_{old} \\ x_i, & otherwise \end{cases} \tag{15}$$

The new position of the *SA* in the exploitation phase, is shown in Eq. (16).

$$P_i\left(x_{i,j}\right)^{exploitation} = x_{i,j} + 0.2\left(1 - \frac{t}{itr_{max}}\right) * \left(1 - \frac{t}{itr_{max}}\right) * x_{i,j} \tag{16}$$

where $x_{i,j}$ is the older location of the SA, t is the current iteration, $itr_{max}$ is the total number of iterations.

Finally, Eq. (17) models the new pelican position, which is subject to acceptance or rejection based on effective updating.

$$X_i = \begin{cases} P_i\left(x_{i,j}\right)^{exploitation}, & P_i\left(x_{i,j}\right)^{exploitation} < ff_i \\ X_i, & otherwise \end{cases} \tag{17}$$

Our protocol employs a dynamic clustering mechanism that periodically selects *CHs* based on them *RE* and proximity to sensor nodes.

## 5 Performance analysis

### 5.1 Simulation setup

An Intel(R) Core (TM) i7-1165G7 processor with four cores and eight logical processors (s) operating at 2.80 GHz, 2803 MHz is used for the simulation experiment, along with the MATLAB2021a simulation environment. The simulation parameters are listed in the Table 3.

### 5.2 Performance parameters and results

We evaluate the protocol using various performance metrics, including network lifetime, energy consumption, ALNs, and total packet received by BS. Two scenarios have been well-thought-out in the simulation. In the scenario one, the BS is positioned inside the environment field at the center (100,100) and in scenario 2, the *BS* is positioned outside of the environment field at (100,250). A total of three cases (100, 200, and 300 nodes) have been chosen for the SN deployments. The all-simulation results are presented in Table 4.

Alive node (ALN): ALN usually refer to sensor nodes that are now in use and functioning as part of the network. When the network contains many active nodes, performance is enhanced.

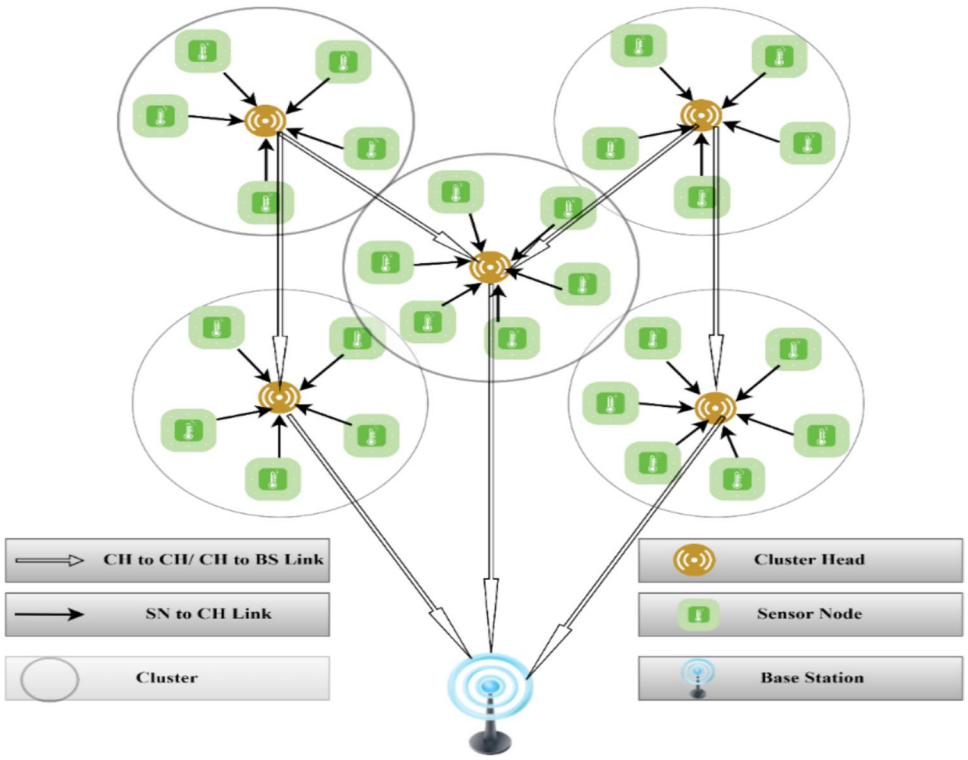



**Fig. 5** Proposed clustering model for routing in WSNs

Average energy consumption (AEC): The total energy consumed by the node throughout each iteration.

Total packets received by BS: The total number of packets that *BS* receives depends on the RE and distance between *CHs* and *BS*.

First node dead (FND): Nodes that have failed or run out of energy are referred to as dead nodes. When the first sensor node in a network break or stops functioning, it is referred to as the FND.

Last node dead (LND): When the final sensor node in a network break or stops functioning, it is referred to as LND.

#### 5.2.1 Case study 1: ALNs versus number of rounds when BS positioned at (100,100) and SN = 100

In the proposed work, the simulation is carried out to check the number of active/alive nodes. The first node is dead after 800 rounds of operations, and all nodes are dead after 4400 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the first node is dead after 800 rounds, but the last node is dead after 2800, 3200, and 3600 rounds of operations, as shown in Fig. 6 and in Table 4.

#### 5.2.2 Case study 2: ALNs versus number of rounds when BS positioned at (100,100) and SN = 200

In the proposed scheme, the first node dies after 3000 rounds of operations, and all nodes are dead after 6600 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the first node is dead after 800, 1200, and 1200 rounds, respectively, but the last node is dead after 2800, 3200, and 3600 rounds of operations, respectively, as shown in Fig. 7 and in Table 4.

#### 5.2.3 Case study 3: ALNs versus number of rounds when BS positioned at (100,100) and SN = 300

In the proposed scheme, the first node dies after 5000 rounds of operations, and all nodes are dead after 11,000 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the first node is dead after 2000 rounds, respectively, but the last node is dead after 7000,8000, and 9000 rounds of operations, respectively, as shown in Fig. 8 and in Table 4.

#### 5.2.4 Case study 4: ALNs versus no of rounds when BS positioned at (100,250) and SN = 100

In the proposed scheme, the first node dies after 1300 rounds of operations, and all nodes are dead after 3000 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the first node is dead after 400,600, and 800 respectively, but the last node is dead after 1800,2100, and 2400 rounds of operations, respectively, as shown in Fig. 9 and in Table 4.

#### 5.2.5 Case study 5: ALNs versus no of rounds when BS positioned at (100,250) and SN = 200

In the proposed scheme, the first node dies after 2000 rounds of operations, and all nodes are dead after 4400 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the first node is dead after 800, 800, and 1200 respectively, but the last node is dead after 2800, 3200, and 3600 rounds of operations, respectively, as shown in Fig. 10 and in Table 4.

#### 5.2.6 Case study 6: ALNs versus no of rounds when BS positioned at (100,250) and SN = 300

In the proposed scheme, the first node dies after 2500 rounds of operations, and all nodes are dead after 5500 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the first node is dead after 1000, 1000, and 1500 respectively, but the last node is dead after 3400,4000, and 4500 rounds of operations, respectively, as shown in Fig. 11 and in Table 4. The LEACH, DEEC, and BOAACO show a gradual decrease in ANL as the number of rounds increases, indicating a decline in network lifetime with time. The result analysis shows that the POA_Proposed model offers superior performance in terms of network lifetime extension compared to LEACH, DEEC, and BOAACO in the given simulation scenario 1.

#### 5.2.7 Case study 7: AEC versus no of rounds when BS positioned at (100,100) and SN = 100

It can be observed that the AEC of the proposed scheme was less compared to LEACH, DEEC, and BOAACO. The residual energy of the proposed model is much better than the other three existing schemes. The AEC is zero after 4400 rounds of operations, as shown in Fig. 12 and Table 4. In

**Table 3** Simulation parameters

| Parameters | Value |
|---|---|
| Sensor area | 200 × 200 Sq. meter |
| No of sensor nodes deployed | 100, 200, 300 |
| BS position | (100,100) and (100,250) |
| Initial energy of the SNs | 0.50 Jules |
| Energy dissipated at receiver and transmitter | 50 nj/bit |
| Packet size | 4000 bits |
| Message size (w) | 500 bits |
| Sensor's communication range | 100 m |
| $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5$ [6] | 0.40,0.20, 0.15, 0.13, 0.12 |
| $\varphi_1, \varphi_2, \varphi_3$ [5] | 0.50, 0.30, 0.20 |
| Base Station coordinates | $BS_1$(100,100) and $BS_2$(100,250) |

**Table 4** ANLs of LEACH, DEEC, BOAACO, and POA_Proposed

| No of rounds | LEACH | DEEC | BOAACO | POA_Proposed |
|---|---|---|---|---|
| *ALNs versus number of rounds when BS positioned at (100,100) and sensor node = 100* | | | | |
| 400 | 100 | 100 | 100 | 100 |
| 800 | 100 | 100 | 100 | 100 |
| 1200 | 70 | 80 | 90 | 100 |
| 1600 | 50 | 70 | 85 | 100 |
| 2000 | 20 | 50 | 70 | 100 |
| 2400 | 10 | 30 | 55 | 90 |
| 2800 | 0 | 10 | 45 | 50 |
| 3200 | 0 | 0 | 30 | 45 |
| 3600 | 0 | 0 | 0 | 30 |
| 4000 | 0 | 0 | 0 | 10 |
| 4400 | 0 | 0 | 0 | 0 |
| *ALNs versus number of rounds when BS positioned at (100,100) and sensor node = 200* | | | | |
| 600 | 200 | 200 | 200 | 200 |
| 1200 | 190 | 200 | 200 | 200 |
| 1800 | 140 | 160 | 180 | 200 |
| 2400 | 110 | 130 | 170 | 200 |
| 3000 | 80 | 100 | 150 | 200 |
| 3600 | 50 | 70 | 130 | 190 |
| 4200 | 10 | 40 | 90 | 150 |
| 4800 | 0 | 30 | 60 | 140 |
| 5400 | 0 | 0 | 10 | 110 |
| 6000 | 0 | 0 | 0 | 50 |
| 6600 | 0 | 0 | 0 | 0 |
| *ALNs versus number of rounds when BS positioned at (100,100) and sensor node = 300* | | | | |
| 1000 | 300 | 300 | 300 | 300 |
| 2000 | 300 | 300 | 300 | 300 |
| 3000 | 240 | 260 | 280 | 300 |
| 4000 | 200 | 260 | 270 | 300 |
| 5000 | 180 | 210 | 250 | 300 |
| 6000 | 10 | 90 | 230 | 290 |
| 7000 | 0 | 30 | 190 | 250 |
| 8000 | 0 | 20 | 160 | 240 |
| 9000 | 0 | 0 | 0 | 185 |
| 10,000 | 0 | 0 | 0 | 100 |
| 11,000 | 0 | 0 | 0 | 0 |
| *ALNs versus number of rounds when BS positioned at (100,250) and sensor node = 100* | | | | |
| 200 | 100 | 100 | 100 | 100 |
| 400 | 100, | 100 | 100 | 100 |
| 600 | 90 | 100 | 100 | 100 |
| 800 | 70 | 90 | 100 | 100 |
| 1200 | 50 | 80 | 90 | 100 |
| 1500 | 30 | 60 | 50 | 90 |
| 1800 | 0 | 30 | 40 | 70 |
| 2100 | 0 | 0 | 30 | 50 |
| 2400 | 0 | 0 | 0 | 35 |

**Table 4** (continued)

| No of rounds | LEACH | DEEC | BOAACO | POA_Proposed |
|---|---|---|---|---|
| 2700 | 0 | 0 | 0 | 20 |
| 3000 | 0 | 0 | 0 | 0 |
| *ALNs versus number of rounds when BS positioned at (100,250) and sensor node = 200* | | | | |
| 400 | 200 | 200 | 200 | 200 |
| 800 | 200 | 200 | 200 | 200 |
| 1200 | 140 | 160 | 200 | 200 |
| 1600 | 110 | 120 | 180 | 200 |
| 2000 | 80 | 90 | 150 | 200 |
| 2400 | 30 | 70 | 130 | 190 |
| 2800 | 0 | 40 | 90 | 170 |
| 3200 | 0 | 0 | 70 | 150 |
| 3600 | 0 | 0 | 0 | 125 |
| 4000 | 0 | 0 | 0 | 100 |
| 4400 | 0 | 0 | 0 | 0 |
| *ALNs versus number of rounds when BS positioned at (100,250) and sensor node = 300* | | | | |
| 500 | 300 | 300 | 300 | 300 |
| 1000 | 300 | 300 | 300 | 300 |
| 1500 | 240 | 260 | 300 | 300 |
| 2000 | 200 | 230 | 270 | 300 |
| 2500 | 80 | 100 | 250 | 300 |
| 3000 | 20 | 70 | 240 | 290 |
| 3500 | 0 | 30 | 190 | 260 |
| 4000 | 0 | 0 | 180 | 250 |
| 4500 | 0 | 0 | 0 | 165 |
| 5000 | 0 | 0 | 0 | 100 |
| 5500 | 0 | 0 | 0 | 0 |

the case of LEACH, DEEC, and BOAACO, the AEC is zero after 2000, 2800, and 3600 rounds of operations.

#### 5.2.8 Case study 8: AEC versus no of rounds when BS positioned at (100,100) and SN = 200

The simulation results, as shown in Fig. 13 and Table 5, indicate that the AEC of the proposed scheme was less compared to LEACH, DEEC, and BOAACO and performed better when the number of sensor nodes increased to 200. The residual energy of the proposed model is much better than the other three existing schemes. The AEC is zero after 4400 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the AEC is zero after 3000, 4200, and 5400 rounds of operations.

#### 5.2.9 Case study 9: AEC versus no of rounds when BS positioned at (100,100) and SN = 300

The simulation results, as shown in Fig. 14 and Table 5, indicate that the AEC of the proposed scheme was less compared to LEACH, DEEC, and BOAACO and performed better than case studies 7, 8 when the number of sensor nodes increases to 300. The residual energy of the proposed model is much better than the other three existing schemes. The AEC is zero after 11,000 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the AEC is zero after 2000, 2800, and 3600 rounds of operations.

#### 5.2.10 Case study 10: AEC versus no of rounds when BS positioned at (100,250) and SN = 100

It can be observed from Fig. 15 and Table 5 that the AEC is more when BS is positioned outside of the environment field, but the AEC of the proposed scheme was less compared to LEACH, DEEC, and BOAACO. The AEC is zero after 3000 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the AEC is zero after 1200, 1800, and 2100 rounds of operations.

#### 5.2.11 Case study 11: AEC versus no of rounds when BS positioned at (100,250) and SN = 200

The simulation results, as shown in Fig. 16 and Table 5, indicate that the AEC is zero after 4400 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the AEC is zero after 2000, 2400, and 3200 rounds of operations, and the performance is better when SNs increase to 200.

#### 5.2.12 Case study 12: AEC versus no of rounds when BS positioned at (100,250) and SN = 200

The simulation results, as shown in Fig. 17 and Table 5, indicate that the AEC is zero after 5500 rounds of operations. In the case of LEACH, DEEC, and BOAACO, the AEC is zero after 2500, 3000, and 4000 rounds of operations, and the performance is better than in the case study 10, 11, when SNs increase to 300. Overall, the simulation results show that the POA_Proposed performs the best in terms of energy efficiency, consistently exhibiting the lowest AEC values across all rounds. LEACH, on the other hand, consistently consumes the highest amount of energy.

#### 5.2.13 Case study 13: total packets received by base station when placed at (100,100)

The findings of LEACH, DEEC, BOAACO, and the proposed system in terms of total data packets received by the BS are displayed in Fig. 18 and Table 6. The total data packets received by BS in the case of LEACH, DEEC, BOAACO, and the proposed method are 19,998, 59,989, 89,879, and 119,986 bytes when SNs are placed inside the environment field, and are 100. The total data packets

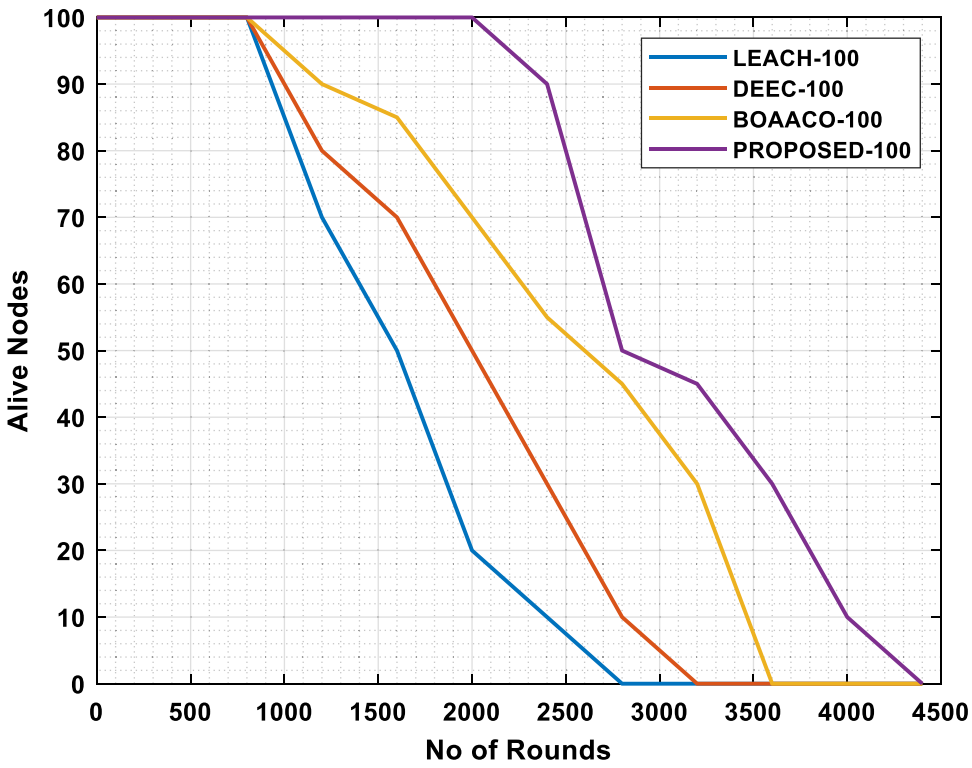



**Fig. 6** ALNs versus number of rounds when BS positioned at (100,100) and sensor node = 100

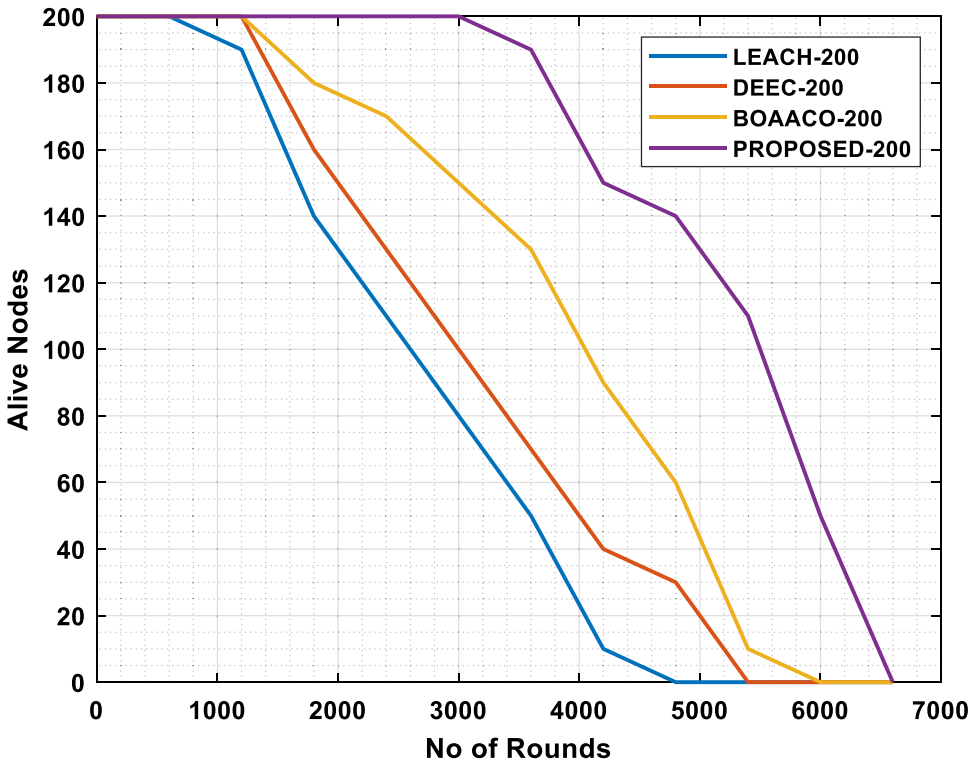



**Fig. 7** ALNs versus no of rounds when BS positioned at (100,100) and sensor node = 200

received by BS in the case of LEACH, DEEC, BOAACO, and the proposed scheme are 49,996,149,989, 169,987, and 239,993 when SNs deployed are 200. The total data packets received by BS in the case of LEACH, DEEC, BOAACO, and the proposed method are 99,989, 169,995, 199,989, and 289,987 Bytes when SNs deployed are 300.

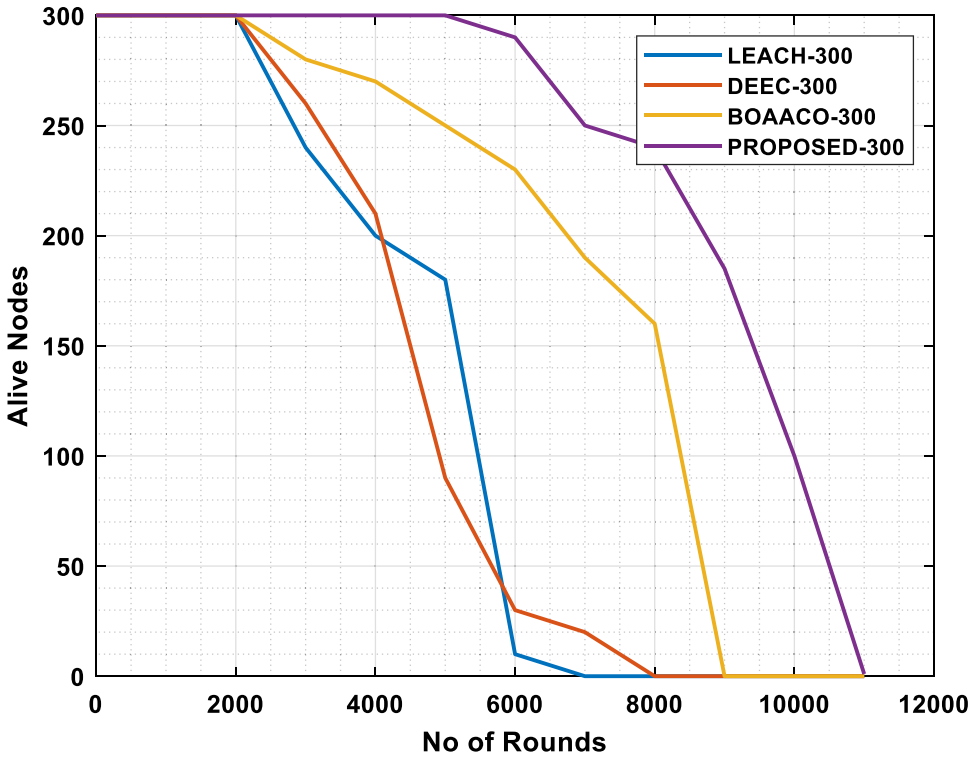



**Fig. 8** ALNs versus no of rounds when BS positioned at (100,100) and sensor node = 300

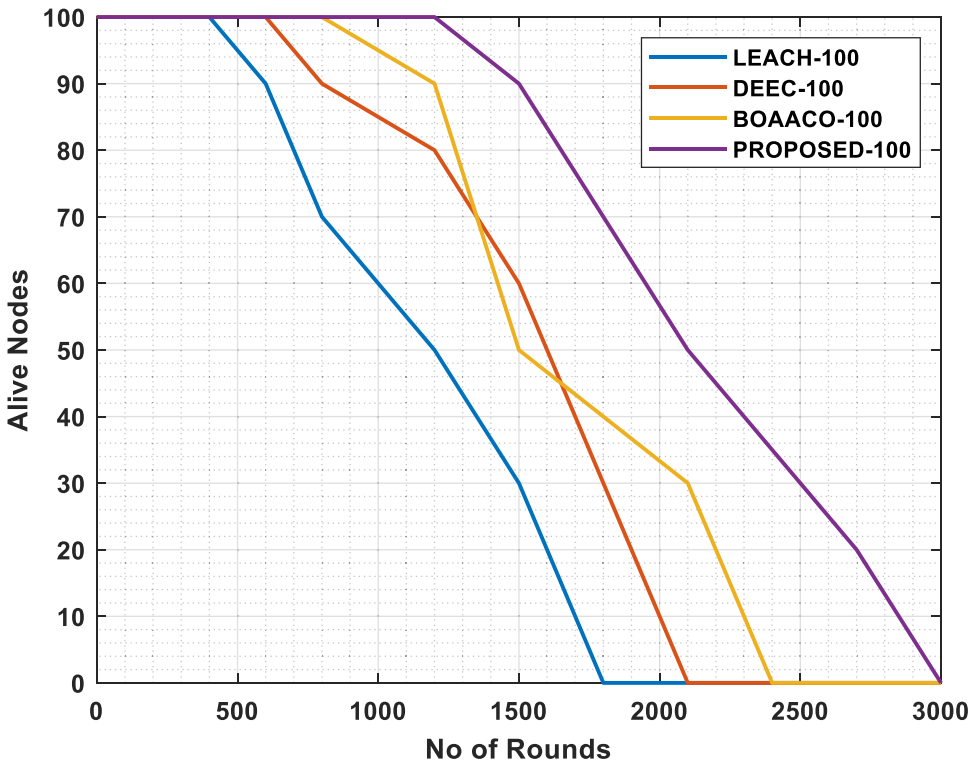



**Fig. 9** ALNs versus no of rounds when BS positioned at (100,250) and sensor node = 100

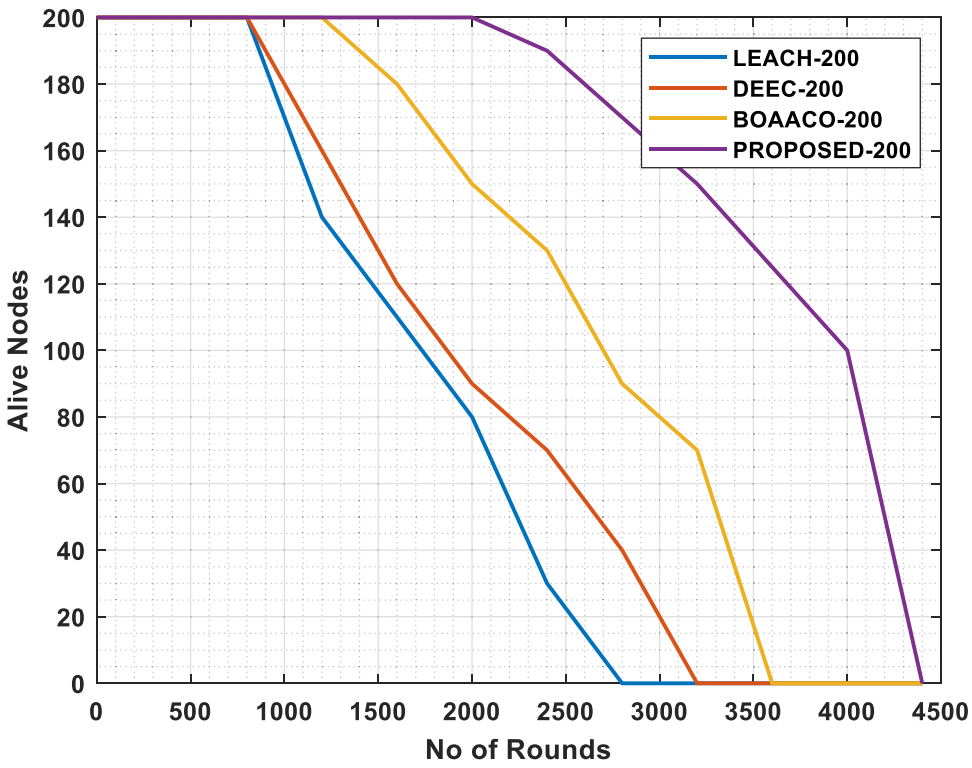



**Fig. 10** ALNs versus no of rounds when BS positioned at (100,250) and sensor node = 200

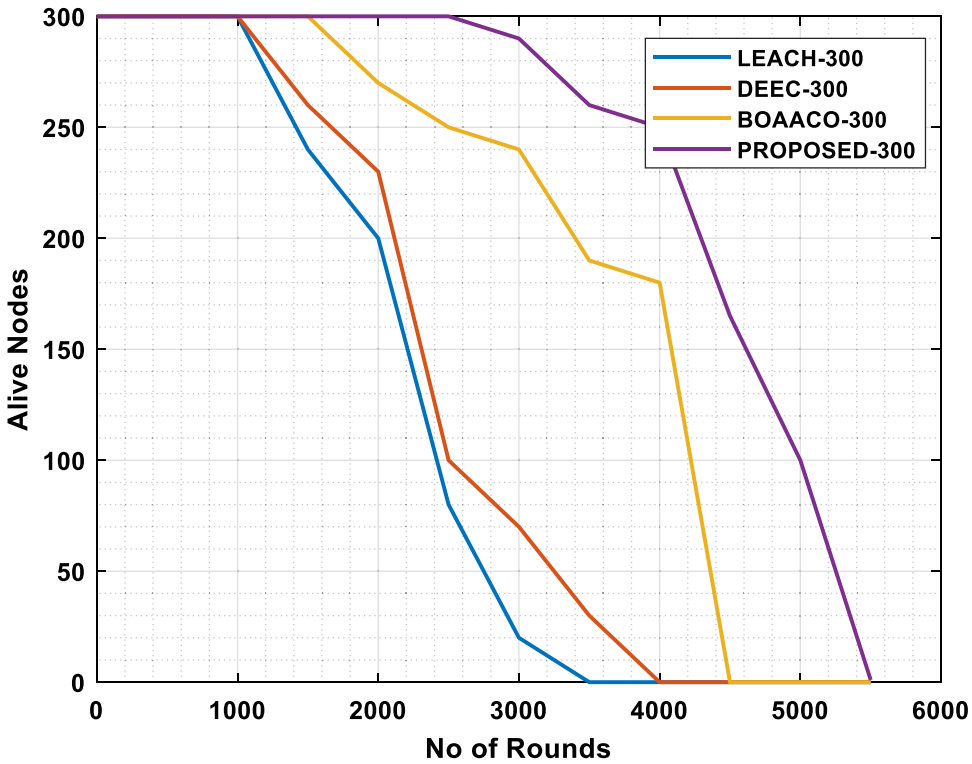



**Fig. 11** ALNs versus no of rounds when BS positioned at (100,250) and sensor node = 300

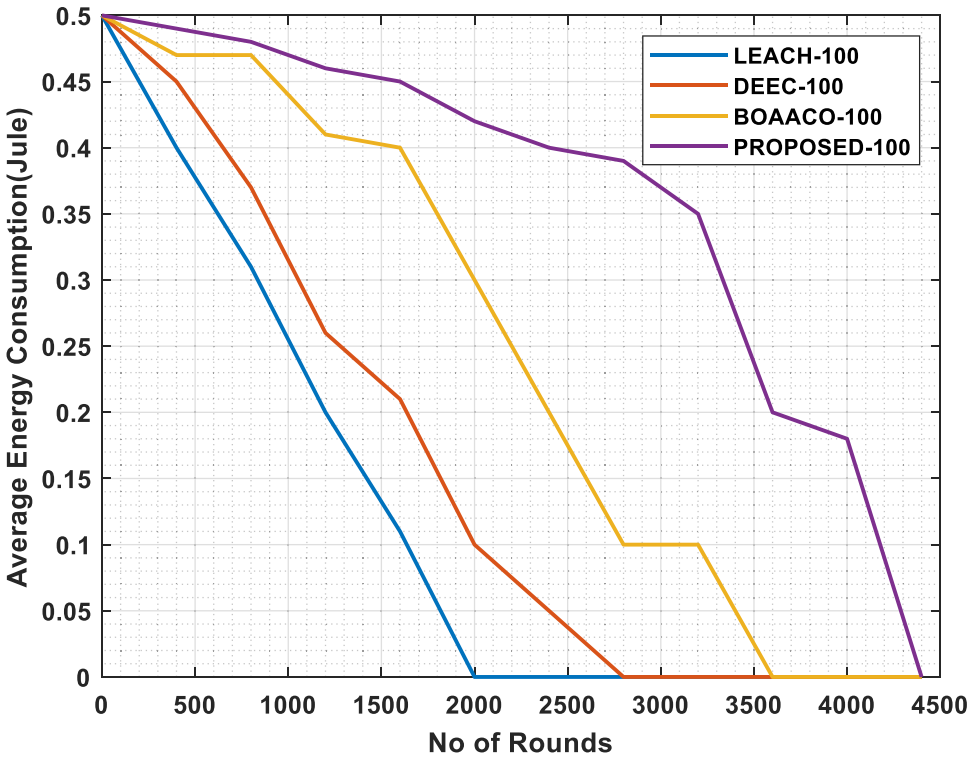



**Fig. 12** AEC versus no of rounds when BS positioned at (100,100) and SN = 100

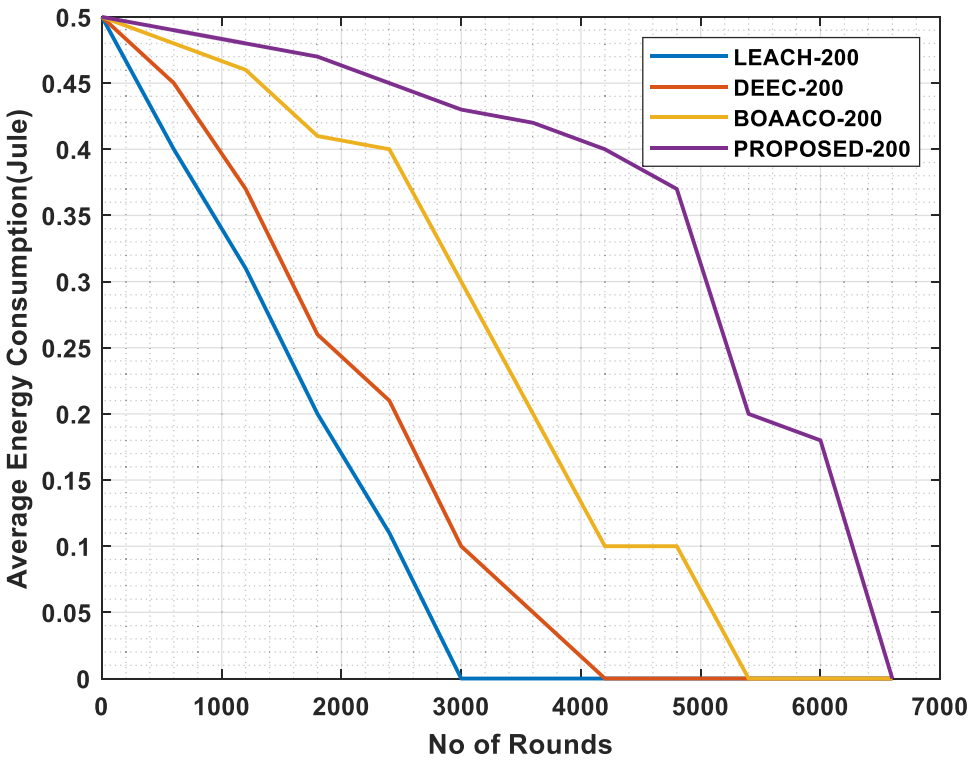



**Fig. 13** AEC versus no of rounds when BS positioned at (100,100) and sensor node = 200

**Table 5** Residual energy of LEACH, DEEC, BOAACO, and POA_Proposed

| No of rounds | LEACH | DEEC | BOAACO | POA_Proposed |
|---|---|---|---|---|
| *AEC versus no of rounds when BS positioned at (100,100) and sensor node = 100* | | | | |
| 400 | 0.4011 | 0.4485 | 0.4785 | 0.4905 |
| 800 | 0.3110 | 0.3711 | 0.4659 | 0.4811 |
| 1200 | 0.2020 | 0.2610 | 0.4109 | 0.4601 |
| 1600 | 0.1110 | 0.2110 | 0.4009 | 0.4501 |
| 2000 | 0.00 | 0.1001 | 0.3009 | 0.4211 |
| 2400 | 0.00 | 0.0510 | 0.2011 | 0.4011 |
| 2800 | 0.00 | 0.00 | 0.100 | 0.3901 |
| 3200 | 0.00 | 0.00 | 0.100 | 0.3491 |
| 3600 | 0.00 | 0.00 | 0.00 | 0.2001 |
| 4000 | 0.00 | 0.00 | 0.00 | 0.1811 |
| 4400 | 0.00 | 0.00 | 0.00 | 0.01 |
| *AEC versus no of rounds when BS positioned at (100,100) and sensor node = 200* | | | | |
| 600 | 0.4001 | 0.4501 | 0.4791 | 0.4910 |
| 1200 | 0.3101 | 0.3678 | 0.4587 | 0.4770 |
| 1800 | 0.2011 | 0.2610 | 0.4110 | 0.4720 |
| 2400 | 0.1101 | 0.2101 | 0.4010 | 0.4501 |
| 3000 | 0.00 | 0.1000 | 0.3001 | 0.4300 |
| 3600 | 0.00 | 0.0501 | 0.2000 | 0.4210 |
| 4200 | 0.00 | 0.00 | 0.1000 | 0.4000 |
| 4800 | 0.00 | 0.00 | 0.1000 | 0.3690 |
| 5400 | 0.00 | 0.00 | 0.00 | 0.2181 |
| 6000 | 0.00 | 0.00 | 0.00 | 0.1810 |
| 6600 | 0.00 | 0.00 | 0.00 | 0.00 |
| *AEC versus no of rounds when BS positioned at (100,100) and sensor node = 300* | | | | |
| 1000 | 0.4401 | 0.4677 | 0.4800 | 0.49 |
| 2000 | 0.3110 | 0.4415 | 0.4579 | 0.4788 |
| 3000 | 0.2200 | 0.3600 | 0.4400 | 0.4700 |
| 4000 | 0.1400 | 0.3100 | 0.4000 | 0.4500 |
| 5000 | 0.00 | 0.2000 | 0.3600 | 0.4300 |
| 6000 | 0.00 | 0.1501 | 0.2900 | 0.4100 |
| 7000 | 0.00 | 0.00 | 0.1900 | 0.4000 |
| 8000 | 0.00 | 0.00 | 0.1001 | 0.3500 |
| 9000 | 0.00 | 0.00 | 0.00 | 0.2800 |
| 10,000 | 0.0 | 0.00 | 0.00 | 0.1801 |
| 11,000 | 0.0 | 0.00 | 0.00 | 0.00 |
| *AEC versus no of rounds when BS positioned at (100,250) and sensor node = 100* | | | | |
| 200 | 0.4101 | 0.4601 | 0.4700 | 0.4900 |
| 400 | 0.3111 | 0.3501 | 0.4101 | 0.4680 |
| 600 | 0.2200 | 0.2300 | 0.3801 | 0.4520 |
| 800 | 0.1000 | 0.2000 | 0.3000 | 0.4401 |
| 1200 | 0.00 | 0.1510 | 0.2800 | 0.4000 |
| 1500 | 0.00 | 0.1000 | 0.1801 | 0.3600 |
| 1800 | 0.00 | 0.00 | 0.0500 | 0.3200 |
| 2100 | 0.00 | 0.00 | 0.00 | 0.3000 |
| 2400 | 0.00 | 0.00 | 0.00 | 0.2401 |
| 2700 | 0.00 | 0.00 | 0.00 | 0.1700 |
| 3000 | 0.00 | 0.00 | 0.00 | 0.00 |
| *AEC versus no of rounds when BS positioned at (100,250) and sensor node = 200* | | | | |
| 400 | 0.4110 | 0.4200 | 0.4700 | 0.4900 |
| 800 | 0.3201 | 0.3520 | 0.4601 | 0.4779 |
| 1200 | 0.2203 | 0.2300 | 0.4111 | 0.4500 |
| 1600 | 0.1000 | 0.2001 | 0.3900 | 0.4200 |
| 2000 | 0.00 | 0.1000 | 0.3200 | 0.3700 |
| 2400 | 0.00 | 0.00 | 0.1800 | 0.3600 |
| 2800 | 0.00 | 0.00 | 0.1500 | 0.3200 |
| 3200 | 0.00 | 0.00 | 0.00 | 0.2601 |
| 3600 | 0.00 | 0.00 | 0.00 | 0.1901 |
| 4000 | 0.00 | 0.00 | 0.00 | 0.1600 |
| 4400 | 0.00 | 0.00 | 0.0 | 0.00 |
| *AEC versus no of rounds when BS positioned at (100,250) and sensor node = 300* | | | | |
| 500 | 0.4110 | 0.4210 | 0.4700 | 0.4900 |
| 1000 | 0.3210 | 0.3500 | 0.4224 | 0.4790 |
| 1500 | 0.2201 | 0.2301 | 0.3800 | 0.4700 |
| 2000 | 0.1000 | 0.2001 | 0.3000 | 0.4500 |
| 2500 | 0.00 | 0.1000 | 0.2810 | 0.4200 |
| 3000 | 0.00 | 0.00 | 0.1800 | 0.3700 |
| 3500 | 0.00 | 0.00 | 0.0500 | 0.3200 |
| 4000 | 0.00 | 0.00 | 0.00 | 0.2610 |
| 4500 | 0.00 | 0.00 | 0.00 | 0.2401 |
| 5000 | 0.00 | 0.00 | 0.00 | 0.1900 |
| 5500 | 0.00 | 0.00 | 0.00 | 0.00 |

The simulation result reveals that the total number of packets received by the BS is large when the SN density increases.

#### 5.2.14 Case study 14: total packets received by base station when placed at (100,250)

The findings of LEACH, DEEC, BOAACO, and the proposed scheme in terms of the total number of data packets received by the base station (BS) are presented in Fig. 19 and Table 6. The base station (BS) received a total of 13,997 data packets in the case of LEACH, 49,989 data packets in the case of DEEC, 69,988 data packets in the case of BOAACO, and 109,993 data packets in the case of the proposed scheme. These data packets were received when the sensor nodes (SNs) were deployed outside of the environment field in the center (100,250), and a total of 100 SNs were deployed. The total data packets received by BS in the case of LEACH, DEEC, BOAACO, and the proposed system are 29,997,109,995, 139,996, and 199,994 bytes when SNs distributed are 200. The base station (BS) received a total

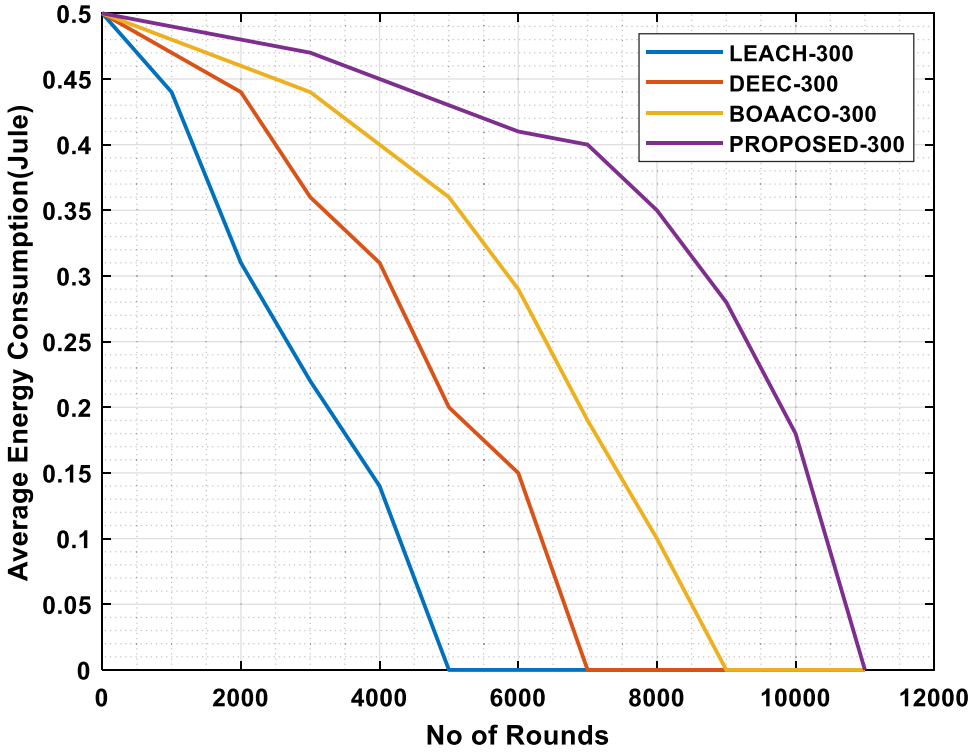



**Fig. 14** AEC versus no of rounds when BS positioned at (100,100) and sensor node = 300

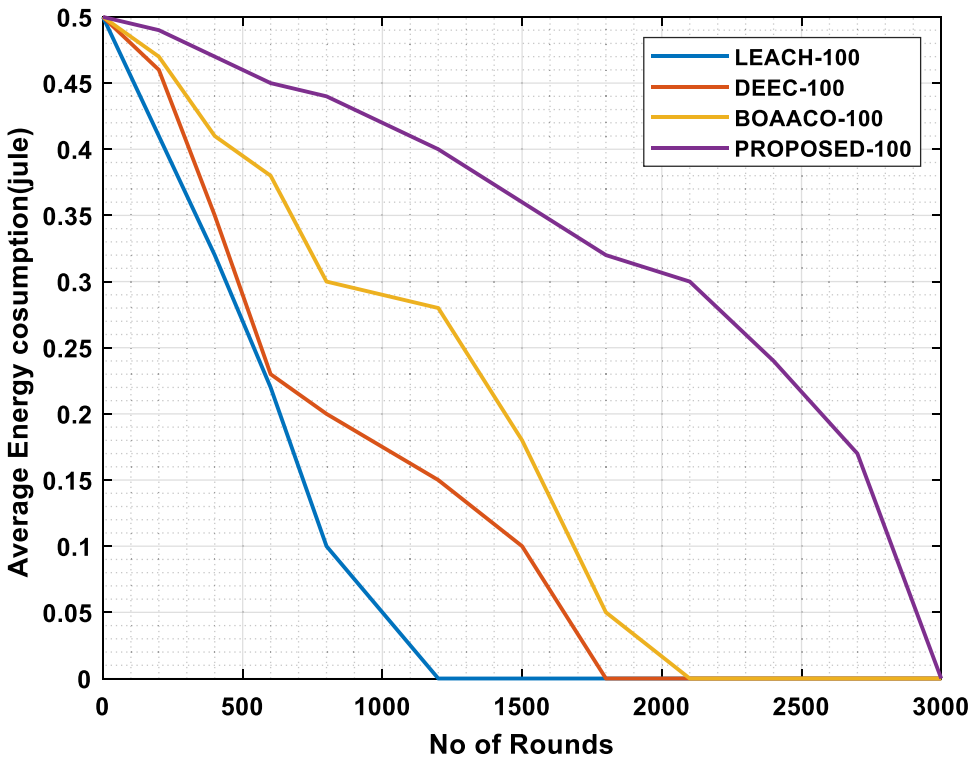



**Fig. 15** AEC versus no of rounds when BS positioned at (100,250) and sensor node = 100

of 79,992 data packets in the case of LEACH, 129,994 data packets in the case of DEEC, 149,993 data packets in the case of BOAACO, and 229,989 data packets in the case of the proposed scheme when 300 sensor nodes (SNs) were installed. As the number of SNs rises, the performance of the suggested scheme and the leaving method increases. In both circumstances, whether BS is deployed at the center or outside of the environment field, the total packets received by the BS increases, and performance is always greater when BS is positioned at the center of the environment field.

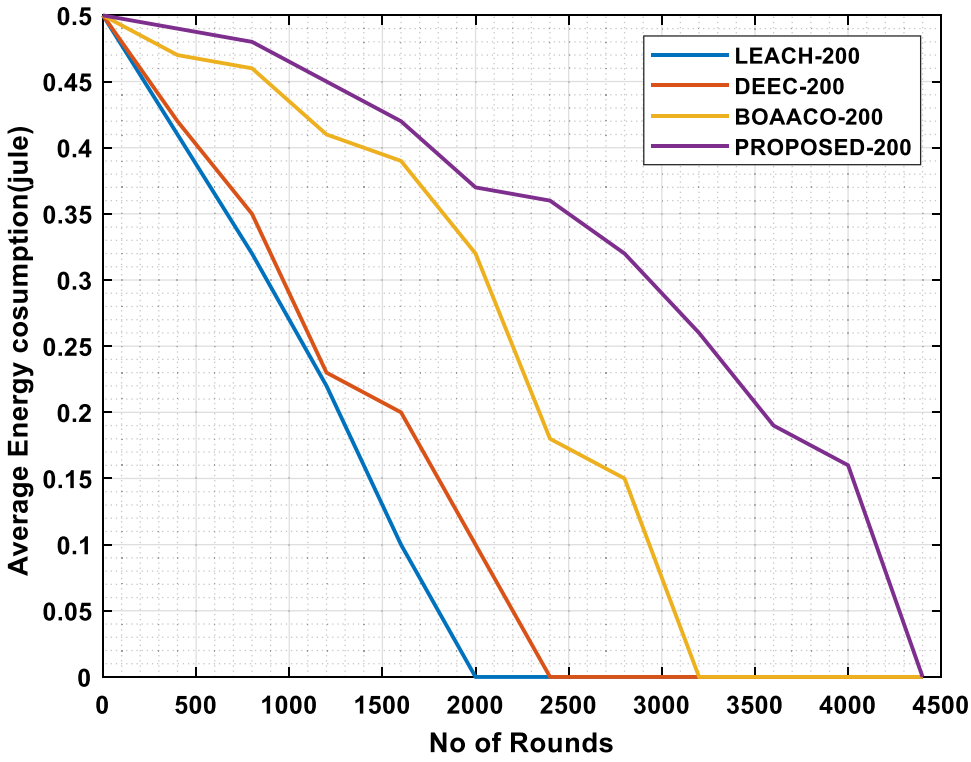



**Fig. 16** AEC versus no of rounds when BS positioned at (100,250) and sensor node = 200

The results demonstrate that the proposed model outperforms existing routing protocols in terms of energy efficiency and network lifetime extension. The impact of the modified POA is discussed on energy efficiency and how it outperforms traditional clustering and routing methods. We analyze the scalability of our proposed scheme in terms of network density, highlighting its adaptability to different deployment scenarios.

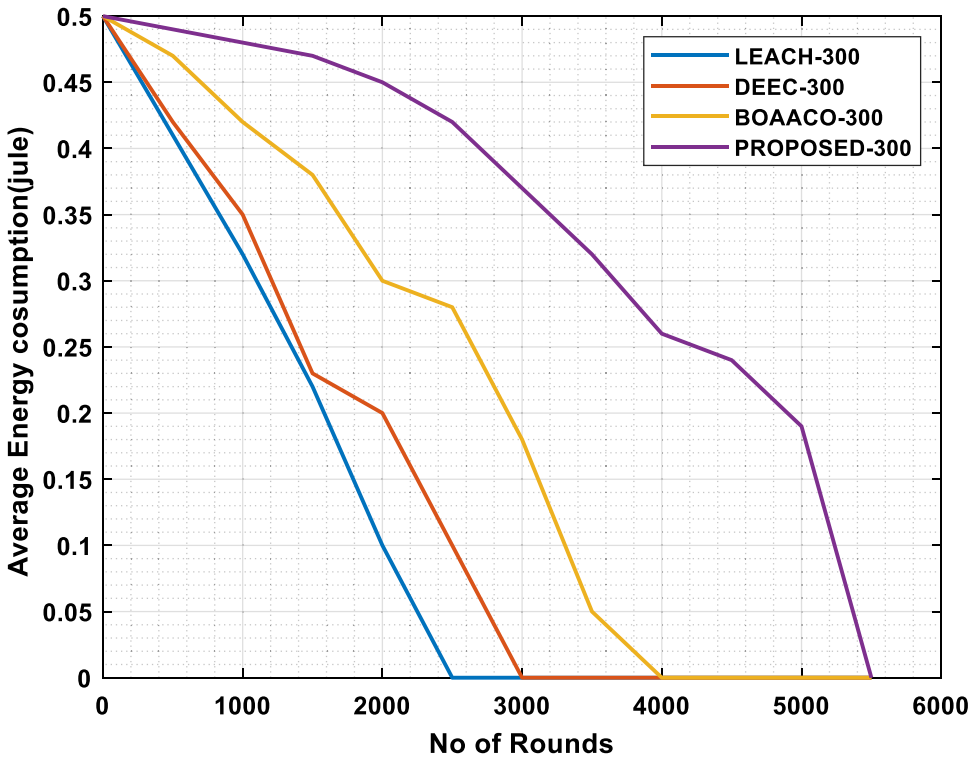



**Fig. 17** AEC versus no of rounds when BS at (100,250) and sensor node = 300

## 6 Conclusion and future work

The cluster-based routing has high computational complexity when the size of the network increases. The proposed scheme (POA_proposed) is modelled using a suitable fitness function to optimize cluster head selection, data routing. A total of fourteen different scenarios are chosen for the simulation of the network. The simulation results show that the proposed model is energy-efficient compared to LEACH, DEEC, AND BOAACO SCHEMES. The extensive simulations, demonstrate that proposed scheme is outperformed in terms of network life and energy consumption compared to existing cluster-based routing schemes. The POA_Proposed scheme holds great promise for various applications, ensuring the longevity of WSNs. This research work may be a valuable contribution to the advancement of cluster-based routing in WSNs and may be essential for various applications, including monitoring the environment, using smart farming techniques, and automating industrial processes.

The potential areas for future research, is further optimization of the meta-heuristic algorithm, adaptation of the protocol for mobile sensor networks, and real-world deployment considerations. The proposed scheme (POA_Proposed) does not consider environmental factors like path loss components and delays in forwarding the data from CHs to the BS. Future work aims to include path loss and delay and design better WSNs.

**Table 6** Packets received by BS of LEACH, DEEC, BOAACO, and POA_Proposed

| No of rounds | LEACH | DEEC | BOAACO | POA_Proposed |
|---|---|---|---|---|
| *Packets received (bytes) by BS when BS positioned at (100,100) and sensor node = 100* | | | | |
| 3000 | 19,998 | 59,989 | 89,879 | 119,986 |
| *Packets received (bytes) by BS when BS positioned at (100,100) and sensor node = 200* | | | | |
| 3000 | 49,996 | 149,989 | 169,987 | 239,993 |
| *Packets received (bytes) by BS when BS positioned at (100,100) and sensor node = 300* | | | | |
| 3000 | 99,989 | 169,995 | 199,989 | 289,987 |
| *Packets received (bytes) by BS when BS positioned at (100,250) and sensor node = 100* | | | | |
| 3000 | 13,997 | 49,989 | 69,988 | 109,993 |
| *Packets received (bytes) by BS when BS positioned at (100,250) and sensor node = 200* | | | | |
| 3000 | 29,997 | 109,995 | 139,996 | 199,994 |
| *Packets received (bytes) by BS when BS positioned at (100,250) and sensor node = 300* | | | | |
| 3000 | 79,992 | 129,994 | 149,993 | 229,989 |

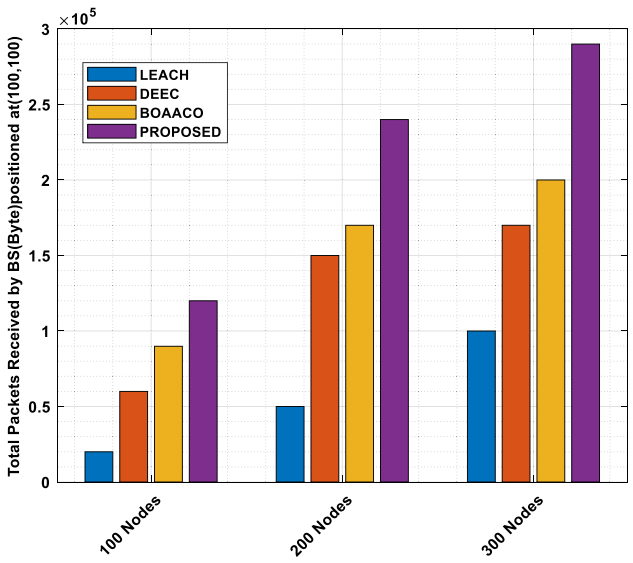



**Fig. 18** No of packets received by BS when BS at (100,100) and sensor nodes deployed are 100,200, and 200

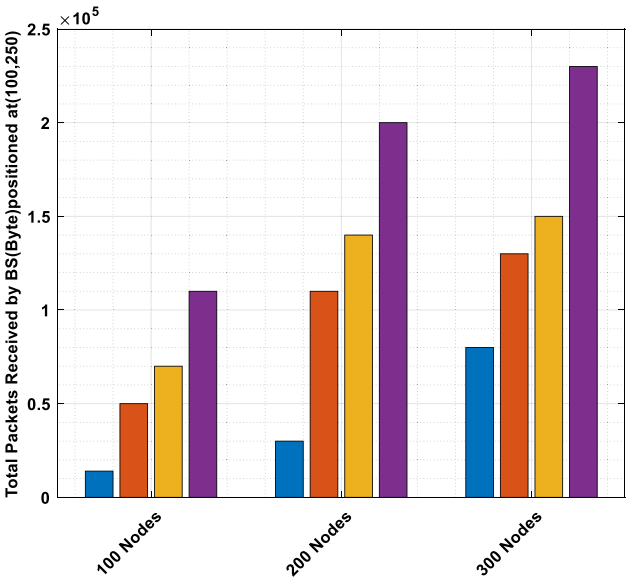



**Fig. 19** No of packets received by BS when BS at (100,250) and sensor nodes deployed are 100,200, and 200

**Author contributions** YM identified and analyzed the issue, a validation supported by RKY and MM. Additionally, YM authored the manuscript under the supervision of RKY and MM, with all authors contributing to the proofreading process.

**Funding** No funding was received to aid in the research of the manuscript.

**Availability of data and materials** No datasets were generated or analysed during the current study.

## Declarations

**Conflict of interest** The authors declare no competing interests.

**Ethical approval** No animal or human testing has been conducted by the authors for this manuscript, thus the ethical approval is not requirement for it.

## References


1. Zachariah, U. E., & Kuppusamy, L. (2022). A hybrid approach to energy efficient clustering and routing in wireless sensor networks. *Evolutionary Intelligence, 15*, 1–13.
2. Maheshwari, P., Sharma, A. K., & Verma, K. (2021). Energy efficient cluster-based routing protocol for WSN using butterfly optimization algorithm and ant colony optimization. *Ad Hoc Networks, 110*, 102317.
3. Lalwani, P., Das, S., Banka, H., & Kumar, C. (2018). CRHS: Clustering and routing in wireless sensor networks using harmony search algorithm. *Neural Computing and Applications, 30*, 639–659.
4. Singh, S. P., & Sharma, S. C. (2015). A survey on cluster-based routing protocols in wireless sensor networks. *Procedia Computer Science, 45*, 687–695.
5. Lalwani, P., Banka, H., & Kumar, C. (2018). BERA: A biogeography-based energy saving routing architecture for wireless sensor networks. *Soft Computing, 22*, 1651–1667.
6. Krishnan, K., Yamini, B., Alenazy, W. M., & Nalini, M. (2021). Energy-efficient cluster-based routing protocol for WSN based on hybrid BSO–TLBO optimization model. *The Computer Journal, 64*(10), 1477–1493.
7. Agarwal, V., Tapaswi, S., & Chanak, P. (2022). Energy-efficient mobile sink-based intelligent data routing scheme for wireless sensor networks. *IEEE Sensors Journal, 22*(10), 9881–9891.
8. Rambabu, B., Reddy, A. V., & Janakiraman, S. (2022). Hybrid artificial bee colony and monarchy butterfly optimization algorithm (HABC-MBOA)-based cluster head selection for WSNs. *Journal of King Saud University-Computer and Information Sciences, 34*(5), 1895–1905.
9. Bongale, A. M., Nirmala, C. R., & Bongale, A. M. (2019). Hybrid cluster head election for WSN based on firefly and harmony search algorithms. *Wireless Personal Communications, 106*, 275–306.
10. Nagarajan, L., & Thangavelu, S. (2021). Hybrid grey wolf sunflower optimization algorithm for energy-efficient cluster head selection in wireless sensor networks for lifetime enhancement. *IET Communications, 15*(3), 384–396.
11. Gupta, S. K., & Jana, P. K. (2015). Energy efficient clustering and routing algorithms for wireless sensor networks: GA based approach. *Wireless Personal Communications, 83*, 2403–2423.
12. Chauhan, V., & Soni, S. (2020). Mobile sink-based energy efficient cluster head selection strategy for wireless sensor networks. *Journal of Ambient Intelligence and Humanized Computing, 11*, 4453–4466.
13. Chaurasia, S., & Kumar, K. (2023). MOORP: Metaheuristic based optimized opportunistic routing protocol for wireless sensor network. *Wireless Personal Communications, 132*(2), 1241–1272.
14. Chaurasia, S., Kumar, K., & Kumar, N. (2023). Mocraw: A metaheuristic optimized cluster head selection-based routing algorithm for WSNs. *Ad Hoc Networks, 141*, 103079.
15. Almasri, M. M., & Alajlan, A. M. (2022). Modified optimization for efficient cluster-based routing protocol in wireless sensor network. *Intelligent Automation and Soft Computing, 33*(3), 1687–1710.
16. Kumar, R., & Kumar, D. (2016). Multi-objective fractional artificial bee colony algorithm to energy aware routing protocol in wireless sensor network. *Wireless Networks, 22*, 1461–1474.
17. Mosavvar, I., & Ghaffari, A. (2019). Data aggregation in wireless sensor networks using firefly algorithm. *Wireless Personal Communications, 104*, 307–324.
18. Al-Kaseem, B. R., Taha, Z. K., Abdulmajeed, S. W., & Al-Raweshidy, H. S. (2021). Optimized energy–efficient path planning strategy in WSN with multiple mobile sinks. *IEEE Access, 9*, 82833–82847.
19. Trojovský, P., & Dehghani, M. (2022). Pelican optimization algorithm: A novel nature-inspired algorithm for engineering applications. *Sensors, 22*(3), 855.
20. Praveen Kumar, R., Raj, J. S., & Smys, S. (2021). Performance analysis of hybrid optimization algorithm for virtual head selection in wireless sensor networks. *Wireless Personal Communications, 123*, 1–16.
21. Pitchaimanickam, B., & Murugaboopathi, G. (2020). A hybrid firefly algorithm with particle swarm optimization for energy efficient optimal cluster head selection in wireless sensor networks. *Neural Computing and Applications, 32*, 7709–7723.
22. Subramanian, P., Sahayaraj, J. M., Senthilkumar, S., & Alex, D. S. (2020). A hybrid grey wolf and crow search optimization algorithm-based optimal cluster head selection scheme for wireless sensor networks. *Wireless Personal Communications, 113*, 905–925.
23. Sengathir, J., Rajesh, A., Dhiman, G., Vimal, S., Yogaraja, C. A., & Viriyasitavat, W. (2022). A novel cluster head selection using Hybrid Artificial Bee Colony and Firefly Algorithm for network lifetime and stability in WSNs. *Connection Science, 34*(1), 387–408.
24. Yogarajan, G., & Revathi, T. (2018). Improved cluster-based data gathering using ant lion optimization in wireless sensor networks. *Wireless Personal Communications, 98*, 2711–2731.
25. Wang, Z., Ding, H., Li, B., Bao, L., & Yang, Z. (2020). An energy efficient routing protocol based on improved artificial bee colony algorithm for wireless sensor networks. *IEEE Access, 8*, 133577–133596.
26. Xu, C., Xiong, Z., Zhao, G., & Yu, S. (2019). An energy-efficient region source routing protocol for lifetime maximization in WSN. *IEEE Access, 7*, 135277–135289.
27. Xiang, W., Wang, N., & Zhou, Y. (2016). An energy-efficient routing algorithm for software-defined wireless sensor networks. *IEEE Sensors Journal, 16*(20), 7393–7400.
28. Sarkar, A., & Murugan, T. S. (2022). Analysis on dual algorithms for optimal cluster head selection in wireless sensor network. *Evolutionary Intelligence, 15*(2), 1471–1485.
29. Mekonnen, M. T., & Rao, K. N. (2017). Cluster optimization based on metaheuristic algorithms in wireless sensor networks. *Wireless Personal Communications, 97*, 2633–2647.
30. Sarkar, A., & Murugan, T. S. (2019). Cluster head selection for energy efficient and delay-less routing in wireless sensor network. *Wireless Networks, 25*, 303–320.

31. Rayenizadeh, M., Rafsanjani, M. K., & Saeid, A. B. (2022). Cluster head selection using hesitant fuzzy and firefly algorithm in wireless sensor networks. *Evolving Systems, 13*, 1–20.
32. Murugan, T. S., & Sarkar, A. (2018). Optimal cluster head selection by hybridization of firefly and grey wolf optimization. *International Journal of Wireless and Mobile Computing, 14*(3), 296–305.
33. Priyanka, B. N., Jayaparvathy, R., & DivyaBharathi, D. (2022). Efficient and dynamic cluster head selection for improving network lifetime in WSN using whale optimization algorithm. *Wireless Personal Communications, 123*(2), 1467–1481.
34. Sivakumar, D., Suganthi Devi, S., & Nalini, T. (2023). Energy aware clustering protocol using chaotic gorilla troops optimization algorithm for Wireless Sensor Networks. *Multimedia Tools and Applications, 83*, 1–19.
35. Wang, Z., Ding, H., Li, B., et al. (2022). Energy efficient cluster based routing protocol for WSN using firefly algorithm and ant colony optimization. *Wireless Personal Communication, 125*, 2167–2200. https://doi.org/10.1007/s11277-022-09651-9
36. Kathiroli, P., & Selvadurai, K. (2022). Energy efficient cluster head selection using improved Sparrow Search Algorithm in Wireless Sensor Networks. *Journal of King Saud University-Computer and Information Sciences, 34*(10), 8564–8575.
37. Wang, Z., Zhang, M., Gao, X., Wang, W., & Li, X. (2019). A clustering WSN routing protocol based on node energy and multipath. *Cluster Computing, 22*, 5811–5823.
38. Arjunan, S., & Sujatha, P. (2018). Lifetime maximization of wireless sensor network using fuzzy based unequal clustering and ACO based routing hybrid protocol. *Applied Intelligence, 48*, 2229–2246.
39. Baskaran, M., & Sadagopan, C. (2015). Synchronous firefly algorithm for cluster head selection in WSN. *The Scientific World Journal, 2015*, 1–7.
40. Wang, Z., Duan, J., Xu, H., Song, X., & Yang, Y. (2023). Enhanced pelican optimization algorithm for cluster head selection in heterogeneous wireless sensor networks. *Sensors, 23*(18), 7711.
41. Adumbabu, I., & Selvakumar, K. (2022). Energy efficient routing and dynamic cluster head selection using enhanced optimization algorithms for wireless sensor networks. *Energies, 15*(21), 8016.
42. Wu, S., Min, X., & Li, J. (2021). Optimal data transmission for wsns with data-location integration. *Symmetry, 13*(8), 1499.
43. Shalika, S., Meena, U., & Agarwal, A. (2023). Optimized routing protocol using metaheuristic techniques for improving energy efficiency in WSN.
44. Zhang, J., Chen, X., Wang, J., et al. (2023). A geodesic distance-based routing scheme for sensor networks with irregular terrain structure. *Wireless Networks, 29*, 3207–3221. https://doi.org/10.1007/s11276-023-03377-7
45. Zheng, B., Zhuo, K., Zhang, H., et al. (2024). A novel airborne greedy geographic routing protocol for flying ad hoc networks. *Wireless Networks, 30*, 4413–4427. https://doi.org/10.1007/s11276-022-03030-9
46. Thi, M. H., Toan, T. T., Vinh, N. T. Q., et al. (2024). An energy efficient routing algorithm in wireless underground sensor networks. *Wireless Networks, 30*, 2297–2311. https://doi.org/10.1007/s11276-024-03666-9
47. Yu, X., Liu, Y., & Liu, Y. (2024). WSN routing algorithm based on node classification and multi-layer transport. *Wireless Networks, 30*, 737–747. https://doi.org/10.1007/s11276-023-03497-0
48. Mohan, Y., Yadav, R. K., & Manjul, M. (2024). Seagull optimization algorithm for node localization in wireless sensor networks. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-18331-8
49. Mohan, Y., Yadav, R. K., & Manjul, M. (2023). PANLS: performance analysis of node localization scheme in WSNs. In *2023 3rd International conference on technological advancements in computational sciences (ICTACS), Tashkent, Uzbekistan* (pp. 950–956). https://doi.org/10.1109/ICTACS59847.2023.10390138.
50. Han, R., Yang, W., Wang, Y., & You, K. (2017). DCE: A distributed energy-efficient clustering protocol for wireless sensor network based on double-phase cluster-head election. *Sensors, 17*(5), 998.
51. Panchal, A., & Singh, R. K. (2021). EHCR-FCM: Energy efficient hierarchical clustering and routing using fuzzy C-means for wireless sensor networks. *Telecommunication Systems, 76*(2), 251–263.
52. Panchal, A., & Singh, R. K. (2021). EOCGS: Energy efficient optimum number of cluster head and grid head selection in wireless sensor networks. *Telecommunication Systems, 78*(1), 1–13.
53. Singh, R. K., Verma, S., Panchal, A., & Dubey, S. (2024). Modified RCH-LEACH (MRCH) for wireless sensor networks (WSN). In *International congress on information and communication technology* (pp. 331–340). Springer, Singapore.
54. Guo, J., Gao, H., Liu, Z., Huang, F., Zhang, J., Li, X., & Ma, J. (2022). ICRA: An intelligent clustering routing approach for UAV ad hoc networks. *IEEE Transactions on Intelligent Transportation Systems, 24*(2), 2447–2460.
55. Zheng, T., Tan, J., Wu, X., Hu, R., Chen, Q., Liu, Z., & Du, Y. (2024). SARM: A network state-aware adaptive routing mutation method for power IoT. *Computer Networks, 255*, 110889.


**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.



**Yogendra Mohan** is an assistant professor at the North Eastern Regional Institute of Science and Technology (Deemed-to-be-University Under the Ministry of Education, Govt. of India), Itanagar Arunachal Pradesh. He is a Dy. Head of Education Technology Cell, and Chief Information Officer, NERIST. He has 18 years of teaching and 2 years of software development experience. He received a B. Tech degree in I.T. from KNIT, Sultanpur, affiliated to AKTU Lucknow, in 2004 and an ME degree in CSE from NITTTR Chandigarh affiliated to Panjab University, in 2015. He is pursuing PhD from DTU, Delhi. He is the author of the book "Essence of Physics, Chemistry and Mathematics." He supervised 25 B.Tech. projects; and 7 M.Tech. thesis. He is a member of IEEE, ISTE, and IEI. His research areas are wireless sensor networks, network security, and meta-heuristic optimization.



**Rajesh Kumar Yadav** is received his B.E. (CSE) from G. B. Pant Engineering College, Pauri (H.N.B. Garahwal University), M.Tech. (CS) from UPTU, Lucknow, and Ph.D. (Computer Engineering) from Delhi Technological University (Formerly Delhi College of Engineering), Delhi, India. Presently, he is working as an Associate Professor in the Department of Computer Science & Engineering at Delhi Technological University (DTU). His areas of research interest are Mobile Ad-hoc Networks, Wireless Sensor Networks, and Internet of Things (IoT).



**Manisha Manjul** received the B.Tech. degree in computer science and engineering from KNIT, Sultanpur, India. She received M.Tech. Degree in computer science and engineering from NIT Jalandhar, and a Ph.D. degree in computer science and engineering from Gautam Buddha University, India. She is an assistant professor in the Department of Computer Engineering, G. B. Pant DESU, New Delhi, India. Her primary research interests include computer networks, network security, multicast communication, and object-oriented programming. She is also a Life Member of CSI, India.

2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3315-2169-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/IATMSI64286.2025.10985441

# Efficient Hardware Implementation of Multi-Armed Bandit Algorithms for RIS-Aided Wireless Networks

Ishaan Sharma[1], Rohit Kumar[1], and Sumit J Darak[2]

[1]Electronics and Communications Department, Delhi Technological University, India-110042
[2]Electronics and Communications Department,IIIT-Delhi, India-110020

***Abstract*—The potential of Reconfigurable Intelligent Surfaces (RIS) to support next-generation wireless communication systems has recently garnered significant attention. This research addresses the challenge of optimizing RIS-aided wireless networks, where RIS technology enhances signal control and coverage. However, in complex scenarios with multiple RIS units and receivers, selecting the optimal RIS blocks to maximize SNR becomes computationally intensive. Traditional methods relying on complete channel state information face high latency and scalability issues in dynamic environments. To overcome this, we propose a synthesizable Multi-Armed Bandit (MAB) algorithm, specifically the Upper Confidence Bound (UCB) algorithm, for RIS selection in real-time deployments on the PolarFire SoC. The focus is on efficient decision-making within constrained computational environments, leveraging a hardware-software co-design approach to optimize resource utilization. The implementation efficiently maps the algorithm to PolarFire SoC resources, including logic, memory, and I/O components, ensuring real-time processing performance in smart wireless environments.**

***Index Terms*—Multiarmed Bandit(MAB), Multi-Reconfigurable Intelligent Surfaces(multi-RIS), Hardware-Software Co-Design (HSCD), Polarfire SoC**

## I. Introduction

RECONFIGURABLE intelligent surfaces (RIS) have recently gained a significant consideration in a very short span enhancing the wireless communication networks by providing a seamless connectivity and high data rates [1]. Several studies has proven that the RIS offers reliability and increased throughput, and is considered as a low-cost solution. RIS has its applications in various domains like security[2, 3], next generation communication [4], minimizing transmit power [5] etc. RIS is being used in providing a secure network with eavesdroppers [2], 3D positioning of a target [6],extending the connectivity for 6G communication[4, 7], for Terahertz (THz) range [8], beam tracking in RIS-aided mmWave mobile system in 6G communication [9], minimizing transmitted signal power and boosting system transmission capacity [5].

Multi Armed Bandit (MAB) algorithms are a subset of reinforcement learning. A simple yet effective framework for decision-making in uncertain and dynamic contexts is offered by MAB algorithms. The objective of the MAB algorithm is to determine the optimal arm from a set of alternatives. In this paradigm, the learner's goal is to maximize cumulative rewards over several trials by strategically and iteratively selecting the optimal arm from a subset. This task is challenging since the statistical properties of each arm is unknown. Since these arms have never been explored before, the algorithm has to carefully strike a balance between exploration and exploitation. The low complexity and analytical tractability has given an upper edge to the MAB algorithms and hence they are used extensively. Our emphasis lies on the stochastic setting [10], using conventional MAB algorithms like UCB , KL-UCB , and Thompson sampling [11, 12] in network applications. MAB algorithms play a crucial role in wireless communication, Internet of Things (IoT) etc.

MAB algorithms have significant applications in beam selection for both RADAR systems[13] and RIS-aided wireless communication[14]. In RADAR systems, where beamforming is used to track and detect targets, MAB algorithms help dynamically select the optimal beam direction without prior knowledge of the environment. By balancing the exploration of different beam angles and the exploitation of previously successful beams, MAB ensures efficient target tracking while minimizing radar signal processing delays. The adaptive nature of MAB is particularly useful in dynamic environments, where the optimal beam direction may change rapidly due to target movement or environmental factors.

Similarly, in RIS-assisted wireless communication, MAB algorithms play a crucial role in optimizing signal reflection. RIS consists of passive reflecting elements that can adjust the phase and direction of incoming signals. MAB helps in selecting the best RIS configuration by evaluating multiple beam and phase shift options, ensuring the strongest possible signal reaches the receiver. With a large number of possible RIS configurations, MAB efficiently explores these options, learning which settings maximize signal strength and network performance over time. In both radar and RIS, MAB's ability to handle complex decision-making under uncertainty enhances system reliability and efficiency.

The need for MAB algorithms in RIS aided wireless systems stems from the challenge of optimizing RIS configurations in real-time to improve signal quality, coverage, and data rates. As RIS technology uses controllable reflective surfaces to direct and reflect wireless signals, these surfaces can be dynamically reconfigured to adapt to changing network con-

979-8-3315-2169-1/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:03:05 UTC from IEEE Xplore. Restrictions apply.

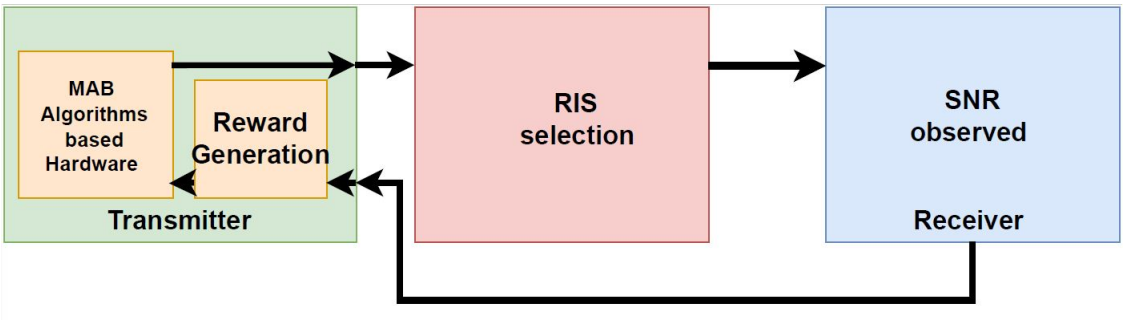



Fig. 1. Flowchart illustrating the implementation of the MAB algorithm hardware for dynamic RIS block selection to maximize SNR in real-time wireless communication.

ditions. However, in environments with multiple RIS units and several users, selecting the optimal combination of RIS configurations for each user becomes increasingly complex. Conventional methods, which require full knowledge of the channel conditions, often demand extensive computation and prior information about the environment. This approach is impractical in dynamic scenarios or for large networks, where obtaining and processing this data quickly enough is challenging. This adaptive approach enables RIS networks to efficiently discover optimal RIS blocks with minimal delay, maximizing throughput and improving communication reliability without needing constant, detailed channel state information. MAB algorithms have already been successfully mapped onto hardware[15–17]. However, MAB algorithms have never been used for RIS applications, especially on the Polarfire SoC. The Polarfire SoC presents several advantages over Zynq SoC, especially for applications like RIS. Polarfire SoC offers lower power consumption, which is crucial for energy-efficient RIS operations, making it more suitable for long-term deployment in wireless communication systems. Additionally, Polarfire SoC boasts higher security features, greater processing capabilities with its multi-core architecture, and superior I/O bandwidth compared to Zynq SoC, which makes it a better choice for handling the complex computations required by MAB algorithms in real-time RIS optimization. These factors position Polarfire SoC as a more efficient, scalable, and secure platform for advanced RIS applications, especially in scenarios requiring high-performance processing and energy optimization. Additionally, its deployment on resource-constrained edge platforms in real-world wireless systems introduces further complexity. These platforms often have limited computational power, memory, and energy, making it crucial for MAB algorithms to be fast and lightweight regarding computational overhead.

This work maps the MAB algorithm, specifically the UCB algorithm, for RIS selection in real-time deployments on the PolarFire SoC. The focus is on optimizing decision-making within constrained computational environments while ensuring efficient resource utilization. A dedicated hardware architecture is designed, employing a hardware-software co-design approach to achieve optimal mapping of the algorithm to the platform. The current implementation highlights the efficient use of PolarFire SoC resources, including logic, memory, and I/O components, to meet real-time processing requirements.

## II. System Model

In a multi-RIS-aided communication system, selecting the optimal RIS is framed as a MAB problem as shown in Fig.2. The aim is to balance exploration and exploitation to maximize the SNR across receivers. In this model, each RIS sub-block is treated as an ”arm,” offering a potential reward in the form of SNR, while the transmitter acts as the ”player,” selecting one RIS sub-block per time slot. The reward for each selection is the observed SNR, drawn from a stationary and independent distribution across different RIS. The transmitter must explore all RIS to gather information on their effectiveness while exploiting the knowledge to maximize SNR based on observed rewards as shown in Fig.1.

## III. Proposed Solution

To facilitate this, we will use hardware-level prototypes of MAB-based learning algorithms at the transmitter. A hardware-software co-design approach will be employed to implement MAB algorithms on a system-on-chip (SoC). As shown in Fig. 3, the Processing System (PS) will handle tasks such as generating arm rewards and conducting regret analysis, while the Programmable Logic (PL) on an FPGA will manage the arm-wise calculation of Q-values and select the optimal arm. This division of responsibilities ensures efficient parallel processing, allowing the PS to focus on higher-level functions and the PL to accelerate core computations. The UCB Q values in time slot $t$ is given as

$$UCB_i(t) = \frac{X_i(t)}{N_i(t)} + 2\sqrt{\frac{\log(t)}{N_i(t)}} \tag{1}$$

where $X_i(t)$ denotes the accumulated reward obtained over $N_i(t)$ time slots and $i$ is the selected arm.

## IV. Performance Analysis

We consider number of RIS as $K = 8$, total time horizon $T = 10000$ and mean probability distribution of the rewards for different RIS as $\text{RIS}_\mu$=$\{0.5, 0.6, 0.8, 0.4, 0.3, 0.9, 0.6, 0.7\}$, and it is assumed that the transmitter does not have any prior knowledge of the mean probability distribution. Thus, the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:03:05 UTC from IEEE Xplore. Restrictions apply.

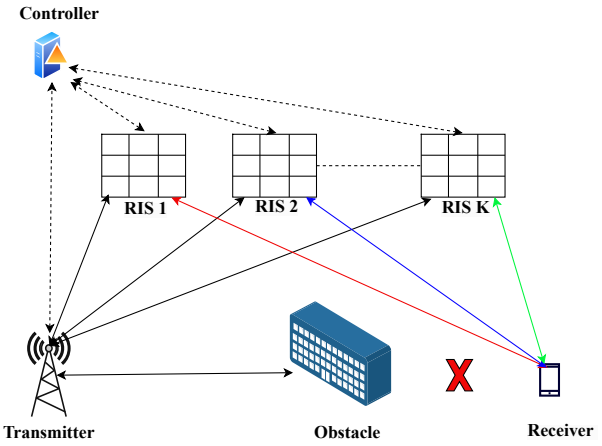



Fig. 2. Illustrations of network model for multi-RIS-aided wireless system with obstacle

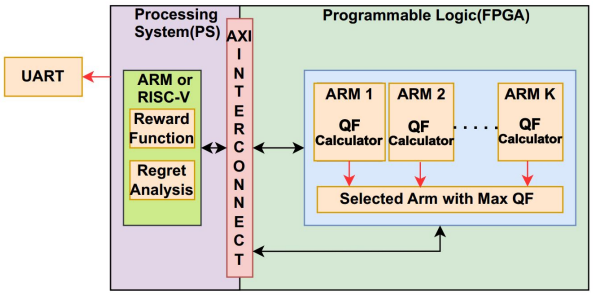



Fig. 3. Hardware software co-design based reconfigurable architecture of MAB algorithms.

aim of the MAB algorithm is to identify the optimal RIS via exploration-exploitation trade-off and choose an appropriate RIS to maximize the SNR.

### A. Regret Analysis

Regret is a statistic that is used to assess the effectiveness of the MAB algorithm and is closely related to the number of times sub-optimal arms are selected by the algorithm, lower regret means lesser number of times the sub-optimal arms are selected [18]. An ideal MAB algorithm guarantees less regret, usually obtaining a logarithmic regret, which is the best possible outcome in this situation. As shown in Fig. 4 (b).For initial stages we have used Upper confidence bound (UCB), by learning which RIS configuration result in the highest reward (e.g., the highest SNR). The agent (transmitter) could be trained to select the configuration with the highest UCB value (Q Factor) at each time step, balancing the trade-off between exploration and exploitation. The cumulative regret over $T$ time steps is given by:

$$\mathrm{R}(T) = \sum_{t=1}^{T} (r^*(t) - r(I_t)) \tag{2}$$

where $r^*(t)$ is the reward from the optimal arm and $r(I_t)$ the reward from the selected arm $I_t$ in time slot $t$.

For each experiment, the horizon is of size 10000-time slots and it can be seen in Fig. 4 (a). that the algorithm selects the RIS with higher average coefficients more times i.e. RIS 6 is selected in 9015-time slots out of 10000-time slots as shown in Fig. 4.

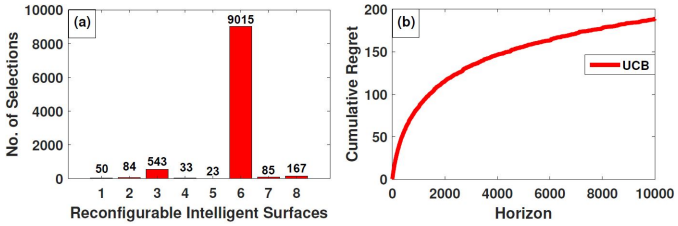



Fig. 4. (a)Selections made by the Transmitter throughout the time horizon (b) Average Cumulative regret of UCB algorithm

To implement the UCB algorithm, we used the PolarFire SoC as the target platform. An IP core for the algorithm was designed and integrated with the RISC-V processor through interconnects, enabling collaboration between the hardware and software components of the SoC. This hardware-software co-design approach allows efficient execution of the UCB algorithm, leveraging the combined processing power of custom hardware and the RISC-V processor, optimized for high performance and low latency on the PolarFire SoC platform. The resource utilization of SoC platform are given in Table I

TABLE I
RESOURCE UTILIZATION IN POLARFIRE SOC

| Resource Type | Total | Utilized | Percentage |
|---|---|---|---|
| **Fabric Interface + 4LUT** | 254196 | 51483+2700=54183 | 21.32 |
| **Fabric Interface + DFF** | 254196 | 54843+2700=57543 | 22.64 |
| **I/O Register** | 432 | 0 | 0.00 |
| **User I/O** | 144 | 65 | 45.14 |
| **uSRAM** | 2352 | 117 | 4.97 |
| **LSRAM** | 812 | 21 | 2.59 |
| **Math** | 784 | 15 | 1.91 |

### B. Resource Utilization

Table I and Fig. 5 shows the resource utilization metrics of the PolarFire SoC. It highlights the types of resources available in the FPGA fabric, their total capacity, the utilized amount, and the corresponding percentage utilization. Fabric Interface + 4LUT refers to the logic elements that combine lookup tables and fabric interfaces. Out of 254,196 total elements, 54,183 are utilized, amounting to 21.32% utilization. Similarly, Fabric Interface + DFF, which represents flip-flops integrated with the fabric interface, utilizes 57,543 out of the same total, reflecting a utilization of 22.64%. The I/O Registers, with a capacity of 432, show no utilization (0%). The User I/O resources, which facilitate input-output connectivity, utilize 65 out of 144, achieving a utilization of 45.14%. Memory resources include uSRAM and LSRAM, with 4.97% and 2.59% utilization, respectively. Of 2,352 available uSRAM

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:03:05 UTC from IEEE Xplore. Restrictions apply.

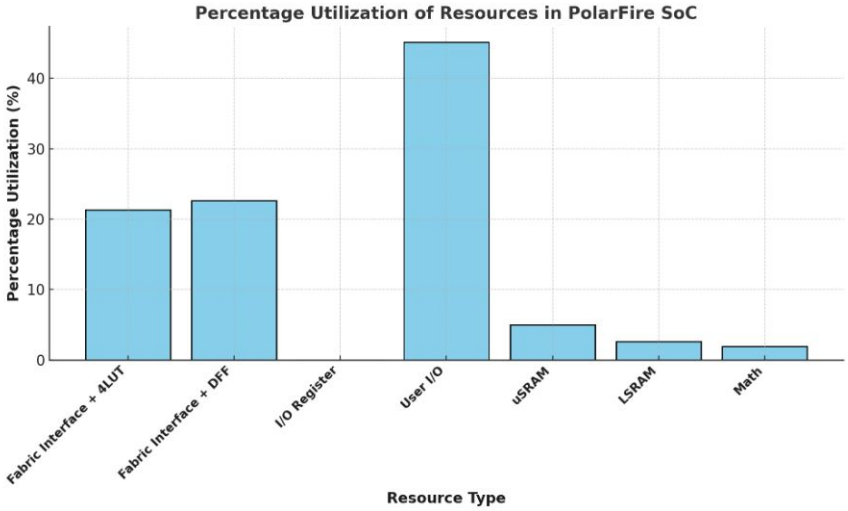



Fig. 5. Percentage Utilization of Resources in PolarFire SoC

blocks, 117 are used, while only 21 out of 812 LSRAM blocks are occupied. Finally, Math blocks, designed for arithmetic operations, show a utilization of 1.91% (15 out of 784).

## V. Conclusions and Future Work

In this work, we present a synthesizable design of the UCB algorithm tailored for RIS selection in edge-level deployment within smart environments. Future work will focus on extending the algorithm to handle a larger number of arms and further optimizing the proposed IP core. Additionally, future work will include the calculation of execution time and a comprehensive comparison of the performance of the PolarFire SoC with the ZYNQ SoC. Furthermore, the work will explore mapping pre-existing algorithms that are more efficient than UCB for scenarios with a large number of arms, providing insights into their adaptability and performance in similar edge-level deployments. These enhancements aim to reduce execution time and resource consumption, enabling detailed performance comparisons in terms of efficiency and scalability on edge platforms.

## References


[1] E. Basar, G. C. Alexandropoulos, Y. Liu, Q. Wu, S. Jin, C. Yuen, O. A. Dobre, and R. Schober, "Reconfigurable intelligent surfaces for 6g: Emerging hardware architectures, applications, and open challenges," *IEEE Vehicular Technology Magazine*, vol. 19, no. 3, pp. 27–47, 2024.

[2] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 375–388, 2020.

[3] W. Khalid, M. A. U. Rehman, T. Van Chien, Z. Kaleem, H. Lee, and H. Yu, "Reconfigurable intelligent surface for physical layer security in 6g-iot: Designs, issues, and advances," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 3599–3613, 2024.

[4] Z. Zhang, L. Dai, X. Chen, C. Liu, F. Yang, R. Schober, and H. V. Poor, "Active ris vs. passive ris: Which will prevail in 6g?," *IEEE Transactions on Communications*, vol. 71, no. 3, pp. 1707–1725, 2023.

[5] A. Khaleel and E. Basar, "Reconfigurable intelligent surface-empowered mimo systems," *IEEE Systems Journal*, vol. 15, no. 3, pp. 4358–4366, 2021.

[6] K. Keykhosravi, M. F. Keskin, S. Dwivedi, G. Seco-Granados, and H. Wymeersch, "Semi-passive 3d positioning of multiple ris-enabled users," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 11073–11077, 2021.

[7] X. Mu, J. Xu, Y. Liu, and L. Hanzo, "Reconfigurable intelligent surface-aided near-field communications for 6g: Opportunities and challenges," *IEEE Vehicular Technology Magazine*, vol. 19, no. 1, pp. 65–74, 2024.

[8] T. L. Nguyen, T. N. Do, G. Kaddoum, D. B. d. Costa, and Z. J. Haas, "Channel characterization for ris-aided terahertz communications: A stochastic approach," *IEEE Wireless Communications Letters*, vol. 11, no. 9, pp. 1890–1894, 2022.

[9] P. Zhang, J. Zhang, H. Xiao, H. Du, D. Niyato, and B. Ai, "Ris-aided 6g communication system with accurate traceable user mobility," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 2718–2722, 2023.

[10] A. Slivkins *et al.*, "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, pp. 235–256, 2002.

[12] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[13] A. Sneh, S. Darak, S. S. Ram, and M. Hanawal, "Radar enhanced multi-armed bandit for rapid beam selection in millimeter wave communications," *IEEE Communications Letters*, vol. 27, no. 9, pp. 2441–2445, 2023.

[14] I. Sharma, R. Kumar, and S. J. Darak, "Online-learning-based multi-ris-aided wireless systems," *IEEE Systems Journal*, vol. 18, no. 2, pp. 1174–1185, 2024.

[15] S. V. S. Santosh and S. J. Darak, "Multiarmed bandit algorithms on zynq system-on-chip: Go frequentist or bayesian?," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.

[16] H. Joshi, S. Santra, S. J. Darak, M. K. Hanawal, and S. V. S. Santosh, "Multiplay multiarmed bandit algorithm based sensing of noncontiguous wideband spectrum for aiot networks," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3337–3348, 2022.

[17] S. Santosh and S. J. Darak, "Intelligent and reconfigurable architecture for kl divergence based online machine learning algorithm," *arXiv preprint arXiv:2002.07713*, 2020.

[18] S. Bubeck, N. Cesa-Bianchi, *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:03:05 UTC from IEEE Xplore. Restrictions apply.

# Empathizing with parents and teachers for educating children with special learning needs

Shraddha Chadha[1] , Ravindra Singh[2] and Lalit Kumar Das[3]

[1] Research Scholar, Department of Design, Delhi Technological University
[2] Assistant Professor, Department of Design, Delhi Technological University
[3] Honorary Distinguished Faculty, Department of Design, Delhi TechnologicalUniversity
chadhashraddha@gmail.com

**Abstract.** It's difficult and demanding to raise a child with exceptional learning needs. The inclusion of children with exceptional learning needs in all facets of their educational journeys must be a priority for society in this dynamic environment. To incorporate as many children on the learning spectrum as possible, the learning methodologies should be varied. Parents and instructors may face a variety of difficulties with the children, which may have a serious effect on their mental health. Parenting children with learning disabilities requires a high level of competence as well as access to resources, knowledge, and help. But in developing countries, these resources and services aren't always available. Here, the main focus is on the challenges that parents encounter. The challenges, coping strategies, and significant insights that parents of children with learning problems shared were the focus of this study. The qualitative study combined a qualitative research approach. The convenience and snowball sampling technique was used to get in touch with parents, special educators, and psychologists of children in India who have exceptional learning requirements. In addition to feeling worn out and having to take on more caregiving responsibilities, the parents also experienced learning difficulties in their daily lives, and some even experienced financial difficulties. The parents' coping strategies also included creating a support network, establishing regular and scheduled activities for their children, and embracing and staying hopeful. The paper's conclusion suggests design interventions to get over obstacles when working with kids who have special needs and establish a supportive teaching and learning environment for all-around development.

**Keywords:** special education, learning, parents, teachers, challenges

© The Author(s) 2025
A. Anand and S. Balasubramanian (eds.), *Proceedings of Sustainability, Entrepreneurship, Equity and Digital Strategies (SEEDS 2024)*, Atlantis Highlights in Economics, Business and Management 3,
https://doi.org/10.2991/978-94-6463-714-4_7

## 1 Introduction

Learning is an integral part of any human. But the capacity and pace to learn can vary from one individual to another. There are various cognitive and psychological factors that are attached to it. The attention needs to be laid on the learners that are not supposedly on the same pace as most of the others. Inclusivity in the society is much required for which steps have been devised too. Attention to details and improvisations has been made a part of various institutions but not made affordable for the mass. When it comes to children they are supervised and are directed at each step, making it essential to train professionals like teachers to equip them in dealing with uniqueness of each child. The parents at the same time need acceptance and right direction to make the child learn considering their strengths and weaknesses. The gaps can be bridged with simple steps and right improvisations.

### 1.1 The objectives of this paper are:

1. To understand challenges of special educators and psychologists while interacting with parents having special children.
2. To understand challenges faced by parents while interacting with children having special needs.
3. To propose design interventions to overcome challenges while interacting children with special needs.

## 2 Literature Study

### 2.1 Child development and early learning

The young children respond to all the actions directed at them as well as react to the words too. The behavior of the children depends more on the early cognitive development. The demonstration of knowledge by children is limited these days because of the conflicting demands on their attention, limited abilities and irregular behavioral aspects. The rapid development of a child in terms of language they learn, the way they mimic people around, developing problem solving skills and critical thinking can be observed through experiential observations and experiments. [1] The learnings based on play and inclusivity can be a very important tactic to encourage children being involved and help raise all round abilities after a certain age. [2] As we are progressing, the education system is being expected to provide a more holistic development towards learning. To help a child thrive in the system, ultimately the society will have to join hands and to reach the full potential the changes will have to be equipped well. [3] The natural curious behaviour of the children can challenge the educators but idea of including play to learn is developmentally appropriate. Along with play therapy and teaching practices need modification in the early years to advance the goals of learning. [2, 4] The challenge that comes along learning through play are regarding the assessments and answerability. There is a gap in developing instruments and consistency while measuring the gains associated with learning through play [5] A few nations are working on creating curricular frameworks with a greater emphasis on academic subjects, goals that can be quantified, and learning outcomes in order to improve accountability and guarantee school preparedness. [6,7] The output children yield in learning is not actually what they experience in person but it is what they have learnt from the people around. The idea can be about the conceptual learning that comes naturally, providing the environment and how each one interacts to grow. [8] There are seen and unseen behaviors, that account for self-contained explanatory theories. The boundaries that we set to the disciplinary issues, leaves less to no space for the innovation or divergence to newer aspects. [9] The complex information can be using visual aids to make it convenient to understand. Further when the language used for communication is simplified it lowers the anxiety levels, and enhances understanding. Learning can be made convenient and more interesting when it is tailored to each student's particular needs and abilities. [10] There are developmental learning difficulties that affect many children, restricting them of the experiences in classroom learning and creating hindrances in many other

aspects of their life. There have been researches that reveal the altered neural networks of the children with developmental difficulties. While there have been interventions and considerations, there are benefits of integrating remedies within the other activities to encourage motivation and improvisations. [11]

## 2.2 Slow Learners and their requirements

An intrinsic part of a nation's development is education. The parents firmly believe that they should be providing a joyful and healthy learning environment to their child. Whereas a teacher contributes to the effectiveness in the classroom, that can contribute in making this a reality. However, there could be a variety of reasons, which can create hindrances in accomplishing this. [13] There are a some decisions that make the learning experience varied like the educational institute's selection, and working through the concerns regarding the day to day practices. [22] These decisions can be based upon the ability of the child to perform academically in the institutions. There are a variety of learners, broadly categorising them seems like the following: Quick learners, Average Learners and slow learners. These could be based on the knowledge gap that they have while learning, psychological lacks, memorising issues and other external challenges. If we discuss in detail about the slow learning children, they need extra assistance in each step. [13, 19] Slow learners are children who need special attention and strategies to cope up. There could be some factors that influence this like family structure, cultural backgrounds, economic factors, poor school facilities or genetic factors as well. There is a need for accommodations in terms of facilities, structures and systems that need inclusion in order to meet the requirements for these special learners. [14]

## 2.3 Role of special education and teachers

Teaching is definitely a demanding profession. The roles of teachers have become more demanding as the global standards expect the outcomes from learning to have a holistic development in terms of skills, knowledge making it a crowded curriculum overall [4] Identifying slow learners in the mainstream setup is an essential task for the teachers to create a foundation for developing strategies to support them. There will be a blended learning environment in the classroom based on the child's requirements and level of insight of the teachers. The identifications and strategies are to be executed by the teachers but special educators play a very important role in helping them develop the same. It is about the collaboration of teachers, parents, special educators and school counsellors. [15, 20] Words of encouragement, building support systems, interactive sessions help in developing individualized learning plans to boost the performance of children. It also encourages them to come at par with the other learners in the mainstream setups. The research also yielded and raised some important points as to how teachers have adopted methods in classrooms to help the children. They used peer learning methods, providing extra time, maintaining a record of their progress, reward system in terms of conversations. [15] In order to increase the learner's overall performance, teachers must practice patience, be progressive in their thoughts, and inspirational. The children classified under moderate learners, slow learners as some of them will not be able to adjust to the teachings taught in the classroom as well as their peers. On their own merits, a good understudy can receive good grades with little effort on the teacher's behalf. Since the teachers have to deal with these kinds of problems in the classroom, various approaches should be taken with the intention of assisting in achieving good grades and graduating successfully. When compared to their peers, ease-back learners should receive more attention and practice. [16]

### 2.4 Role of parents

A child has a natural need to bond with the parents, to make them feel secure both emotionally and psychologically, children who experiences care and has been given a helping hand at most of the steps would increase the chances of having a secure attachment pattern. [17] This key factor of attachment helps the child despite the hardships in severe situations. [21] The young minds have a feeling of protection, having a positive impact on their mental health too. It has shown through research that parents who are aware and displaying positive attitude makes the child's journey easy [17, 21] The process of identifying the special needs in children also has an impact on the parents journey both emotionally and physically. Parents have a major role to play in facilitating safety, quality and inclusivity of children with special needs. [22] It came to the notice through this research that mothers were able to devote more time to the children in comparison to the father. The fathers were only able to spend quality time on the weekends. Also, when they showed positive behavior in accepting the conditions it made the situation easier to cope up with. [21] There are a few decisions that need the attention of both the parents, when it comes to the decisions of educational requirements, selection process of school, raising concerns regarding the practices definitely has an impact in the later stages of a child's life. [22]

## 2.5 Role of psychologists

Psychology as a subject has been a part of the major contribution in providing solutions for special needs. It works on the perspectives of considering all the stake holders as a part of the system. [24] Collaboration amongst all in relation to the child increase the social interaction and aspects related to it. Assistance of the psychologist to the parents helps them effectively to deal with various situations and the stress that comes along. [21]

## 2.6 Role of collaborative efforts

The research has put focus on the collaborative efforts of the major stake holders in the life of a child for their development; parents, teachers, psychologists and special educators. [20, 23] It is essential to collaborate in order to formulate special education services and implementation of programs. If they meet with the requirements of the children's needs, the collaboration can progress to inclusive education too. [20, 23] Challenges always emerge, but if the bond is nurturing it anchors trust, motivation and a great psychological environment. Creating an inclusive environment and collaboration is influenced by some crucial factors including funds and diagnosis, procedures for communication, practical collaborations, considerations to inclusions. [18, 23] Requirements usually include services like training the professionals to enhance skills from time to time. Understanding the setting, attempting to experiment with the process of extensive learning in a mainstream set up. [19] Making special learning needs a part of the mainstream education can be challenging but necessary and need of the hour. The ultimate goal is to motivate the stake holders to modify the learning parameters, improvising the strategies and cultivating a positive surrounding for the child. [18, 23]

## 2.7 Design intervention for children

The attention spans of the children are short, and they have trouble focussing. The early understanding of how tools work most likely are developed to stop the spread of false information or improper usage of items, which can obstruct technology. [27] The results includes, being drawn to the products with intriguing combinations, functionalities, structures, and appearances. Based on the needs of children's intellectual growth at various levels, as well as on consumption, entertainment, and improving interactive experiences, the focus on intelligence aims to develop the brain and hand in tandem with the puzzle effect. Product design should encourage children's thinking, foster their capacity for imaginative and creative thought, and reflect creativity and imagination. Additionally to the above having the child as a co-designer broadens the mindset and the opportunities along with the constraints. This process of co-designing could be an essential part to accomplish desired results. The cognitive growth of the child should be taken into account by designers, which further can help modify the techniques so that they can be viewed as useful communication tools rather than barriers. [25, 26, 28] While designing for children apart from co-designing opportunities, comfortability, convenience, safety concerns, and other factors should be considered in products; larger focus area should be placed on the intricate design, market demands, and production costs. The communication between designers and children being a part of the process can be beneficial but challenging. Both the participants will be viewing the design and situation from the type of experiences that they have come across in their lives. The ultimate product/ ideation should be away from complications and details in the designs should not create fatigue for the target user, it should be flexible cultivating interest and conscious learning. [25, 26]

Through the literature studied there are multiple gaps that can be addressed in order to create a congenial environment for special learning needs. Though the needs are varied but there are a few systems that can be taken into account. There is a need to have personalized learning plans for children who are identified in the classrooms as slow learners. There is further a need to have continuous teacher training programs to equip them well for handing these children in the classroom set up without making them feel isolated. The inclusivity would not only be due to the special educators but in the classroom set ups as well. Though there are existing programs that suggest the same but are not in regular practices. Along with the teachers there should be awareness programs as well for parents to make them aware about the conditions and sensitize them towards the same. Along with academics there should be experiential learning for children to enhance their sensory developments in the institutional setups.

## 3 Methodology

Personal Interview based study with phenomenological approach was conducted with 25 participants; special educators, psychologists, parents of children with special needs and teachers. All participants belong to Delhi NCR, India. The interaction was conducted through online mode and in-person interview. The interaction with each participant was conducted separately. The special educators had a varied experience in teaching children, psychologists had been dealing with children as well as educating the parents regarding the problems faced by children as well helping them understand the coping mechanisms that could work well with the regulating the behavioral challenges.

A semi structured questionnaire-based interaction was conducted with the individuals to understand the challenges, coping mechanisms as well day to day interactions with the children. The study conducted was qualitative and was framed in order to go deep into the aspects which need attention. It also helped in understanding the scope of research further.

In order to understand the gaps found through literature, it was important to have conversations and discussions for the understanding of the challenges faced by the stake holders. The stake holders were requested to share their experiences, opinions, methods modified/ adopted, their daily observations that impact the child in one manner or the other.

Content analysis was carried out with the primary data that was collected through interviews with the stake holders. This has been mentioned in the next section and table 2.0 and 3.0

**Table 1.0:** Tabular representation of research methodology

| Type of Research | Qualitative Research |
|---|---|
| Method of Data Collection | In-depth interviews (online and in-person) |
| Tool of Data Collection | Semi-structured questionnaire |
| Target Respondents | Psychologists, Special Educators, parents and teachers in Delhi-NCR |
| Sampling Technique | Convenience and Snowball Sampling |

| Sample Size | 25 (10 psychologists; 5 parents; 5 teachers; 5 special educators) |
|---|---|

## 4 Finding and Analysis

### 4.1 Major themes that were generated out of interviews with parents

1. Discomfort with labeling – The parents were uncomfortable with the labels provided by the society. With the societal settings like family, friend and other social gatherings they are often considered different, separate and treated not normally. Few members of the society are considerate; However, majority is not congenial.
2. Financial implications – The people taking their children for extra support face financial implications. The education system in today's time is already considered expensive and extra support creates drainage too. There are many stake holders who withdraw their children due to this reason.
3. Schedule building – Majority of parents feel that the child needs a schedule in order to sustain. The schedule should include physical activity as well as sufficient time for them to revise their educational curriculum. Extra-curricular activities sometimes could take a back seat due to the busy packed schedules of children with special learning needs.
4. Acceptance – The stake holders look out to the society in order to feel accomplished and sufficiently carrying out tasks for their child. The acceptance of society still holds utmost importance for the stake holders.

### 4.2 Major themes that were generated out of interviews with psychologists

1. Support system building – The psychologists work in collaboration with teachers, learners and parents working on bridging the gap of information and creating a path of congenial learning environment.
2. Interventions and modifications – The learners are assessed and personalized plan for them can be devised understanding the gaps.
3. Timely Interactions and conversations – Time are one of the essential components when it comes to providing support. If the conversations yield positive responses, they can be converted into positive results which are visible in the learning graph of the child.
4. Formal assessments – Assessments done with the formal tools are to understand the clinical psychological developments of the child according to their age to provide them with required support.

### 4.3 Major themes that were generated out of interviews with special educators

1. Behavior Modifications – The special educators are trained professions who work with the children on one-to-one basis, regularizing their emotions in different situations, ultimately working through a positive change in their behavior.

2. Personalized attention – The personal attention given to the child makes the areas that need support clearer, it helps in devising provisions required for the child
3. Support system building – The special educators work in collaboration with psychologists, learners and parents. They work with children on individual basis and bridge the gap of congenial learning environment.
4. Creating congenial learning environment – The environment that provides a sense of safety, joy and learning is considered as a congenial environment. This can only be created when one understands the proper requirements and considers inclusivity.

Table 2.0: Tabular representation of the theme identification through interviews conducted

| S. No. | PARENTS | PSYCHOLOGISTS | SPECIAL EDUCATORS |
|---|---|---|---|
| 1 | Discomfort with labeling | Support system building | Behavior Modifications |
| 2 | Financial implications | Interventions and modifications | Personalized attention |
| 3 | Schedule building | Timely Interactions and conversations | Support system building |
| 4 | Acceptance | Formal assessments | Creating congenial learning environment |

Based on the data analyzed earlier, the research formulated relationships to describe the study's major attributes. Table 3.0 gives a detailed review of these qualitative correlations.

Table 3.0: Tabular representation of the relationship between various attributes of the study

| S. No | Attribute 1 | Attribute 2 | Impact (Positive/ Negative/ Both) |
|---|---|---|---|
| 1 | Stress | Behavior | Negative |
| 2 | Finance | Facilities | Positive |
| 3 | Time | Comfort | Positive |
| 4 | Behavior | Attention | Both |
| 5 | Acceptance | Situation | Both |
| 6 | Labels | Acceptance | Positive |
| 7 | Interventions | Learning | Positive |
| 8 | Support | Learning | Positive |
| 9 | Restrictions | Learning | Negative |
| 10 | Environment | Holistic growth | Positive |

Qualitative correlation 1 – Stress influences behavior in a negative manner when the parent/ teacher loses their temper on the children due to their adamant behavior, which further leads to breakdowns and tantrums.

Qualitative correlation 2 – Finance influences facilities provided in a positive manner when the stake holders are able to get the child with the extra learning assistance apart from the basic education and help the schools provide.

Qualitative correlation 3 – Time influences comfort received in a positive manner when the increased availability with the parents and teachers influences the behavior due to the love, attention and support provided is more.

Qualitative correlation 4 – Behavior influences attention in a positive manner when the energy channelized yields results where the child become congenial and attentive to their surroundings but becomes negative when the child has increased tantrums due to pampering.

Qualitative correlation 5 – Acceptance influences situations in a positive manner when there is increased acceptance, the child feels more comfortable in their environment and learning is improved where as it has a negative impact when the parent shields the condition rather than taking steps for improvement.

Qualitative correlation 6 – Labels influence acceptance in a positive manner where the interventions could be proper and help can be received. Less acceptance sometimes delays the process of learning and improvement.

Qualitative correlation 7 – Interventions influences learning in a positive manner when an early identification is done. It helps and support given to the children at the initial stages results in a positive change in the long run.

Qualitative correlation 8 – Support influences learning in a positive manner where the child feels safe in the system, receiving the required provisions to improve their skills.

Qualitative correlation 9 – Restrictions influence learning in a negative manner as it leads to less exposure and reduced interventions. Restrictions hold the child back from exploring their skills and learning through other mediums.

Qualitative correlation 10 – Environment influences holistic growth in a positive manner where the child gets all the required support and exposure to enhance their skills. When the environment is congenial it helps the child to be more comfortable in opening up about their learnings.

## 5 Discussion

The purpose of the study was to understand the challenges and coping mechanisms the stake holders face in relation to children with special learning needs. The stake holders include teachers, psychologists, special educators and parents. It is essential to create an inclusive environment for children to excel in their lives well irrespective of the methodology and medium of learning. This environment needs support and collaborative efforts of the society as a whole. The interviews conducted helped get insights about the situations and perspectives. The findings indicate the importance of the support system around the child. The parents along with special educators can create individualized learning plans and methodologies for the child, the psychologists can guide the parents to work through the behavioral difficulties. The results suggests that the labels given to the children make the parents uncomfortable, addressing the difficulties faced by the child in public is still a stigma. Beyond that the financial implications that parents face seems to be a challenge in long term as well as for the people in the society who are deprived of the facilities because of the same. The already existing knowledge gave overview of the problems as well as raised a need to come up with design solutions to make learning fun, along with the secondary research and the primary research conducted has given a few design direction which will be further worked upon.

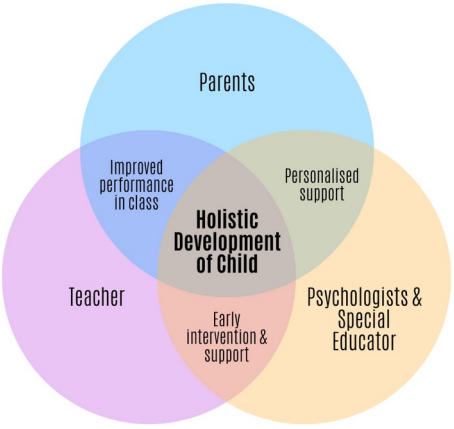



Fig 1.0: Diagrammatically representing the understanding the findings of the given research

The Figure is a representation of how each one influences the learning growth of the child with special learning needs. The interventions at each end will be beneficial at various stages.

## 6 Contribution of the paper

The paper contributes to the insights that need more attention. The semi structured questionnaire helped probing into the factors that contribute to the development of children with special needs. It raises questions on the current trends and understanding of society about the special requirements. Every child is unique but the system being followed still expects them to be confined to the structured and standardized norms of the society. All the stake holders contribute in the development of the child holistically. The paper aims to address the importance of the same and increase awareness about the help that can be provided by the society at large.

## 7 Scope of further interventions

### 7.1 Recommendations for Designers

The designers can work in collaboration with the stake holders that have a major role to play in the field. The designers can support by developing material that works on the universal inclusivity of learners. There can be mobile applications, interactive books, improvised learning games, toys, animations and much more. The interventions already existing can be modified keep in mind the visual elements that work in favor of the learners.

### 7.2 Recommendation for parents/ teachers

The parents and teachers can foster support by being regular with their interactions with the children. Supporting them by words of encouragement, patience and constant reassurance. They can establish routines that include a mix of academic learning, taking them to public places, enrolling them in a physical activity. The routine suggested should include working through multisensory activities and learning programs to keep them engaged and far from getting bored easily.

### 7.3 Recommendation for psychologists/ special educators

The psychologists and special educators play a very important role as they can help the child in providing emotional and psychological support when they are away from their homes. They have to collaborate with other stake holders like family members

and friends which gives them the power to improvise learning strategies and provide personalized support and focus on long term outcomes.

## 8 Limitations of the research

The research conducted had a few limitations that could be further given importance in order to improvise the results and include more population to the study. The stake holders approached were limited in number, convenient sampling was employed. The data collected was qualitative, varied and more concerned with behavioral aspects. The parents were asked limited questions pertaining to the child, which need more attention and detail in order to probe more into the scenarios. Further the research could also be taken in the direction of quantitative analysis for an in-depth study.

**Acknowledgments.** Authors would like to extend gratitude to all the participants for their contribution to this research.

**Disclosure of Interests.** The authors have no competing interests. This research is a part of the PhD research work.

References

1. Committee on the Science of Children Birth to Age 8: Deepening and Broadening the Foundation for Success; Board on Children, Youth, and Families; Institute of Medicine; National Research Council; Allen LR, Kelly BB, editors. Transforming the Workforce for Children Birth Through Age 8: A Unifying Foundation. Washington (DC): National Academies Press (US); 2015 Jul 23. 4, Child Development and Early Learning. Available from: https://www.ncbi.nlm.nih.gov/books/NBK310550/
2. Parker R, Thomsen BS and Berry A (2022) Learning Through Play at School – A Framework for Policy and Practice. Front. Educ. 7:751801. doi: 10.3389/feduc.2022.751801
3. Foundation for Young Australians [FYA] (2017). The New Work Order. Ensuring young Australians have skills and experience for the jobs of the future, not the past. Melbourne: FYA.
4. Darling-Hammond, L. (2006). Constructing 21st-century teacher education. J. Teach. Educ. 57, 300–314. doi: 10.1177/0022487105285962
5. DeLuca, C., and Hughes, S. (2014). Assessment in early primary education: an empirical study of five school contexts. J. Res. Child. Educ. 28, 441–460. doi: 10.1080/02568543.2014.944722
6. Karlsson Lohmander, M. (2022). Quality early childhood education for all children? European Early Childhood Education Research Journal, 30(5), 653–657. https://doi.org/10.1080/1350293X.2022.2109857
7. Moss, P., G. Dahlberg, S. Grieshaber, S. Mantovani, H. May, A. Pence, and M. Vandenbroeck. 2016. "The Organization for Economic co-Operation and Development's International Early Learning Study: Opening for Debate and Contestation." Contemporary Issues in Early Childhood 17(open in a new window): 343–351. doi:10.1177/1463949116661126.

8. Gelman SA. Learning from others: children's construction of concepts. Annu Rev Psychol. 2009;60:115-40. doi: 10.1146/annurev.psych.59.103006.093659. PMID: 18631027; PMCID: PMC2829654.
9. Reiss AL. Childhood developmental disorders: an academic and clinical convergence point for psychiatry, neurology, psychology and pediatrics. J Child Psychol Psychiatry. 2009 Jan;50(1-2):87-98. doi: 10.1111/j.1469-7610.2008.02046.x. PMID: 19220592; PMCID: PMC5756732.
10. Mohammed Elmadani, Simon Klara, Evans Kasmai Kiptulon, Diego Andrade, Mate Orsolya, A systematic review of educational interventions to enhance cancer awareness and screening in individuals with intellectual disabilities: A mixed-methods approach, Research in Developmental Disabilities, Volume 155, 2024, 104867, ISSN 0891-4222, https://doi.org/10.1016/j.ridd.2024.104867.
11. Moreau D, Waldie KE. Developmental Learning Disorders: From Generic Interventions to Individualized Remediation. Front Psychol. 2016 Jan 12;6:2053. doi: 10.3389/fpsyg.2015.02053. PMID: 26793160; PMCID: PMC4709759.
12. John, Philip. Learning and other developmental disorders in India. Indian Journal of Psychiatry 52(Suppl1):p S224-S228, January 2010. | DOI: 10.4103/0019-5545.69236
13. Koirkana, Appaji. (2020). "SLOW LEARNERS- A UNIVERSAL PROBLEM AND PROVIDING EDUCATIONAL OPPORTUNITIES TO THEM TO BE A SUCCESSFUL LEARNER". PEOPLE: International Journal of Social Sciences. 6. 29-42. 10.20319/pijss.2020.61.2942.
14. A Vasudevan. Slow learners – Causes, problems and educational programmes. Int J Appl Res 2017;3(12):308-313.
15. Dibia, N., & Ajoku, M. (2018). TEACHERS' ROLE IN IDENTIFICATION AND SUPPORT FOR SLOW LEARNERS AT THE LOWER BASIC EDUCATION LEVELS: IMPLICATIONS FOR COUNSELLING. European Journal of Education Studies, 0. doi:http://dx.doi.org/10.46827/ejes.v0i0.1527
16. Pandey, S., & Kurian, B. J. (2016). Role of Teacher's: Helping Slow Learners to Bring out Hidden Skills. In IOSR Journal of Research & Method in Education, IOSR Journal of Research & Method in Education (Ver. VII, Vol. 6, Issue 6, pp. 23–26) [Journal-article]. https://doi.org/10.9790/7388-0606072326
17. Bowlby, J. 1969. Attachment and Loss. Attachment, New York: Basic Books.
18. Ana Garcia-Melgar, Nerida Hyett, Kerryn Bagley, Carol McKinstry, Jo Spong, Teresa Iacono, Collaborative team approaches to supporting inclusion of children with disability in mainstream schools: A co-design study, Research in Developmental Disabilities, Volume 126, 2022, 104233, ISSN 0891- 4222, https://doi.org/10.1016/j.ridd.2022.104233.m(https://www.sciencedirect.com/science/article/pii/S0891422222000634)
19. Hernandez, S. J. & Hofstra University. (2013). Collaboration in Special Education: Its History, Evolution, and Critical Factors Necessary for Successful Implementation. *US-China Education Review B*, *3*(6), 480–498. https://files.eric.ed.gov/fulltext/ED544122.pdf
20. Yell, M., & Bateman, D. (2024). The Critical Nature of Collaboration in Special Education: An Analysis of the IDEA. *Journal of Educational and Psychological Consultation*, 1–18. https://doi.org/10.1080/10474412.2024.2391329
21. T, P. (2017). Role of Parents of Special Needs Children: A Qualitative Study. In Dr.Shailaja Shastri, *Jain University, Bangalore: Vol. Volume 9, Issue 5, 2019* [Journal-article]. https://www.pramanaresearch.org/gallery/prj-p889.pdf
22. Mann, G., Kennedy-Wood, L., & Gilmore, L. (2024). The role of parents in facilitating safe, quality, inclusive education for students with intellectual disabilities: the position of the Disability Royal Commission. *Research and Practice in Intellectual and*

*Developmental Disabilities*, *11*(1), 107–120. https://doi.org/10.1080/23297018.2024.2337756

23. Issabayeva, Z., Sapargaliyeva, A., Shubayeva, G., Shalabayeva, K., & Ismagambetova, G. (2024). Psychological and Pedagogical Support for Parents of Children with Special Educational Needs. *Continuity in Education, 5(1),* 66–75. https://doi.org/10.5334/cie.103
24. Hanurawan, F. & Universitas Negeri Malang. (2017). The Role of Psychology in Special Needs Education. In *Journal of ICSAR* (Vol. 1, Issue 2) [Journal-article].
25. Zhang, L. (2018). Study on Children Product design and Development based on fashion consumption. In *7th International Conference on Social Science and Education Research (SSER 2017)* (Vol. 132, p. 194) [Conference-proceeding].
26. Hansen, A. S. & Department of Design. (2021). *Co-Design with Children* [Journal-article].https://www.ntnu.edu/documents/139799/1279149990/13+Article+Final_anjash_fors%C3%B8k_2017-12-07-20-11-11_Co-Design+with+Children+-+Final.pdf/b8dd19c4-d2b1-4322-a042-718e06663e13
27. Fong, F. T., Puebla, G., & Nielsen, M. (2024). The role of conventionality and design in children's function judgments about malfunctioning artifacts. *Journal of Experimental Child Psychology*, *240*, 105835. https://doi.org/10.1016/j.jecp.2023.105835
28. Schepers, S., Schoffelen, J., Zaman, B., Dreessen, K., Social Spaces, LUCA School of Arts / KU Leuven, Social Spaces, LUCA School of Arts / KU Leuven, & Meaningful Interactions Lab (Mintlab), KU Leuven. (2019). Children's roles in Participatory Design processes: making the role of process designer 'work.' In *Interaction Design and Architecture(S) Journal: Vol. N.41* (pp. 87–108) [Journal-article].

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial 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 license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license 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.



ARTHAVAAN
Volume 7, Issue 1, December 2024
© Bharati College
10.71322/arthavaan.vol.7.issue.1.01

# Empirical Investigation into the Interlinkages Between the Price Movements of Energy Commodities and Stock Market in India

**Dr. Nishi Sharma* and Tarandeep Kaur****

***Abstract***

*The present study is an attempt to empirically inquire into the behaviour of and integration between energy commodity prices and the benchmark index of the Indian securities market. For this, two prime energy commodities are considered, i.e., crude oil and natural gas and the Nifty 500, the broader market index of India's National Stock Exchange (NSE). The inquiry is based on data for the daily spot price (FUTCOM) of crude oils and natural gas in the evening session from January 1, 2017, to September 30, 2024, from the Multi Commodity Exchange of India and the daily close price of Nifty 500 index from the National Stock Exchange of India. Firstly, the series are tested for stationarity and transformed in the log return form. Next, a preliminary analysis is performed using graphical and descriptive methods. Then, the times series approach of the Vector Autoregressive (VAR) model, the Granger causality and the Johansen cointegration are applied to investigate the inter-relation among the three variables in the short and long run. The variables are found to be sensitive to external effects of the Covid-19 pandemic. Results establish bidirectional causality between the crude oil prices and the Nifty 500 index. Long-term equilibrium relationships among all three variables are discovered using the cointegration test. The study concludes that the connectedness among the trio variables is time-varying. The results have practical implications for the benefit of investors, traders, arbitrageurs, and managers in hedge and mutual funds, as they can consider the findings while designing and reframing risk management and hedging strategies. Similarly, the empirical results can benefit policymakers when drafting and designing policies governing the commodities and equity markets.*

***Keywords****: energy commodity, crude oil, natural gas, cointegration, stock market, causality*

## 1. Introduction

Crude oil and natural gas are the two prominent sources of energy that serve as a pivotal factor in an economy's development. Globally, India is the third largest oil-consuming country after the USA and China (Terra Nova, T. 2024). Further, it is experiencing a tremendous growth in consumption of energy from 26,822 Petajoules (PJ) in 2013-14 to 35,159 Petajoules (PJ) in 2022-23 (Provisional). Moreover, 6.48% of total energy consumption increased from 2021-22 to 2022-23 (Ministry of Statistics and Program Implementation, 2024). Natural gas consumption in India is also rapidly growing owing to population growth and economic development, reflecting a significant shift towards cleaner energy. India relies heavily on imports to fulfil its constantly growing oil and natural gas demand. India imports approximately 83% of its crude oil and 47% of its natural gas. Further, about 30% of India's total energy consumption is met by crude oil. The price of these two commodities is driven by the demand, supply and geopolitical factors that significantly impact other

**Associate Professor, Department of Commerce, Dr. Bhim Rao Ambedkar College, University of Delhi, Delhi. (also, the corresponding author) Email ID: nishidandeva@gmail.com*
***M.A. economics, Delhi Technological University, Delhi. Email ID: tarandeepk2002@gmail.com*

sectors of the economy. Crude oil and natural gas prices are subject to more significant uncertainty that might result in low investment, less production and reduced aggregate output. It affects the performance of various industries and sectors, ultimately affecting the stock market performance, a barometer of the economy. Thus, the volatility in oil and natural gas prices significantly impacts economic and financial structures and influences stock market returns. Hence, it is important to identify the underlying dynamics between the energy commodity and equity markets.

Many studies have shown that volatility in the prices of these commodities affects the financial market. Many traders dealing in the commodity markets shift to stock markets to hedge investment risks and optimize returns. There is growing interest among researchers in understanding the contagion and interconnection between the commodity and stock markets driven by increasing uncertainty in both sectors and the significance of these markets for portfolio diversification and hedging risk (Mensi W. et al., 2022).

Extensive literature is available on assessing the inter-relationship between the commodities and stock market as aggregate and at sectoral levels in developed and emerging economies, with a few studies covering Indian markets. We were motivated to identify how the securities market in India responds to price shocks, particularly in the two leading energy commodities, crude oil and natural gas, which play an important role in the development of the economy.

This study aims to assess the overall performance of the equity market and the energy commodity market. It will also examine the cointegration and causal effect between the price movements of two prime energy commodities, crude oil and natural gas, and the stock market, focusing on India.

The significance of this research lies in providing informed knowledge on the performance and interlinkages between the energy and equity markets in India to investors, portfolio managers, and policymakers with empirical evidence using Vector Autoregressive (VAR) model, cointegration, and causal methodologies on daily data spanning from January 2017 to September 2024, covering more than seven years that includes the periods before, during, and after COVID-19.

The contribution of our study can be summarised as follows. Initially, we provide the performance evaluation of the stock market by considering a broader benchmark index, i.e., the Nifty 500 index and the energy commodity market represented by crude oil and natural gas based on risk and return parameters. Then, considering the significance of the trio of crude oil, natural gas and equity investments for short-term (speculators and arbitrageurs) as well as long-term (mutual and hedge funds) investors, we build the cointegrating and causality models helpful in framing mixed assets portfolio and effectively hedging the risk in the financial market.

The structure of the paper is as follows: Section 2 provides a review of the relevant literature. Section 3 outlines the data used in the study. Section 4 details the research methodology. Section 5 presents the results along with a discussion. Lastly, Section 6 concludes the study, offering policy recommendations and suggesting directions for future research, followed by references and appendix section.

## 2. Literature Review

Many studies that focused on similar variables using various statistical techniques are reviewed to get the knowledge of existing literature of the interconnectedness of commodity and stock markets across nations. Anna Creti et al. (2013) identified that the correlations between the commodity and stock markets are highly volatile, especially from the financial crisis of 2007-08 onwards, which led to the growing financialization of commodity markets. They investigated the correlation between the price returns of 25 commodities and the stock market using GARCH methodology involving the study period from January 2001 to November 2011. Similar results were obtained by Marco J. L. et al. (2016),

who examined the correlation between the returns from commodity and equity and its implications for asset allocation using the time-varying Bayesian Dynamic Conditional Correlation Model. They identified higher volatility between the two variables and concluded that joint modelling of equity and commodity prices yields accurate forecasts. Syzdykova, A. & Azretbergenova, G. (2024) examined the asymmetric effect of oil prices on exchange rates and stock market index returns in Kazakhstan by applying nonlinear ARDL cointegration technique over monthly data from January 2010 to February 2024. It was found that changes in oil prices, whether positive or negative, significantly affected the Kazakhstan stock exchange (KASE) returns and the exchange rates.

Numerous recent studies have examined spillover effects between the various sectors and markets. One such study by Elsayed, A.H. et al. (2020) analyzed the volatility spillover mechanism between the energy markets and stock prices of seven major global securities markets during 2000-2018 using time domain interlinkages techniques of Diebold and Yilmaz. They identified an insignificant contribution of volatility in the energy sector to the global securities market. In the clean energy market, the major transmitters of volatility were the returns from the World Stock Index and the World Energy Index. Further, their study identified a strong impact of the energy market on the global financial markets. Maghyereh, Aktham I., et al. (2016) explored the oil and equities directional connectedness using the implied volatility indexes, taking samples from the eleven global stock exchanges from 2008-2015. Their research established the bi-directional information spillovers between the two markets across the studied countries. However, the transmission from oil to equity markets significantly contributed to the linkages between the two.

To understand the impact of COVID-19, Lucey, Brian and Ren, Boru (2023) examined dynamic connectedness between the sustainability indices, equities and energy assets employing the CAViaR-TVP-VAR technique on daily data about the period October 14, 2014 – August 14, 2022. They observed that total risk connectedness was at a medium level. COVID-19 also had a mild effect in the short run. Another research by Hanif W. et al. (2023) evaluated the time-frequency dependence and risk connectivity between the green stocks and oil shocks by applying wavelet coherence and frequency connectedness analysis. Their study discovered tighter dependency relationships between the two variables at mid and long-term scales. Lead lag patterns were mixed and time-varying. They further noticed a substantially significant risk spillover from the oil market in the green stock market. They also concluded that various global crises, viz., the Great Recession, the oil price collapse and the COVID-19 pandemic, significantly enhanced the magnitude of risk spillovers.

Mensi W. et al. (2022) examined the interlinkages between the sectoral markets and crude oil and gold in the USA to comprehend the impact of oil prices on various sectoral markets. They also studied the implications of their connectedness on portfolio management using Diebold and Yilmaz (2012) and Barunik and Krehlik's approaches. They discovered that gold, oil, financials, utilities, communications services and health care were the net recipients of the spillover effects, and the rest of the sectors were the net contributors irrespective of the frequencies. Furthermore, a reduced spillover effect was observed when gold and oil were included in the sector portfolio.

In India, Shahani, R. et al. (2023) investigated the cointegrating behaviour of crude oil and natural gas in the Indian commodity market using daily returns from April 1, 2017, to March 31, 2022. Based on ARDL (with structural break) and nonlinear ARDL methods, their study evidenced the long-run cointegration for natural gas only and not for crude oil. Further, they found that natural gas, concerning crude in the short run, was inelastic while highly elastic in the long run. There was no asymmetric impact of crude on natural gas in the short run, whereas, in the long run, it was found only at a 10% significant level. Furthermore, the negative and

significant Error Correction term (using VECM) for natural gas reflected the stable movement from short-run disequilibrium to long-run equilibrium, though the speed of adjustment was 3 percent per period. Another study by Bhullar Pritpal et al. (2024) investigated the volatility spillover between the energy commodities, viz. oil and natural gas and the stock indices, including India, USA and Japan, during 2001-2023 using DCC-GARCH models. They found strong interdependence between the crude oil and all the stock indices under study compared to natural gas.

Thus, body of literature has discussed the association between the various commodities and equity markets by applying numerous techniques, viz. correlation, GARCH model, wavelet analysis, and so on, across developed and emerging economies, including India. However, scant literature is available in the Indian context. Thus, there is scope of research to assess the commodity and equity market nexus. This study is one-step addition to the empirical evidence on it in the Indian context.

## 3. Data

The data for this study is sourced from the Multi Commodity Exchange (MCX) India and the National Stock Exchange (NSE) India databases. From the NSE, daily closing prices for the Nifty 500 Index are obtained. From the MCX, spot market prices for crude oil (per barrel) and natural gas (per Metric Million British Thermal Unit, mmBtu) are used. All values are reported in Indian Rupees (INR). The analysis period spans from January 2017 to September 2024, using daily price data for all series, though data is only available for business days, excluding weekends and national holidays. It results in three-time series with irregular intervals.

## 4. Methodology

The Time series data is analyzed using Graphical, Descriptive and econometric methods. Time plots are made to represent data graphically in order to understand the trends over the period. The original data series were checked for stationarity as a pre-condition for time series analysis. Stationarity implies that the mean and variance remain constant over time. Additionally, the covariance between two time periods depends solely on the interval between them, rather than the specific times at which it is calculated (Enders, 2004). The original series were transformed into log return series to make the data stationary and the analysis more meaningful. Log returns possess multiple advantages over simple returns. They can be easily summed over multiple periods, which allows for straightforward calculations of cumulative returns across different time frames. The distribution tends to be more symmetrically distributed, providing advantages for statistical analysis. The descriptive statistics are analyzed for absolute values and log return values.

Stationarity testing is performed using the Augmented Dickey-Fuller (ADF) test. The ADF test is a unit root test whose null hypothesis is the presence of a unit root, which is non-stationarity.

A Vector Autoregression (VAR) model is employed to capture the dynamic interactions among multiple time series variables. Unlike univariate models that analyze a single time series, the VAR model examines the interdependencies by treating all variables in the system as endogenous (dependent on each other) (Enders, 2004). Further, the causal effects are checked using the Granger Causality test. Following that, the long-run relationship is checked using the Johansen Cointegration test.

The analysis is performed using the latest version of the statistical software R. R is an open-source program for statistical analysis and visualization.

## 5. Results

Figure 1 below shows the line plots of the series. Panel A shows the Nifty 500 closing prices; the series was stable below Rs. 10,000 before 2020.

[1] *A barrel (BBL) equals 42 US gallons or 159 litres.*
[2] *A Metric Million British Thermal Unit (mmBtu) is commonly used to measure energy value, particularly for natural gas.*

After that, it witnessed a sharp decline in 2020, attributed to the COVID-19 pandemic. The series started rising in value and stabilized for 2022. In 2023, the series started rising till 2024 and reached above Rs. 24,000. The series showcases the significant growth of the Indian stock market during the period under consideration, thereby showcasing the robust growth of the economy.

Panel B shows the line plot of crude oil spot price series. The series depicts a fluctuating trend, which is rising overall. The series also dropped significantly in 2020, around the time lockdowns were announced in the nation as a preventive measure for the pandemic. Till 2022, the series increased rapidly above Rs. 8000, thereafter noting a decline coming down to Rs. 6000.

Panel C shows the natural gas spot price series. The series was stable with minor fluctuations until the middle of 2021, after which it observed peaks. The series stabilized at its original level after 2022.

**Figure 1: Line Plots of actual series**

**Panel a) Nifty 500 Closing Price (Rs.)**




*Source: Authors' computation*

**Panel b) Crude Oil Spot Price (Rs.)**



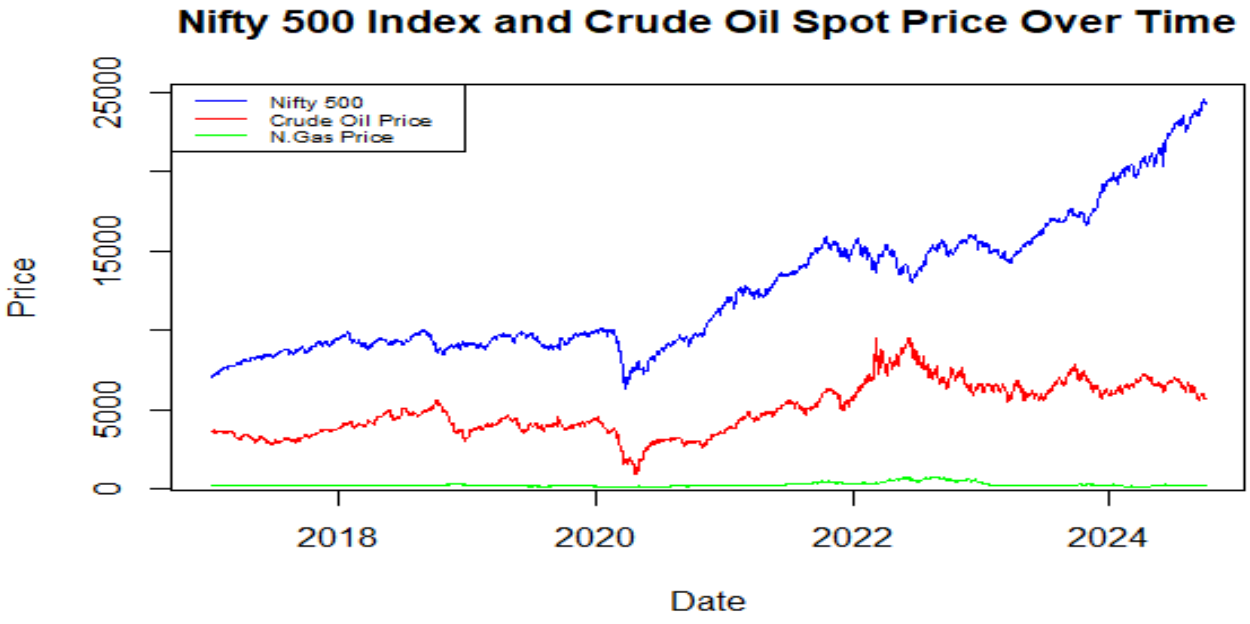

*Source: Authors' computation*

**Panel c) Natural Gas Spot Price (Rs.)**




*Source: Authors' computation*

**Figure 2: Line plots of all series**




*Source: Authors' own computation*

Figure 2 provides us with a comparative view of the three series. It displays that the Benchmark Index, Nifty 500, holds the highest value, followed by Crude oil price in the spot market, and natural gas price takes the lowest value. Moreover, it can be observed that the Nifty 500 and crude oil prices followed similar trends until 2022. However, post 2022, the Nifty 500 value has increased, whereas Crude oil prices have decreased slightly, indicating the growth of the stock market.

### 5.1 Checking for Stationarity

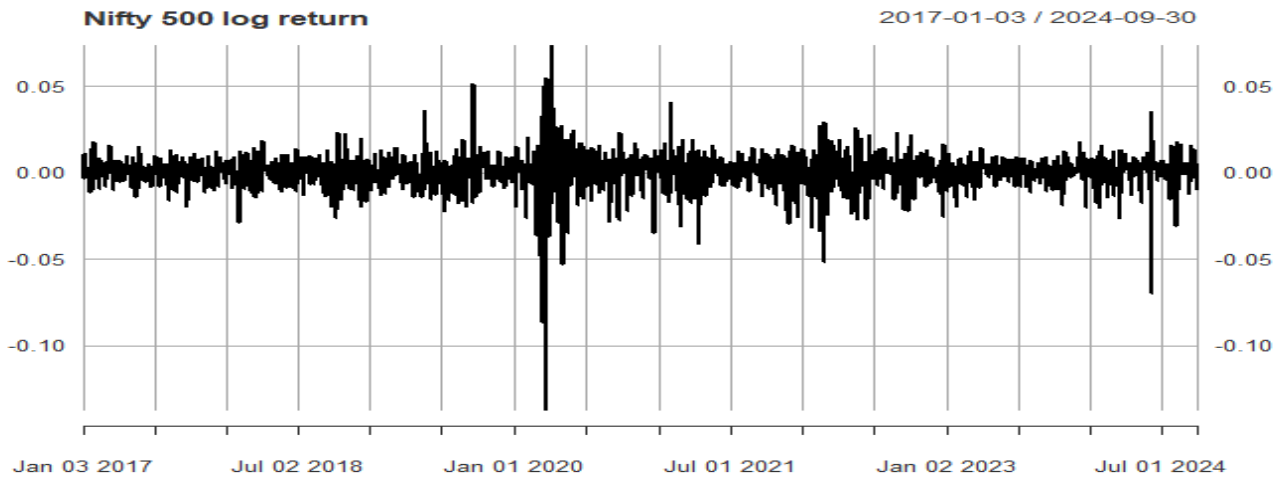
After plotting the series, it was observed that they

showed trends that indicated non-stationarity, as it did not revert back to the mean. It is tested for using the formal testing procedure of the unit root test. It is one of the commonly used methods for assessing stationarity. The presence of a unit root signifies that the data series is non-stationary. The stationarity of the series was evaluated using the formal ADF test, which confirmed that the series was non-stationary.

The series was then transformed into a log return form to perform the analysis further. The log return series were found to be stationary, reverting to the mean (see Figure 2). The ADF test also confirms this (Table A2).

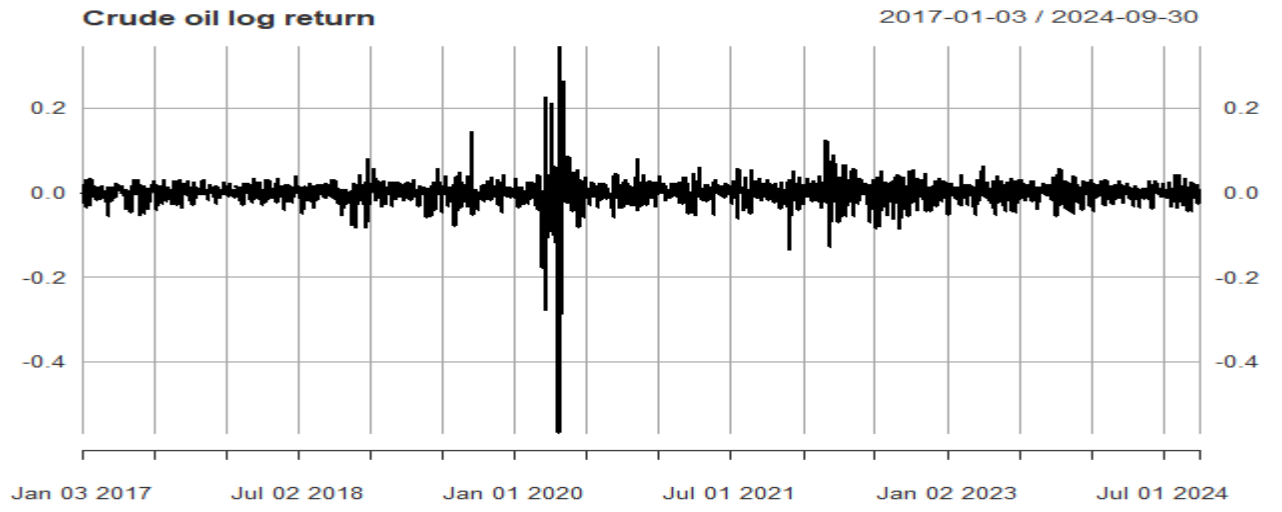
**Figure 3: Line Plots of Log returns**

**Panel a) Nifty 500 Index**



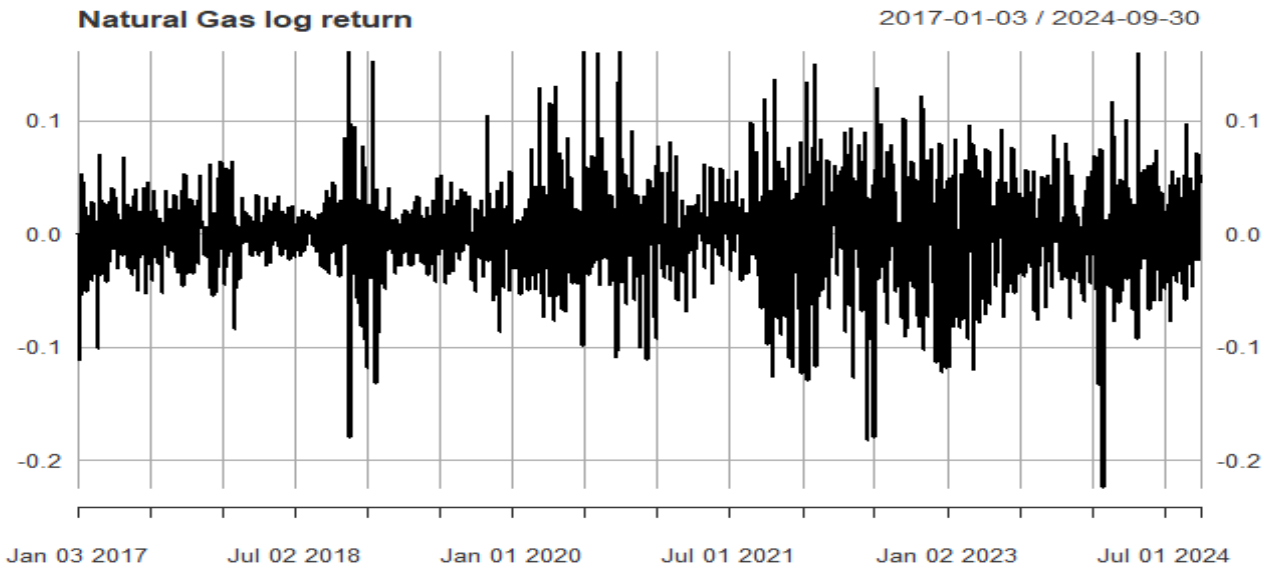

*Source: Authors' own computation*

**Panel b) Crude Oil Spot Price**




*Source: Authors' computation*

**Panel c) Natural Gas Spot Price**




*Source: Authors' computation*

## 5.2 Descriptive Statistics

Descriptive statistics are discussed for both the actual and the series with log differencing transformations. The actual series is important as it depicts the overall trends in the variables during the time period. Log return series are also discussed as analysis is based on them.

Table 1 showcases the descriptive statistics of the actual series. The actual series, Nifty 500, shows a mean of Rs. 12,572 with a standard deviation of Rs. 4,237. The minimum value of the series is Rs. 6,243, and the maximum value is Rs. 24,497. Series is slightly positively skewed with a skewness of 0.84. The kurtosis of the distribution is -0.18.

The crude oil price series shows a mean of Rs. 4,964, with a standard deviation of Rs. 1,617. The minimum value is Rs. 887, and the maximum value is Rs. 9,510. Series is slightly positively skewed, with a skewness of 0.31. The kurtosis of the distribution is -0.65.

The natural gas spot price averaged Rs. 244, with a standard deviation of Rs. 124. The series' minimum value is Rs 111, and its maximum value is Rs. 773. Series

is highly positively skewed, with a skewness of 2.14. The distribution is peaked, with a high kurtosis of 4.16.

**Table 1: Descriptive Statistics of Actual values of series**

| Descriptive Statistics of Absolute values | | | |
|---|---|---|---|
| | **Nifty 500** | **Crude Oil Price** | **Natural Gas Price** |
| **Source** | NSE India | MCX India | MCX India |
| **Unit** | Rs./ INR | Rs. /INR | Rs./INR |
| **Mean** | 12572 | 4964 | 244.4 |
| **SD** | 4236.87 | 1616.67 | 124.44 |
| **Min** | 6243 | 887 | 111.2 |
| **Max** | 24497 | 9510 | 773.1 |
| **Range** | 18253.9 | 8623 | 661.9 |
| **1st Quartile** | 9151 | 3688 | 179.7 |
| **Median** | 10697 | 4664 | 200 |
| **3rd Quartile** | 15280 | 6324 | 245.5 |
| **Skewness** | 0.84 | 0.31 | 2.14 |
| **Kurtosis** | -0.18 | -0.65 | 4.16 |
| **Observations** | 1933 | 1933 | 1933 |

*Source: Authors' computation*

Table 2 presents the descriptive statistics for the log return values of the series. The average return of 0.00064 is provided by the Nifty 500 series, which is the highest. Crude oil has shown an average return of 0.00023 over the period in consideration. Natural gas shows a negative return with a minimal magnitude. The risk, measured by the standard deviation, is also lowest for the Nifty 500, followed by crude oil. The highest risk is associated with Natural gas. Risk to return reward, measured by the Coefficient of Variation (CV), is also the lowest for the Nifty 500. These characteristics make the Indian stock market a better investment avenue than these commodities.

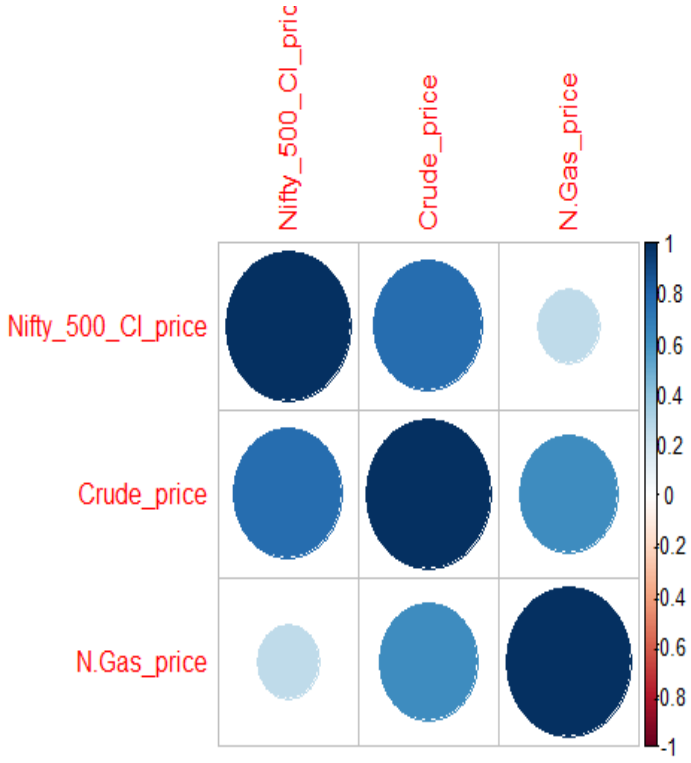
**Table 2: Descriptive Statistics of Log Return values of series**

| Descriptive Statistics of Log return values | | | |
|---|---|---|---|
| | **Nifty 500** | **Crude Oil Price** | **Natural Gas Price** |
| **Mean** | 0.00064 | 0.00023 | -0.00002 |
| **SD** | 0.01000 | 0.03000 | 0.04000 |
| **Coefficient of Variation** | 15.55694 | 129.87013 | -1860.46512 |
| **Range** | 0.21000 | 0.92000 | 0.39000 |

*Source: Authors' computation*

### 5.3 Correlation analysis

**Figure 4: Correlation plot**




*Source: Authors' computation*

The correlation plot reveals the degree and direction of association between variables. In our case, the series are positively associated. The Nifty 500 index and Crude oil have a 0.76 association. The natural gas price and Crude oil price have a 0.62 association. The Nifty 500 index and natural gas prices have a low association 0.26.

### 5.4 Econometric Analysis

R suggested models with different lag lengths based on four criteria. The four criteria choose the lags to minimize the function laid down. The Schwarz Criterion (SC) suggested the model with two lags, also known as the Bayesian Information Criterion (BIC).

The first model with 2 lags was selected to keep it simple and retain degrees of freedom. The model was then checked using a diagnostic check.

#### 5.4.1 Diagnostic Check

A stability check and autocorrelation test were performed as diagnostic checks. The VAR model with two lags, as Schwarz Criterion (SC) suggested, is stable but has autocorrelated residuals. Hence, another model with five lags was considered, as suggested by Hannan Quinn's (HQ) criterion. This model is also found to be stable and does not have the problem of autocorrelated residuals.

**Stability Check**

The stability test for VAR models is based on the eigenvalue stability condition. Checks roots of the characteristic polynomial to evaluate whether the system satisfies the stationarity condition. This test ensures that all roots of the characteristic polynomial are strictly within the unit circle in the complex plane

(i.e., their moduli are less than 1). A stable VAR model indicates that the dynamic relationships among the variables are well-behaved over time without exhibiting explosive or non-stationary behaviour.

Our model is found to be stable.

**Figure 5: Stability Test Result**

```
> stability_check2 <- roots(var_model2)
> print(abs(stability_check2))
 [1] 0.7356676 0.7356676 0.6642790 0.6061909 0.5786935 0.5786935 0.5756241
 [8] 0.5756241 0.5644191 0.5644191 0.5334809 0.5334809 0.5147053 0.4514878
[15] 0.4514878
> # Check if all roots are within the unit circle
> if (all(abs(stability_check2) < 1)) {
+   cat("The VAR model is stable.\n")
+ } else {
+   cat("The VAR model is unstable.\n")
+ }
The VAR model is stable.
```

*Source: Authors' computation*

**Residual Autocorrelation Test**

To assess the adequacy of the estimated Vector Autoregression (VAR) model, the residuals were tested for autocorrelation by using the Portmanteau test. This test evaluates whether the model's residuals exhibit white noise, a crucial assumption for the validity of VAR models.

The Portmanteau test assesses the null hypothesis that the residuals have no autocorrelation up to a specified lag, with the alternative hypothesis indicating the existence of autocorrelation. The test statistic is calculated as a function of the sum of squared residual autocorrelations and follows a chi-squared ($\chi$^2) distribution under the null hypothesis.

For this analysis, the test was conducted for up to 12 lags using the asymptotic version of the Portmanteau test. The test results include the chi-squared statistic, degrees of freedom, and the associated p-value. A p-value exceeding 0.05 suggests that the null hypothesis cannot be rejected, indicating that the residuals are uncorrelated and that the model is adequately specified. Conversely, a p-value of 0.05 or lower would indicate significant residual autocorrelation, necessitating a re-evaluation of the model specification.

The model with two lags is found to have auto-correlated residuals with a p-value of approximately zero. Models with higher numbers of lags, 5 (suggested by HQ criteria) and 7 (suggested by AIC and FPE criteria) lags, were also checked, but the autocorrelation problem remained. The reason for this might be that the number of variables under consideration is not enough to explain each other, as some macroeconomic and global factors and governance-related factors also affect our variables.

Hence, a model with 5 lags is considered to conduct the study. A drawback is that the results of the Granger causality test based on this model may be biased.

**Figure 5: Autocorrelation test: VAR model with five lags**

```
> print(autocorr_test)

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object var_model2
Chi-squared = 136.51, df = 63, p-value = 2.397e-07

$serial

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object var_model2
Chi-squared = 136.51, df = 63, p-value = 2.397e-07


>   cat("No significant autocorrelation in residuals (p-value =", autocorr_test$serial$p.value,
").\n")
No significant autocorrelation in residuals (p-value = 2.396967e-07 ).
```

*Source: Authors' computation*

### 5.4.2 VAR Model with five lags

The following is the table for the model with five lags. The model is found to be stable, but its residuals show autocorrelation. This model was suggested by the Hannan-Quinn (HQ) Criterion.

**Table 3: VAR Model estimation results**

```
Estimated coefficients for equation Crude_Return:
=================================================
Call:
Crude_Return = Nifty_500_Return.l1 + Crude_Return.l1 + N.Gas_Return.l1 + Nifty_500_Return.l2 +
Crude_Return.l2 + N.Gas_Return.l2 + Nifty_500_Return.l3 + Crude_Return.l3 + N.Gas_Return.l3 + N
ifty_500_Return.l4 + Crude_Return.l4 + N.Gas_Return.l4 + Nifty_500_Return.l5 + Crude_Return.l5
+ N.Gas_Return.l5 + const

Nifty_500_Return.l1     Crude_Return.l1     N.Gas_Return.l1 Nifty_500_Return.l2
       0.1157528607        0.0214911793        0.0005797115        0.0978987041
    Crude_Return.l2     N.Gas_Return.l2 Nifty_500_Return.l3     Crude_Return.l3
      -0.1167176477       -0.0059791624       -0.0780075786       -0.0495371358
    N.Gas_Return.l3 Nifty_500_Return.l4     Crude_Return.l4     N.Gas_Return.l4
      -0.0037455495       -0.1447788153        0.1210396784        0.0046219840
Nifty_500_Return.l5     Crude_Return.l5     N.Gas_Return.l5               const
       0.3104544454        0.0257427404       -0.0253615148        0.0005241040


Estimated coefficients for equation N.Gas_Return:
=================================================
Call:
N.Gas_Return = Nifty_500_Return.l1 + Crude_Return.l1 + N.Gas_Return.l1 + Nifty_500_Return.l2 +
Crude_Return.l2 + N.Gas_Return.l2 + Nifty_500_Return.l3 + Crude_Return.l3 + N.Gas_Return.l3 + N
ifty_500_Return.l4 + Crude_Return.l4 + N.Gas_Return.l4 + Nifty_500_Return.l5 + Crude_Return.l5
+ N.Gas_Return.l5 + const

Nifty_500_Return.l1     Crude_Return.l1     N.Gas_Return.l1 Nifty_500_Return.l2
       0.0723348098        0.0017029589       -0.0582488004        0.0123801692
    Crude_Return.l2     N.Gas_Return.l2 Nifty_500_Return.l3     Crude_Return.l3
       0.0742712367       -0.0322804062       -0.0835535682       -0.0369274026
    N.Gas_Return.l3 Nifty_500_Return.l4     Crude_Return.l4     N.Gas_Return.l4
       0.0481661082        0.0287579864       -0.0305494858        0.0160044131
Nifty_500_Return.l5     Crude_Return.l5     N.Gas_Return.l5               const
       0.0993599606       -0.1032775170       -0.0208872880        0.0008079393
```

*Source: Authors' computation*

The Vector Autoregression (VAR) analysis with five lags reveals complex dynamic relationships between the Nifty 500 index, crude oil, and natural gas returns. For the Nifty 500 returns equation, we observe that its own first lag has a negative coefficient (-1.706), while the second lag shows a positive effect (0.0374), suggesting short-term mean reversion in stock market returns. The impact of crude oil returns on the Nifty 500 is relatively modest, with the first lag showing a negative coefficient (-0.1283) and the second lag displaying a positive coefficient (2.28), indicating a volatile transmission mechanism from oil markets to stock returns. All subsequent variables have a positive effect; only lag 5 of crude and natural gas have a negative impact.

In the crude oil returns equation, we find significant feedback effects from the stock market, with the first leg of Nifty 500 returns showing a positive coefficient (0.1157) and the impact being stronger (0.31) at the $5^{th}$ lag. It suggests that stock market movements have a substantial bi-directional relationship with crude oil returns. The own-lag effects for crude oil are relatively small, with the first lag showing a positive coefficient (0.0214) and the second lag of crude oil lag showing a similar magnitude but a negative coefficient (-0.1167). Subsequent lags show diminishing effects. The impact of natural gas is comparatively smaller as its coefficients are small.

For natural gas returns, the results indicate moderate sensitivity to stock market movements, with the first leg of Nifty 500 returns showing a positive coefficient (0.0723). The own-lag effects for natural gas are notably negative for the first lag (-0.0582), suggesting mean-reverting behaviour in natural gas prices. The relationship between crude oil and natural gas appears relatively weak, with small coefficients across different lags.

These findings suggest that while significant interconnections exist between Indian stock markets and energy commodities, the relationships are complex and time-varying. Compared to natural gas, the stronger bidirectional relationship between stock markets and crude oil highlights the importance of oil prices for Indian financial markets.

### 5.4.3 Granger Causality Test Results

It is important to note that the Granger causality test is designed to identify linear causal relationships, which implies that the detected causality is based on the predictive ability of one variable in relation to another within the framework of a linear model.

The test results reveal significant causal relationships among the variables under consideration (refer to Table A4). Crude oil prices have a statistically significant causal effect on the Nifty 500 index and natural gas prices (p-value $< 0.01$). Similarly, the Nifty 500 index demonstrates a statistically significant causal effect on crude oil and natural gas prices (p-value $< 0.01$). However, natural gas prices do not exhibit a statistically significant causal effect on the Nifty 500 index or crude oil prices (p-value = 0.65).

### 5.4.4 Johansen cointegration test

An enquiry into the long-term relationship among the variables is made using the Johansen cointegration test. "According to Engel and Granger (1987), if the variables are found to be cointegrated, they would not drift apart over time, and the long-run combination amongst the non-stationary variables can be established." The test results reveal a significant long-term equilibrium relationship among natural gas prices and the Nifty index (Table A5, panel a). The test identifies two cointegrating relationships, as the trace statistic for $r = 0$ is 1614.13, far exceeding the critical value of 24.60 at the 1% significance level, and for $r \leq 1$, it is 738.23, also well above the critical value of 12.97. The first cointegrating vector indicates that a unit increase in natural gas prices corresponds to a 1.189-unit decrease in the Nifty index in the long run. The adjustment coefficients show that natural gas prices exhibit a stronger correction mechanism with a loading value of -1.001, compared to the Nifty index's weaker adjustment at 0.0816. These findings confirm a robust cointegrating relationship,

highlighting the interdependence of energy prices and stock market movements.

The test confirms a long-term equilibrium relationship between crude oil prices and the Nifty index, suggesting that changes in crude oil prices significantly influence stock market movements (refer to Table A5, panel b). The test indicates the presence of at least one cointegrating relationship, demonstrating that these variables move together over the long term despite short-term fluctuations. The cointegrating vectors reveal that an increase in crude oil prices is associated with a decrease in the Nifty index, highlighting the inverse relationship between energy prices and market performance. The loading matrix suggests that crude oil prices dominate in correcting deviations from the equilibrium while the stock market adjusts more slowly. These findings underline the critical impact of global energy price dynamics on financial markets.

The test indicates a significant long-term relationship among natural gas and crude oil prices (Table A5, panel c). The test results confirm at least one cointegrating vector, suggesting that these energy commodities are closely tied in their long-term price movements. The coefficients in the cointegrating relation highlight that a change in crude oil prices has a measurable impact on natural gas prices, reflecting the interdependence of these energy markets. As revealed by the loading matrix, the adjustment speeds show that crude oil prices play a larger role in correcting deviations from the equilibrium, while natural gas prices adjust more slowly. These findings underscore the integrated nature of energy markets and the influence of crude oil prices on related energy commodities.

Hence, all the series are cointegrated and showcase a long-term relationship.

## 6. Conclusion and Policy Recommendations

The study's main objective was to check the interaction between the Indian stock market and commodity market, taking values of a benchmark index, Nifty 500, from NSE and two important energy sector commodities, crude oil and natural gas, from MCX. This study provides compelling evidence of the intricate relationships between India's stock market and key energy commodities. The empirical analysis reveals several crucial findings. First, both markets demonstrated significant vulnerability to external shocks, as evidenced by the sharp declines during the COVID-19 pandemic. Second, the analysis through the risk-return framework depicts nifty 500 as the best investment avenue with minimal risk and high return. Crude oil can be seen as a possible avenue. Natural gas cannot be seen as a potential commodity for investment due to the negative average returns offered during the last 7-8 years.

Third, the Granger causality tests establish bidirectional causality between crude oil prices and the Nifty 500 index, while natural gas prices do not show a causal influence on the Nifty 500 and Crude Oil prices. Fourth, the Johansen cointegration analysis confirms the existence of long-term equilibrium relationships among all three variables, highlighting their fundamental interconnectedness in the Indian financial system. The inverse relationship between energy prices and market performance is found to exist.

These findings put forward important implications for policymakers, regulators, and market participants. From a risk management perspective, there is a clear need to establish an integrated market monitoring system that tracks real-time correlations between energy commodities and stock market movements. This should be complemented by early warning systems for potential market disturbances based on commodity price volatility. The government should also strengthen its strategic petroleum reserves management to buffer against external shocks.

On the regulatory front, we recommend implementing mandatory disclosure requirements for listed companies regarding their commodity price exposure. It should include standardized reporting formats for commodity

risk assessment and specialized regulatory guidelines for companies with significant exposure to energy commodities. The strong interconnection between markets suggests that regulatory bodies should adopt a more coordinated approach to oversight.

Market infrastructure development is another crucial area for improvement. Creating specialized derivative instruments would enable better hedging of commodity-related risks, while improved price discovery mechanisms for energy commodities would enhance market efficiency. The market infrastructure should be enhanced to facilitate seamless integration between commodity and stock markets, reducing cross-market transactions and information flow friction.

Information dissemination and education also require significant attention. Regulatory bodies should create dedicated channels for disseminating commodity market information to stock market participants. It should be supported by comprehensive educational programs focusing on the interconnections between commodity and equity markets. Enhanced transparency and understanding enable market participants to make well-informed investment choices and more effectively manage their exposure to risk.

The present study limits the analysis to include only two energy commodities and a benchmark index from NSE. This limitation can be addressed by including certain control variables to capture the effects of macroeconomic and global factors. Future research can explore the impact of other commodities on the behaviour of stock market using a different benchmark index, like Sensex from the Bombay Stock Exchange and investigate the transmission mechanisms through several commodity price shocks affecting the Indian economy. Such research would further enhance understanding of these complex market dynamics and help in developing more effective policy responses.

## REFERENCES

Anna Creti, Marc Joëts, Valérie Mignon. 'On the links between stock and commodity markets' volatility', Energy Economics, Volume 37, May 2013, Pages 16-28,

Bhullar Pritpal, Gupta, P.K. & Tandon, D. (2024). 'Interdependence Between Energy Commodities and Global Stock Indices: A Volatility Transmission Approach,' FIIB Business Review, November 2024.

Elsayed, Ahmed H. & Nasreen, Samia & Tiwari, Aviral Kumar, 2020. "Time-varying co-movements between energy market and global financial markets: Implication for portfolio diversification and hedging strategies," Energy Economics, Elsevier, vol. 90(C).

Enders, W. (2004). Applied Econometric Time Series. 2nd edition, Wiley Series in Probability and Statistics.

Engel, R. F., Granger W.J. 1987. Cointegration and Error Correction: Representation, Estimation, and Testing. Econometrica 55:2, 251-276.

Hanif, W., Teplova, T., Rodina, V., and Alomari, M. (2023). 'Volatility Spillover and frequency dependence between oil price and green market stocks', Resource Policy, 85 (2023), 103860.

Lucey, Brian, and Ren, Boru (2023), 'Time-varying tail risk connectedness among the sustainability-related products and fossil energy investments', Energy Economics, 126 (2023) 106812.

Maghyereh, A. I. & Awartani, B. & Bouri, E., (2016). 'The directional volatility connectedness between crude oil and equity markets: New evidence from implied volatility indexes,' Energy Economics, Elsevier, vol. 57(C), pages 78-93.

Marco J. Lombardi, & Francesco Ravazzolo (2016). 'On the correlation between commodity and equity returns: Implications for portfolio allocation,' Journal of Commodity Markets, June 20, Journal of Commodity Markets 2(2016) 45–57.

Mensi, W., Al Rababa'a, A. R., Alomari, M., Vo, X. V., & Kang, S. H. (2022). Dynamic frequency volatility spillovers and connectedness between strategic commodity and stock markets: US-based sectoral analysis: US-based sectoral analysis. Resources Policy, 79, Article 102976.

Naik, P. K. (2013). Does the stock market respond to economic fundamentals? Time-series analysis from Indian data. Journal of Applied Economics and Business Research, 3(1), 34–50.

Nova, T., (2024). 'India's energy profile: View from the South', Observer Research Foundation, July 2024, online

Peng, Q., & Bai, M. (2022). Does oil price uncertainty matter in stock market volatility forecasting? PLoS One, 17(12), e0277319. Enders, W. 2004. Applied Econometric Time Series. 2nd edition, Wiley Series in Probability and Statistics.

Shahani, R., Sharma, A.K. and Chhugani, G. (2023). 'Natural gas and crude: evidence of nonlinear cointegration from Indian commodities market,' Int. J. Energy-Growth Nexus, Vol. 1, No. 1, pp.40-62.

Syzdykova, A. & Azretbergenova, G. (2024). ' Asymmetric Effect of Oil Prices on Kazakhstan's Stock Market Index and Exchange Rate', International Journal of Energy Economics and Policy, 2024, 14(6), 15–23.

**Websites**

https://energyeducation.ca/encyclopedia/Barrel#:~:text=A%20barrel%20(bbl)%20is%20a,%22price%20of%20the%20barrel%22.

https://www.adanigas.com/png-commercial/explore-mmbtu

https://www.mcxindia.com/market-data/spot-market-price

https://www.nseindia.com/reports-indices-historical-index-data

https://www.orfonline.org

https://www.researchgate.net/post/Why_do_we_use_log-retuns_in_financial_time_series_modelling

https://www.scirp.org/html/6-2410229_79227.htm#f1

www.mospi.gov.in

## Appendix

**Table A1: Data Series in Absolute Numbers**

| | Date | Nifty_500_Cl_price | Crude_price | N.Gas_price |
|---|---|---|---|---|
| 1 | 2017-01-02 | 7002.5 | 3651 | 253.1 |
| 2 | 2017-01-03 | 7028.7 | 3651 | 253.1 |
| 3 | 2017-01-04 | 7030.7 | 3563 | 226.5 |
| 4 | 2017-01-05 | 7106.9 | 3631 | 222.7 |
| 5 | 2017-01-06 | 7083.1 | 3644 | 221.9 |
| 6 | 2017-01-09 | 7085.1 | 3669 | 223.2 |
| 7 | 2017-01-10 | 7132.2 | 3543 | 211.6 |
| 8 | 2017-01-11 | 7214.1 | 3459 | 223.1 |
| 9 | 2017-01-12 | 7231.8 | 3565 | 220.0 |
| 10 | 2017-01-13 | 7228.3 | 3609 | 230.5 |
| 11 | 2017-01-16 | 7247.9 | 3573 | 233.3 |
| 12 | 2017-01-17 | 7243.6 | 3573 | 233.3 |
| 13 | 2017-01-18 | 7269.0 | 3571 | 232.2 |
| 14 | 2017-01-19 | 7287.7 | 3470 | 224.3 |
| 15 | 2017-01-20 | 7202.7 | 3502 | 229.6 |

**Table A2: Data series in Log Return Values**

| | dataactual.Date | Nifty_log_ret | Crude_log_ret | N.Gas_log_ret |
|---|---|---|---|---|
| 1 | 2017-01-03 | 3.734539e-03 | 0.0000000000 | 0.0000000000 |
| 2 | 2017-01-04 | 2.845072e-04 | -0.0243982160 | -0.1110397226 |
| 3 | 2017-01-05 | 1.077987e-02 | 0.0189052058 | -0.0169193707 |
| 4 | 2017-01-06 | -3.354478e-03 | 0.0035738870 | -0.0035987443 |
| 5 | 2017-01-09 | 2.823224e-04 | 0.0068371660 | 0.0058414006 |
| 6 | 2017-01-10 | 6.625755e-03 | -0.0349453195 | -0.0533705305 |
| 7 | 2017-01-11 | 1.141770e-02 | -0.0239942959 | 0.0529224015 |
| 8 | 2017-01-12 | 2.450524e-03 | 0.0301845239 | -0.0139925551 |
| 9 | 2017-01-13 | -4.840907e-04 | 0.0122666718 | 0.0466233161 |
| 10 | 2017-01-16 | 2.707895e-03 | -0.0100251466 | 0.0120743166 |
| 11 | 2017-01-17 | -5.934514e-04 | 0.0000000000 | 0.0000000000 |
| 12 | 2017-01-18 | 3.500410e-03 | -0.0005599104 | -0.0047261098 |
| 13 | 2017-01-19 | 2.569265e-03 | -0.0286910747 | -0.0346146277 |
| 14 | 2017-01-20 | -1.173204e-02 | 0.0091796399 | 0.0233542229 |
| 15 | 2017-01-23 | 5.510495e-03 | 0.0342441543 | -0.0509265910 |

**Table A3: Augmented Dickey-Fuller (ADF) Test Results**

```
        Augmented Dickey-Fuller Test

data:  datalogret$Nifty_log_ret
Dickey-Fuller = -11.373, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  datalogret$Crude_log_ret
Dickey-Fuller = -12.176, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  datalogret$N.Gas_log_ret
Dickey-Fuller = -11.668, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary
```

**Table A4: Granger's Causality test results**

```
> print(grangertest_crude_to_nifty_gas$Granger)

        Granger causality H0: Crude_Return do not Granger-cause Nifty_500_Return
        N.Gas_Return

data:  VAR object var_model2
F-Test = 5.7046, df1 = 10, df2 = 5733, p-value = 1.451e-08

> print(grangertest_nifty_to_crude_gas$Granger)

        Granger causality H0: Nifty_500_Return do not Granger-cause Crude_Return
        N.Gas_Return

data:  VAR object var_model2
F-Test = 3.4834, df1 = 10, df2 = 5733, p-value = 0.0001376

> print(grangertest_gas_to_crude_nifty$Granger)

        Granger causality H0: N.Gas_Return do not Granger-cause Nifty_500_Return
        Crude_Return

data:  VAR object var_model2
F-Test = 0.77126, df1 = 10, df2 = 5733, p-value = 0.6569
```

**Table A5: Johansen Cointegration test results**

**Panel a) Natural Gas and Nifty 500 Index**

```
Test type: trace statistic , without linear trend and constant in cointegration

Eigenvalues (lambda):
[1] 0.3648115 0.3178480 0.0000000

Values of teststatistic and critical values of test:

           test 10pct  5pct  1pct
r <= 1 |  738.23  7.52  9.24 12.97
r = 0  | 1614.13 17.85 19.96 24.60

Eigenvectors, normalised to first column:
(These are the cointegration relations)

                      timeseries_n.gas_re.l2 timeseries_Nifty_re.l2
timeseries_n.gas_re.l2           1.000000e+00             1.00000000
timeseries_Nifty_re.l2          -1.189063e+00            13.29157659
constant                         5.503856e-05            -0.01005018
                       constant
timeseries_n.gas_re.l2  1.0000000
timeseries_Nifty_re.l2 -0.9685869
constant                8.0383949

Weights W:
(This is the loading matrix)

                     timeseries_n.gas_re.l2 timeseries_Nifty_re.l2
timeseries_n.gas_re.d            -1.00165328            -0.08639188
timeseries_Nifty_re.d             0.08167139            -0.06653928
                         constant
timeseries_n.gas_re.d -2.435938e-20
timeseries_Nifty_re.d  5.785462e-21
```

**Panel b) Crude oil and Nifty 500 Index**

```
Test type: trace statistic , without linear trend and constant in cointegration

Eigenvalues (lambda):
[1] 4.050380e-01 3.126414e-01 5.551115e-17

Values of teststatistic and critical values of test:

           test 10pct  5pct  1pct
r <= 1 |  723.56  7.52  9.24 12.97
r = 0  | 1725.72 17.85 19.96 24.60

Eigenvectors, normalised to first column:
(These are the cointegration relations)

                      timeseries_crude_re.l2 timeseries_Nifty_re.l2
timeseries_crude_re.l2           1.000000e+00             1.000000000
timeseries_Nifty_re.l2          -1.037431e+00            10.449226592
constant                        -6.454950e-06            -0.008000836
                       constant
timeseries_crude_re.l2  1.000000
timeseries_Nifty_re.l2 -7.358074
constant               54.168231

Weights W:
(This is the loading matrix)

                     timeseries_crude_re.l2 timeseries_Nifty_re.l2
timeseries_crude_re.d            -1.04801233             -0.07725963
timeseries_Nifty_re.d             0.08608218             -0.08434812
                        constant
timeseries_crude_re.d 2.67560e-21
timeseries_Nifty_re.d 1.51703e-21
```

**Panel c) Natural gas and Crude Oil**

```
Test type: trace statistic , without linear trend and constant in cointegration

Eigenvalues (lambda):
[1] 4.029607e-01 3.563966e-01 1.110223e-16

Values of teststatistic and critical values of test:

           test 10pct  5pct  1pct
r <= 1 |  850.50  7.52  9.24 12.97
r = 0  | 1845.94 17.85 19.96 24.60

Eigenvectors, normalised to first column:
(These are the cointegration relations)

                      timeseries_n.gas_re.l2 timeseries_crude_re.l2
timeseries_n.gas_re.l2           1.000000000             1.000000000
timeseries_crude_re.l2          -3.072654751             0.376549965
constant                         0.001469408            -0.001049497
                       constant
timeseries_n.gas_re.l2 1.00000000
timeseries_crude_re.l2 0.04575895
constant               8.22144132

Weights W:
(This is the loading matrix)

                     timeseries_n.gas_re.l2 timeseries_crude_re.l2
timeseries_n.gas_re.d             -0.1400168              -0.9523032
timeseries_crude_re.d              0.3245633              -0.3247392
                        constant
timeseries_n.gas_re.d 6.428159e-21
timeseries_crude_re.d 1.035276e-20
```

**Discover** Computing

Research

# Enhanced software change proneness prediction in android applications using balanced data techniques and advanced language models

**Ruchika Malhotra**[1] · **Jyoti Patidar**[1]

Received: 22 June 2024 / Accepted: 11 April 2025
Published online: 27 April 2025
© The Author(s) 2025 OPEN

**Abstract**
Software change proneness prediction in Android applications is crucial in enhancing software maintenance and evolution processes. Inadequate consideration of data imbalance may compromise the reliability and robustness of change proneness prediction models. The study proposes a novel approach that leverages advanced techniques to address data imbalance while improving predictive accuracy. The methodology integrates state-of-the-art techniques to address key challenges. SMOTE (Synthetic Minority Over-sampling Technique) is employed for handling class imbalance, Self-Attention and RoBERTa embeddings are used to extract rich contextual information from text inputs, and Recursive Feature Elimination (RFE) is applied for feature selection to enhance model interpretability and performance. Moreover, we employ XLNet coupled with Butterfly optimization to reduce model weight while maintaining high predictive accuracy effectively. Our hybrid approach offers a robust framework for software change proneness prediction in Android applications by combining advanced language representation learning with optimization techniques. The efficiency of the presented technique is analyzed based on various metrics and research work compared with different state-of-the-art works.

**Keywords** Proneness prediction · Synthetic minority over-sampling technique · Self-attention and RoBERTa · XLNet · Butterfly optimization

**List of symbols**

| | |
|---|---|
| $x_{min}$ | Minimal value |
| $x_{max}$ | Maximum value |
| $d_i$ | Dataset |
| M | Sample's number |
| $\sigma$ | Standard deviation |
| z | Number of minority class |
| r | Number of majority class |
| $C_0$ | Class 0 |
| $C_1$ | Class 1 |
| $x_{syn}^{j}$ | Synthetic sample |
| $\lambda_j$ | Random number |

✉ Jyoti Patidar, jyotipatidar_2k21phdcs01@dtu.ac.in | [1]Department of Software Engineering, Delhi Technological University, Delhi 110042, India.

| | |
|---|---|
| $x_h$ | Character's encoding |
| $\chi_{token_emb}$ | Token embedding |
| $x_{Seg_emb}$ | Segment embedding |
| $W_i^I$, $W_i^K$, $W_i^F$, and $W^O$ | Attention mechanism's weight matrices |
| $X_N$: $N^{th}$ | Butterfly's position |
| $X_i'$ | Opposite solution |
| $d_i^t$ | Solution vector |

## 1 Introduction

One of the most unexpected events that can occur throughout a system's lifetime is software change. It is more closely related to software design theories that recommend best practices than they are to detailed instructions on how a design has to be built[1–3]. Perhaps the most current study topic in software engineering is software defect prediction, which is crucial to ensuring the quality of software [4, 5]. A key component of software success is software quality. The primary factors influencing quality are its qualities, which may be broadly classified into two distinct groups: internal and external. While exterior features like maintainability must be evaluated indirectly and frequently rely on internal attributes for prediction, internal attributes like class cohesiveness may be measured directly from the software [5]. By identifying potentially problematic code, defect forecasting techniques try to assist developers in prioritizing their testing and debugging efforts [6, 7]. Consequently, the development of models to predict and improve change- and bug-prone components, during their early lifetime, would be highly beneficial to lessen future maintenance efforts [8]. Due to, e.g., meeting customers' demands, quality enhancement, bugs fixing, changes in technology, etc. [9].

It has given rise to machineries in the form of many research studies that are already conducted to identify the modules which are prone to change in the earlier project phases for several reasons. Unified modeling language (UML) is a graphical language for visualizing, specifying, constructing, and documenting software intensive systems [10]. UML has limitations, such as complexity and the absence of domain-specific customization, along with ambiguity in larger systems. Some works have been proposed using machine learning techniques to predict change-prone classes, including Bayesian networks [11], neural networks [12], and ensemble methods [13, 14]. Data quality, feature selection and generalizability across different software systems may limit these methods.

Other scholars focused on isolating those entities that are prone to change from those that are not change-prone. They modeled change-proneness prediction as a binary classification problem and applied the classification methods in data mining [15]. Therefore, conventional classification methods are incapable of offering satisfactory performance in change-proneness prediction. [16, 17] employed traditional classifiers without addressing class imbalance issue. However, reference [18] only considering the cost-sensitive learn methods and resampling method on class imbalance problem. In the early days an approach was proposed in [19] that predicts a probability of a software change using a regression model. However, regression models are flawed; assuming linearity cannot represent intricate relationships within software evolution metrics. Proposed XLNet is used for Proneness Prediction along with Butterfly optimization to effectively minimize the model weight to solve these problems.

### 1.1 Motivation

Change proneness prediction in Android apps is a critical component of software maintenance and evolution, significantly impacting the effectiveness and quality of software development processes. However, an increasingly common issue arises from unbalanced datasets, where the proportion of change-prone elements is significantly greater than that of non-change-prone elements. This imbalance poses significant challenges for traditional prediction models, frequently resulting in biased outcomes and diminished predictive accuracy. This challenges the prediction model to identify the components that are prone to change with the presence of imbalanced data, and motivates unique techniques to be developed in this field. Researchers and practitioners are thus using state-of-the-art methodologies, which do not resort to over-sampling or under-sampling techniques such as SMOTE, that alleviate the class imbalance problem to improve the effectiveness of software maintenance practices. The final research aim is the provision of the ability to proactively identify potential issues in Android applications, as well as to optimize resource allocation and improve software quality and user satisfaction fostering an integrated and adaptable development process within the field of Android programming.

### 1.2 Problem statement and contributions

Proneness prediction focuses on the prediction of software components that are likely to contain defects which can help allocate quality assurance efforts appropriately. A variety of approaches have been explored, such as graph-based neural networks, line-level defect prediction frameworks, and cross-project defect prediction models, each achieving varying levels of effectiveness. Nonetheless, such approaches commonly encounter limitations including low generalization capability across heterogeneous datasets, underperformance for modeling long-range dependency, and inefficiency for characterizing bidirectional context correlation in source code. To overcome these limitations, this study applies XLNet, a transformer-based model, to predict proneness. The proposed approach aims to enhance the accuracy and robustness of defect prediction across different datasets and software systems by capitalizing on XLNet's ability to capture sequential as well as bidirectional context.

The key contribution of the proposed work is,

- We propose an innovative framework specifically designed to enhance software change proneness prediction in Android applications, addressing challenges in predictive reliability and performance.
- We utilize Synthetic Minority Oversampling Technique (SMOTE) in our approach to defeat the challenge of class imbalance that has been a longstanding issue in prior work. This helps in creating a balanced dataset which greatly increases the robustness of the predictive models.
- Our approach utilizes combinations of self-attention and RoBERTa embeddings to acquire complex, contextual patterns from textual data, which is often missed in traditional methods.
- Utilize Recursive Feature Elimination (RFE) to systematically prune the dataset to retain only the most informative features. This increases model interpretability and retains better performance, solving the black box problem of most complicated machine learning models.
- XLNet for its superior contextual understanding capabilities and couple it with the Butterfly Optimization Algorithm to minimize model weight. This innovation reduces computational overhead while preserving high predictive accuracy, directly addressing the efficiency limitations of existing solutions.
- By combining advanced language representation learning techniques with optimization strategies, our hybrid approach establishes a robust framework. This addresses gaps in prior research by simultaneously enhancing predictive accuracy and computational efficiency.
- The proposed framework directly supports improved software maintenance and evolution processes in Android applications, providing practical utility for developers by predicting change proneness with high reliability.

## 2 Literature review

Šikić et al. [20] Prediction of defective software modules before testing is a beneficial process which helps to minimize the time and cost of software testing as well. Proposed an end-to-end model based on a convolutional graph neural network (GCNN) for defect prediction that can have its architecture tuned according to the software being analyzed, enabling projects of different sizes to process it with the same granularity. On the PROMISE dataset, DP-GCNN manner better than classical SDP models built upon static code characteristics in terms of AUC and F-score, and can achieve comparable predictive performance with state-of-the-art AST-based models. This model is yet to be validated on different programming languages and source code files of varying sizes.

Wattanakriengkrai et al. [21] Defect prediction models were introduced to help a team set their priorities from source code files that require Software Quality Assurance (SQA) depending on whether they have defects or not. In this work, they introduced a model-agnostic approach, IMAGE-DP to basically find defective lines. By achieving higher recall with lower false alarm levels in identifying defective lines, LINE-DP significantly outperforms baseline approaches, effectively prioritizing SQA effort within manageable computation time. Execution time for LINE-DP was higher than that of some baseline methods, though still manageable, and performance may vary based on particular software systems and defect types.

Ni et al. [22] applied to target project defect prediction models built on the source project that had been active research were specific types of cross-project defect prediction (CPDP). Hence, they propose that the Manual Down should be used as the baseline in studies considering non-effort-aware performance measures (NPMs) and the ManualUp should be treated as the baseline when considering effort-aware performance measures (EPMs) in future CPDP studies.

In almost all cases, EASC surpasses the unsupervised method ManualUp, especially in limited inspection cost scenarios, with observations that are consistent across two different classifiers. EASC's empirical findings need to be validated on larger and more diverse datasets, like actual commercial projects.

Dalla Palma et al. [23] An Infrastructure-as-code (IaC) would be the DevOps practice introduced, that allowed us manage and provision of infrastructure through the provisioning of machine-readable definition files, hereinafter will be called IaC scripts. They proposed a fully integrated machine learning framework for IaC Defect Prediction which enables repository crawling, metrics collection, model building, and evaluation. The proposed approach is able to predict failure-prone IaC scripts, yielding a median AUC-PR of 0.93 (very high for this type of experiments), for Random Forest (well-known ML algorithms), and thus, the authors provide a good basis (and data set) for researchers and DevOps. Beyond that, the approach does not yet know if it can be applied to other configuration orchestration languages or how it performs with semi-supervised/unsupervised learning approaches.

Hassouneh et al. [24] The proposed Software fault prediction (SFP) was a challenging process which any efficient software should cope with to ensure that all software components are fault-free. They also proposed an improved Whale Optimization Algorithm (WOA) that works in conjunction with a single point crossover scheme. TBWOA outperformed all state-of-the-art methods used for SFP applications, showing the highest overall AUC and effectively improving feature selection with natural selection schemes. It is yet to be studied whether TBWOA can be extended to tackle other real world optimization problems like traveling salesman and educational timetabling.

Tumar et al. [25] Software fault prediction (SFP) can be a challenging task that developers face during the software development life cycle. For predicting software faults, they proposed Enhanced Binary Moth Flame Optimization (EBMFO) with Adaptive synthetic sampling (ADASYN) in this research. Thus, it can be concluded that the BMFO-based approach with ADASYN plays a significant role in enhancing the performance of classifiers for the prediction of software faults. Amongst the classifiers, the LDA is achieving the maximum AUC whilst kNN has shown the best performance in terms of execution time. Feature selection is significant in improving classifier performance and SFP model accuracy; hence it is an area that needs further exploration in upcoming work.

Yuan et al. [26] CPDP-Methods can be used when the target project was a new project or there are not enough labelled program modules in the target project. They first proposed ALTRA, a new method that can leverage both active learning and TrAdaBoost to help resolve such disadvantage in this article. The ALTRAjen method can significantly enhance cross-project defect prediction by leveraging active learner and TrAdaBoost, outperforming seven state-of-the-art baselines on large-scale open-source projects. However, the performance of the method may be limited by the quality and diversity of the labeled modules and would be need further fine-tuning for more sophisticated defect prediction applications.

Yedida et al. [27] It is claimed that the deep learning allows automatic feature engineering avoids the laborious manual feature engineering before doing teh DL This problem is solved via a preprocessing stage (a new oversampling technique, "fuzzy sampling" is introduced as part of pipeline GHOST (Goal-oriented Hyper-parameter Optimization for Scalable Training) and they can outperform DL State of the art in 14/20 defect datasets. Pros: Unlike existing methods, GHOST harmonizes deep learning with adaptive oversampling, weighted loss function and fuzzy sampling to enhance defect prediction performance by providing state-of-the-art performance in several datasets. This only highlighted the application of the method on defect prediction, thus it needs further exploration in order to quantify its applicability across other software analytics.

Khan et al. [28] focused Software testing was an essential part of software development activities and to be used most of the resource that is time, cost and effort. Although a model was proposed that predicted software bugs, using machine learning classifiers with Artificial Immune Network (AIN) to optimize hyper-parameter for better accuracy in predicting bugs. The application of hyper-parameter optimization through opt-aiNet has led to a significant increase in the prediction accuracy and AUC performance of machine learning classifiers for predicting software bugs with respect to classifiers such as SVM-RBF and LDA showing prominent improvement after hyper-parameter tuning. The effects of hyper-parameter optimization were different across the classifiers, and more studies using more different classifiers and optimization techniques, particularly deep learning and resampling methods, are needed to gain an understanding of the effects on the prediction of software bugs.

Zhao et al. [29] MT has supported proposals, Android mobile apps have been essential in our daily life and work. They presented a new Imbalanced Deep Learning model to perform the JIT defect prediction task for Android mobile applications. Class imbalance is one of the primary challenges in defect prediction for JIT text in Android mobile software, especially with JIT patterns, and today we show that our novel cost-sensitive cross-entropy loss function (IDL method) outperforms 23 imbalanced learning methods for JIT cases in Android mobile apps. The effectiveness of the method was assessed for selected Android apps, but it requires further work to generalize it to cross-project scenarios to make it more broadly applicable.

The existing literature on defect prediction models predominantly focuses on utilizing various machine learning techniques to identify defective software components. While these models have shown promising results, there remains a gap in the research regarding the optimization of hyperparameters for these classifiers. Additionally, there is limited exploration into the effectiveness of integrating advanced oversampling techniques, such as fuzzy sampling, within deep learning pipelines to enhance defect prediction accuracy. Contextual relationships in defect prediction were neither evaluated nor used, thus existing defect prediction models could be further enhanced by exploring these gaps. Table 1 demonstrates the summary of the literature reviewed in this study.

## 3 Proposed method

The technique for software change proneness prediction in the context of Android applications comprises three main components: data pre-processing, feature extraction and model training and optimization. Data cleaning is performed in a few steps during the data pre-processing phase. It handles the missing values with various methods to ensure the completeness of the dataset, and normalizes and scales the data excellent way to help the model perform better. Other preprocessing steps included encoding categorical variables to make them into numerical formats to be used in machine algorithms. The class imbalance problem in change proneness data is handled in this case using SMOTE. This helps to create a balanced dataset that reduces the chance of bias in the prediction model by generating synthetic samples of the minority class. This entire pre-processing is indeed important to improve the model accuracy of less occurrences of class. Step 2: Feature ExtractionIn this stage we employ Self-Attention mechanisms and the RoBERTa embeddings to capture the meaningful features that represent contextual information regarding software change proneness. These approaches can effectively learn deep contextual relations from the code and documentation. The RFE implemented approach was responsible for providing the features for informative properties selection at the stage of training and optimization of the model, thereby increasing its interpretability and improving the performance of the developed models. We use XLNet, a state-of-the-art language model, as the basis of our prediction model, and we optimize it with the lightweight Butterfly optimization algorithm to maintain high total predictive accuracy with a smaller model weight. The proposed work is a hybrid approach of learning representation of language and optimization, finally offering a strong and promising framework for prediciting change proneness in Android applications. Figure 1:Flow Diagram of Software Change Proneness Prediction.

### 3.1 Data preprocessing

#### 3.1.1 Handling missing values

Since missing or null data might cause any classifier to provide an inaccurate prediction, it is necessary to fill them in. According to Eq. (1), missing or null values that are extended are imputed to the mean attribute values in the suggested design, instead of omissions. Imputation is used to improve the data while avoiding the introduction of outliers into the continuous data.

$$I(x) = \{mean(x), ifx = null/missedx, otherwise \tag{1}$$

This feature vector $x$ consists of events that take place in space, $x \in j^n$

#### 3.1.2 Normalization and data scaling

### 3.2 Normalization

Missing values are filled, and normalization is performed during the preprocessing stage. The maximum of each feature in a dataset that is utilized in model training has a distinct distribution. Artificial neural networks struggle to match the facts in certain scenarios. It makes an effort to modify every characteristic to make the range of the real numbers set comparable. There is a plethora of methods available to address this issue. Once the data has been transmitted, each symbolic property is given a numerical value. Data scaling, also known as data normalization, is used to standardize the range of data attributes. Utilizing maximum-minimum normalization, scale the feature values. a certain range of [– 1, 1] since every feature value is normalized.

Discover Computing (2025) 28:51 | https://doi.org/10.1007/s10791-025-09552-y

**Table 1** Summary of literature survey

| Ref. No | Technique | Merits | Demerits |
|---|---|---|---|
| Šikić et al. [20] | Convolutional Graph Neural Network (GCNN) for Defect Prediction (DP-GCNN) | Reduces testing costs, Adaptable to sizes, High predictive performance | Requires large datasets, Complexity in implementation, Language-specific limitations |
| Wattanakriengkrai et al. [21] | LINE-DP (Model-agnostic defect prediction framework) | High recall rate, Prioritizes SQA efforts, Manageable computation time | Higher computation time, Varies by systems, Defect type sensitivity |
| Ni et al. [22] | EASC (Effort-Aware Supervised Classification for Cross-project Defect Prediction) | Outperforms unsupervised methods, Consistent across classifiers | Limited dataset validation, Generality unproven, May require adaptation |
| Dalla Palma et al. [23] | Machine-learning framework for IaC Defect Prediction | High AUC-PR performance, Comprehensive repository analysis, Valuable baseline dataset | Limited language applicability, Unsupervised methods unexplored, Requires large datasets |
| Hassouneh et al. [24] | TBWOA (Whale Optimization Algorithm with single point crossover) | Effective feature selection, Enhances optimization process | Limited real-world applicability, Unexplored problem domains |
| Tumar et al. [25] | EBMFO with ADASYN (Enhanced Binary Moth Flame Optimization with Adaptive Synthetic Sampling) | High AUC with LDA, Efficient execution time | Computational complexity concerns |
| Yuan et al. [26] | ALTRA (Active Learning and TrAdaBoost for Cross-project Defect Prediction) | Handles large-scale datasets | Quality-dependent performance |
| Yedida et al. [27] | GHOST (Goal-oriented Hyper-parameter Optimization for Scalable Training) | Integrates adaptive techniques | Limited domain applicability |
| Khan et al. [28] | Machine learning with Artificial Immune Network (AIN) and opt-aiNet for bug prediction | Improved prediction accuracy | Limited classifier exploration |
| Zhao et al. [29] | Imbalanced Deep Learning (IDL) with cost-sensitive cross-entropy loss function for JIT defect prediction | Outperforms imbalanced methods | Cross-project adaptation needed |

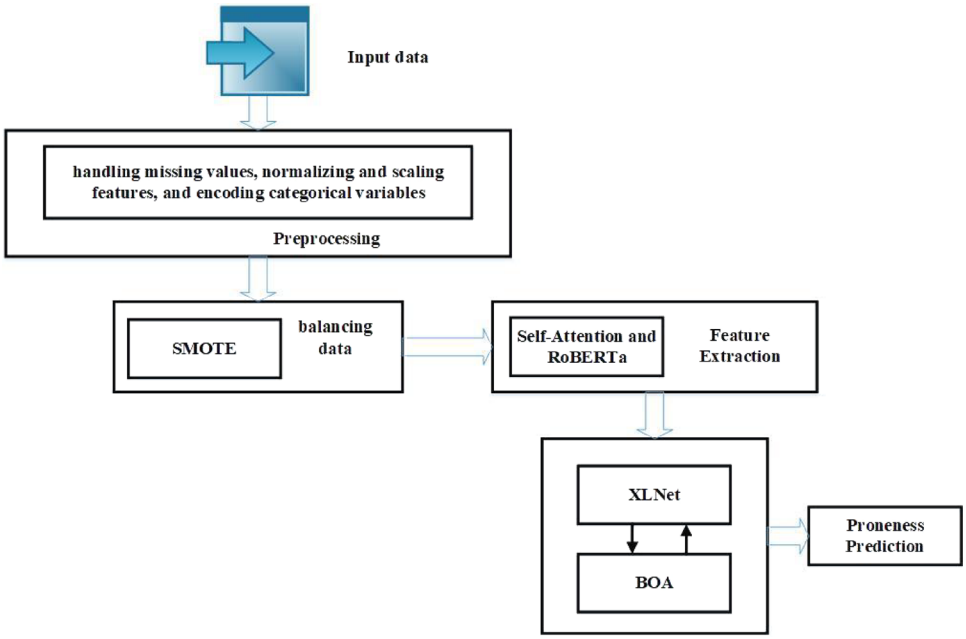



**Fig. 1** Flow diagram of Software Change Proneness Prediction

$$\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Here $x$ shows the value of the sample, $x_{min}$ indicates minimal value, and $x_{max}$ denotes each feature's maximum value.

The standard normal distribution is described by the equation, which may be used to determine the data's mean and standard deviation.

$$\hat{x} = \frac{x - \mu}{\sigma} \tag{3}$$

Here is the meaning of $\mu$:

$$\mu = \frac{1}{M}\sum_{i=1}^{M} d_q \tag{4}$$

Their standard deviation $\sigma$ is:

$$\sigma = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(d_i - \mu\right)^2} \tag{5}$$

Here $M$ indicate the sample's number and $d_i$ denote the dataset's $i^{th}$ $_{\text{of}}$ element.

Logarithmic normalization is used to apply the logarithmic scale to the data in the following equation.

$$\hat{x} = log(x) \tag{6}$$

### 3.3 Data scaling

Data scaling is particularly important for methods that utilize distance measures, such as nearest-neighbor classification and clustering. In addition, artificial Neural Network models require the input data to be normalized, so that the learning process can be more stable and faster.

#### 3.3.1 Encoding categorical variables

When using deep learning techniques, the initial input must always be encoded to extract the pertinent vector or matrix and add it to the learning model. For example, the date and departure time are continuous values, while the aircraft number, airport ID, and temperature are all provided as discrete variables. The next part will introduce the various techniques for encoding various types of data.

(a) Count encoding: Frequency statistics are used to define count encoding, which makes use of how frequently a particular characteristic category takes the place of the category. Under this coding method, different categories will appear at identical frequency; low model learning accuracy will arise from coding two unique categories into the same number. However, dimension explosion will not occur.
(b) Label encoding: In label encoding, a feature attribute's category sequence is used; that is, consecutive integers ranging from 1 to n-1 are used to replace the category. This encoding approach brings order to high-based, non-sized data without causing a dimensionality bursting, however will also result in low algorithm model correctness.
(c) CatBoost encoding: CatBoost encoding is a supervised coding technique that uses statistics to transform high-base information into numerical data. It can prevent dimension explosion and prevent the encoded data from being sequential. Simultaneously, the encoded data falls within the same dimensional range.

#### 3.3.2 Synthetic minority over sampling technique (SMOTE) for balancing data

To manage supplied imbalance data, the SMOTE is used. The data is pre-processed to remove missing values before being normalized with the conventional scalar approach. Following that the SMOTE approach is used to the provided datasets to address the data imbalance problem. SMOTE is a well-known classifier-building approach for the unbalanced dataset. An unbalanced dataset is made up of unequally distributed subordinate output classes. SMOTE is often utilized in imbalance dataset classification challenges. This contributes to categorization generalization. SMOTE is a powerful approach for coping with balanced datasets. It is one of the most extensively used strategies for dealing with difficulties caused by an uneven number of categorization cases in a dataset. To cope with the unbalanced dataset, the suggested algorithm was utilized for SMOTE implementation.

In the first step, they use the SMOTE technique to create a synthetic instance using the equation below:

Here preprocessed misbalancing data $(e)$ is the input.

$$N = 2 \times (r - z) + z \tag{7}$$

where z is the number of minority class samples, r is the number of majority class samples and N is the initial synthetic instance number (newly produced).

In the second stage, they remove the artificial samples that are closer to the majority than the minority and the artificial cases that are nearer the boundary created by SMOTE. The detailed steps of the A-SMOTE method are as follows:

Step 1 The SMOTE-generated synthetic instances are subject to acceptance or rejection based on two criteria, which align with the first step: Supposed to $\hat{e} = \left\{\hat{x}_1, \hat{x}_2, \ldots \hat{x}_N\right\}$ be the collection of fresh artificial examples and $\hat{e}_i^{(j)}$ is $j^{th}$ attribute value of $\hat{e}, j \in [1.M]$. Let $S_m = \left\{S_{m1}, S_{m2}, \ldots S_{mz}\right\}$ and $S_a = \left\{S_{a1}, S_{a2}, \ldots S_{ar}\right\}$ sets representing majority and alternative samples, respectively. They determine the distance between deciding whether to accept or reject the suggestion and $S_{mk}, DD_{minority}\left(\hat{e}_i, S_{mk}\right)$ and distance between $\hat{e}_i$, and the $S_{al}, DD_{majority}\left(\hat{e}_i, S_{al}\right)$ respectively. Compute these distances step-by-step for i from 1 to N using the formula below:

$$DD_{minority}\left(\hat{e}_i, S_{mk}\right) = \sum_{j=1}^{M} \sqrt{\left(\hat{e}_i^{(j)}, k \in [1, z]\right)} \tag{8}$$

$$DD_{majority}\left(\hat{e}_i, S_{al}\right) = \sum_{j=1}^{M} \sqrt{\left(e_i^{(j)} - S_{al}^{(j)}\right)}, k \in [1, r] \tag{9}$$

then get 2 arrays $A_{minority}$ $A_{majority}$ and, which have the following definitions:

$$A_{minority} = \left(DD_{minority}\left(\hat{e}_i, S_{m1}\right), \ldots, DD_{minority}\left(\hat{e}_i, S_{mz}\right)\right) \tag{10}$$

$$A_{majority} = \left(DD_{majority}\left(\hat{e}_i, S_{a1}\right), \ldots, DD_{majority}\left(\hat{e}_i, S_{arz}\right)\right) \tag{11}$$

Afterwards, select the min value $A_{minority}$, $min\left(A_{minority}\right)$, and the min value $A_{majority}$, $min\left(A_{majority}\right)$. If $min\left(A_{minority}\right)$ there are fewer $min\left(A_{majority}\right)$, new artificial samples are accepted; if not, they are eliminated.

$$min\left(A_{minority}\right) < min\left(A_{majority}\right)(accepted) \tag{12}$$

$$min\left(A_{minority}\right) \geq min\left(A_{majority}\right)(rejected) \tag{13}$$

If ten consecutive attempts to meet the acceptance requirement result in a synthetic sample that does not meet the requirement, they accept the synthetic sample closest to the minority to prevent delays brought about by the creation of fake samples that are unable to progressively achieve acceptance conditions and stop the algorithm from working slowly. This sample is treated as a fresh synthetic example since it satisfies the acceptance requirements. Thus, they can produce synthetic examples that are close to the minority class.

Step 2: determine the separation between $\left(\hat{S}_i, S_a\right)$ and every initial majority $S_a$, $Maj_{Rap}\cdot\left(\hat{S}_i, S_a\right)$, expressed as follows:

$$Maj_{Rap}\cdot\left(\hat{S}_i, S_a\right) = \sum_{l=1}^{r}\sum_{j=1}^{M}\sqrt{\left(\left(\hat{S}_i^{(j)} - S_{al}^{(j)}\right)^2\right.} \tag{14}$$

where $Maj_{Raj.}$ is the proximity of the samples to each majority. The equation after that leads us to H, which has the following definition:

$$H = \sum_{i=1}^{n}\left(Maj_{Raj.}\left(\hat{S}_i, S_a\right)\right) \tag{15}$$

Next, they remove half of the artificial samples with the closest distances between $S_a$ and $\hat{S}_i$ to acquire pure data. The dataset now has an equal number of samples for each class, such as majority class and minority class, due to these processes.

Majority Class: $C_0$: Class 0 (Not change-prone).

Minority Class: $C_1$: Class 1 (Change-prone).

In our dataset, we define two classes: $C_0$ (Not change-prone) and $C_1$ (Change-prone). Before applying SMOTE, there are 100 instances of $C_0$ and 20 instances of $C_1$. After applying SMOTE, we generated 80 synthetic instances $C_1$, resulting in a balanced dataset with 100 instances in each class.

#### 3.3.3 Step 3: synthetic sample acceptance in A-SMOTE

Let $x_i$ be the minority class sample, and $x_{i,nn}$ be one of its nearest neighbors. During the process of generating a synthetic sample $_{syn}$, we attempt to generate a valid sample by interpolating between $x_i$ and $x_{i,nn}$.

For each attempt j, the synthetic sample is generated as follows:

$$x_{syn}^{j} = x_i + \lambda_j \times \left(x_{i,nn} - x_i\right) \tag{16}$$

where:

- $x_{syn}^{j}$is the synthetic sample generated in the jth attempt.
- $\lambda_j$ is a random number between 0 and 1 for the jth attempt.

### 3.4 Acceptance criterion

After generating the synthetic sample $x_{syn}^{j}$, it is evaluated against an acceptance criterion $C\left(x_{syn}^{j}\right)$. If $C\left(x_{syn}^{j}\right)$ is satisfied, the sample is accepted. Otherwise, the algorithm attempts to generate another sample.

If after 10 consecutive attempts, the criterion is not met, the algorithm accepts the synthetic sample that is closest to the original minority sample $x_i$. The closest sample is determined by the smallest Euclidean distance between $x_i$ and $x_{syn}^j$ for $j \in \{1,2,\ldots,10\}$.

### 3.5 Closest sample selection

$$x_{syn}^{final} = argmin_{j\in\{1,2,\ldots,10\}} \|x_{syn}^j - x_i\| \tag{17}$$

where:

$x_{syn}^{final}$ is the final accepted synthetic sample.
$\|x_s yn^j - x_i\|$ represents the Euclidean distance between the original sample $x_i$ and the synthetic sample $x_{syn}^j$.

If no acceptable sample is found in 10 attempts, the sample with the smallest distance $x_i$ is accepted as the final synthetic sample to prevent delays. Below pseudo-code for generate synthetic samples using A-SMOTE.

Pseudo-code for Generate synthetic samples using A-SMOTE

```
function generate_A-SMOTE_samples (minority_samples, majority_samples, k, num_samples): synthetic_samples = []
# For each minority sample
for minority_sample in minority_samples:
neighbors = find_neighbors(minority_sample, minority_samples, k)
# Generate synthetic samples
for i = 1 to num_samples:
neighbor = pick_random(neighbors)
synthetic_sample = create_sample(minority_sample, neighbor)
if is_valid_sample(synthetic_sample, majority_samples):
synthetic_samples.append(synthetic_sample)
return synthetic_samples
# Find k nearest neighbors of a sample
function find_neighbors(sample, samples, k):
distances = calculate_distances(sample, samples)
return get_k_closest(distances, k)
# Randomly pick one neighbor
function pick_random(neighbors):
return random_choice(neighbors)
# Create a new sample by blending two samples
function create_sample(sample1, sample2):
alpha = random_between(0, 1)
return sample1 + alpha * (sample2—sample1)
# Check if the synthetic sample is valid
function is_valid_sample(sample, majority_samples):
distance_to_majority = min_distance_to_majority(sample, majority_samples)
return distance_to_majority > threshold
# Calculate distances from a sample to all other samples
function calculate_distances(sample, samples):
return [distance(sample, s) for s in samples]
# Get the k closest neighbors
function get_k_closest(distances, k):
return select_top_k(distances, k)
# Calculate the Euclidean distance between two samples
function distance(sample1, sample2):
return euclidean_distance(sample1, sample2)
# Find the minimum distance from a sample to the majority class
function min_distance_to_majority(sample, majority_samples):
return min(calculate_distances (sample, majority_samples))
```

### 3.6 Self-attention and RoBERTa for feature extraction

Self-Attention mechanisms in models like RoBERTa extract features for Software Change Proneness Prediction by analyzing the relationships between different code elements, identifying patterns indicative of areas likely to undergo modifications. This allows RoBERTa to learn structural and syntactical properties of code snippets, paving the way toward discovering features that predict the proneness of a software component to change. The self-attention layer of RoBERTa allows the model to focus on the relationship between tokens in parts of code snippets, capable of recognizing the characteristics related to program entities that tend to change and thus improve the prediction accuracy of the model. This feature extraction process leverages the inherent structure of code to inform predictions about future modifications, contributing to more effective software maintenance planning.

To be more precise, the embedding representation of the text for the input content (balanced data) is the result of combining segment embedding, position embedding, and token embedding; the formula for creating text embedding looks like this:

$$x_h = x_{token_emb} + x_{seg_emb} + x_{pos_emb} \tag{18}$$

Every character's encoding representation is denoted by $x_h$, its token embedding is represented by $x_{token_emb}$, its segment embedding is represented by $x_{seg_emb}$, and its place in the embedding is indicated by $x_{pos_emb}$.

Subsequently, the embedding model is introduced into the RoBERTa model's middle layer, which includes an encoder within the 12-layer transformer for gathering semantic data. The multi-head attention mechanism, which is made up of several self-attention layers, is the transformer's most important module. The dot product of attention is used to determine the equivalent output of a single self-attention mechanism, and concatenating all heads yields multiple self-attention. The computation for multi-head attention is then obtained by using a linear transformation. Ultimately, the following formulas can be used to determine multi-head attention and self-attention:

$$head_i = Attention(IW_i^I, KW_i^K, FW_i^F) \tag{19}$$

$$MultiHead(I, K, F) = Concat(head_1, head_2, \ldots, head_n)W^O \tag{20}$$

Here $W_i^I$, $W_i^K$, $W_i^F$, and $W^O$ represent the attention mechanism's weight matrices, and $I$, $K$, *and* $F$ stand for the mechanism's input vectors.

Ultimately, the phrase vector found in the RoBERTa model's output layer provides global semantic and contextual information that may be used to ascertain if the short text and the candidate entity's description text share the same semantic context. Figure 2 shows the model.

#### 3.6.1 Input

The balanced dataset with 200 instances (100 Class 0 and 100 Class 1). Each instance might be a text or code snippet in the case of software change proneness prediction.

#### 3.6.2 Example instances

Instance 1 (Class 0): "This code is stable and hasn't changed in multiple releases."

Instance 2 (Class 1): "This component frequently undergoes modifications due to feature updates."

#### 3.6.3 Using RoBERTa for feature extraction

RoBERTa uses self-attention mechanisms to extract features from text data.

#### 3.6.4 Steps:

#### 3.6.5 Tokenization

RoBERTa tokenizes each text instance into tokens that the model can process.
For example:
Instance 1: ["This", "code", "is", "stable", "and", "hasnt", "changed", "in", "multiple", "releases", "."].
Instance 2: ["This", "component", "frequently", "undergoes", "modifications", "due", "to", "feature", "updates", "."].

#### 3.6.6 Embedding

RoBERTa converts tokens into dense vectors (embeddings).
Every token has a high-dimensional vector representation.

#### 3.6.7 Self-Attention mechanism

RoBERTa applies self-attention to capture contextual relationships between tokens. This mechanism assigns different weights to different tokens based on their importance in the context of the entire instance.

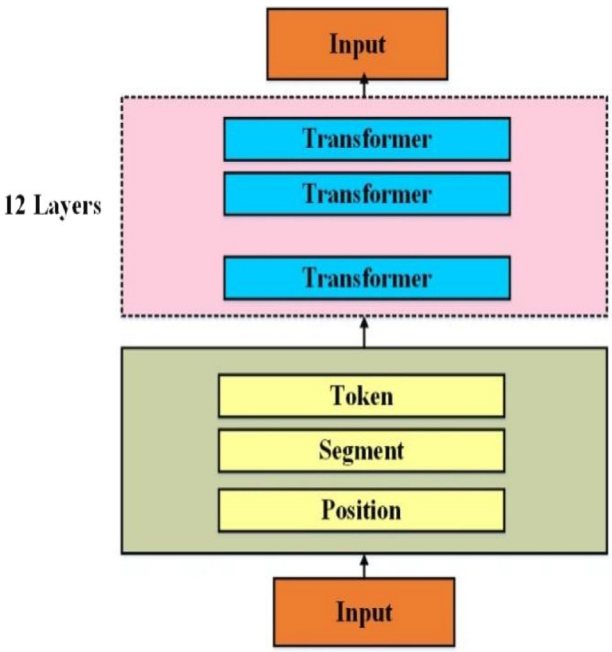
#### 3.6.8 Output features

The final output from RoBERTa for each instance is a set of feature vectors. These vectors represent the contextualized embeddings of the tokens. For simplicity, let's represent the feature vector for each instance as $f_i$.
Example Output:
Instance 1 (Class 0): Feature vector $f_1$ (e.g., first-dimensional vector).
Instance 2 (Class 1): Feature vector $f_2$ (e.g., another dimensional vector).

### 3.7 Recursive feature elimination (RFE) feature selection

They employed the RFE approach to extract the most essential aspects of a prediction in this method. The approach is frequently used because of its ease of use and deployment, as well as its efficiency in selecting feature training datasets

**Fig. 2** RoBERTa Model



that are critical for predicting target variables and removing weak features. The percent RFE approach is used to identify the most crucial traits by determining a high degree of adjustment between a particular variable and goals (labels).

The missing values that are calculated, need to get the relevant attributes whose relationship to the features is strong and positive. Vector feature extraction stops the development of workable models by eliminating characteristics that aren't needed or important for prediction. We used a RFE based approach in this search for determining the most important components in a prediction. RFE method is most widely used as it is a simple method that is easy to implement, and it works well for selecting features in training datasets that are important for predicting target variables while eliminating weak features. The RFE method is used to explore the best correlations between every single feature and the target feature, and then prioritize the feature and assign percentage on each feature based on the correlation.

*Feature Matrix F*: The matrix containing feature vectors for all instances, with dimensions $n \times m$.

*Label Vector y:* The vector containing the class labels for all instances, with dimensions $n \times 1$.

The input to RFE is represented as the pair (*F*,) where *F* is the feature matrix and *y* is the label vector. Represent the input to RFE as follows:(*F*,*y*).

*Initialize a Model:* Choose a model for evaluating the importance of features.

*Feature Importance Ranking:* Fit the model to the data and rank the features based on their importance scores.

*Elimination:* Iteratively remove the least important feature and refit the model. Repeat until the desired number of features *k* is reached:

$(t+1) = (t)$- {least important feature}.

Where (*t*) is the feature matrix at iteration t.

*Final Selected Features:* The remaining feature matrix $F_{\text{selected}}$ after selecting the top*k*w features:

$$F_{\text{selected}} = (\text{w}) \tag{21}$$

### 3.8 XLNet for Proneness Prediction

One of the newest pre-trained models, XLNet, shows promise in several proneness prediction tasks, including sentiment analysis, question responding, and document rating.

Let $w = [w_1, w_2, \ldots, w_T]$ the given input text (selected feature) with length T.

The XLNet learns bidirectional context information through positional encoding and decomposition order. That model's function of objective has a theoretical definition of,

$$\max_{\theta} U_{z \sim Z_T} [\sum_{t=1}^{T} logp_{\theta}(w_{z1} \mid W_{z<t}] \tag{22}$$

where $Z_T$ represents the collection of every conceivable way the input sequence may be changed, with length T. indicating the present $(t - th)$ element. $z < t$ denotes the previous elements $(t - 1)$ of a variation$z \in Z_T$. To forecast the present component or token, the intended use uses input context from the preceding elements. Specifically, the objective function takes into account the order of factorization based on permutations instead of the order of the input sequence [30].

Furthermore, this type of permutation language modeling is accomplished by the transformer-based XLNet architecture through the usage of two-stream self-attention methods. Initially, the context and content are encoded via the content stream self-attention mechanism using the typically hidden state representations shown in the equation. The query stream self-attention mechanism is responsible for encoding the geographical and contextual information, as demonstrated by the following equation.

$$h_{z_t}^{(n)} \leftarrow Attention\left(Qu = h_{z_t}^{(n-1)}, KF = h_{z \le t}^{(m-1)}; \theta\right) \tag{23}$$

$$g_{z_t}^{(n)} \leftarrow Attention\left(Qu = g_{z_i}^{(n-1)}, KF = h_{z<t}^{(m-1)}; \theta\right) \tag{24}$$

where *Qu*, *K*, and *F* indicates the number of attention levels, and represents the query, key, and value.

Additionally, XLNet leverages proneness forecasting in an Android application to reduce the complexity of the autoregressive language model's optimization. This predicts a factorization sequence's last elements. The resulting log-likelihood of reaching the intended subsequence $(z > c)$ is optimized, subject to a non-target subsequence condition $(z \le c)$. In this

case, c represents a subsequence's cutting point. ($Z > c$) Contains the most relevant context in the factorization order ($z$). Additionally, XLNet uses Transformer-XL's segment recurrence method and relative positional encoding. With the use of these two methods, XLNetcan enhances the input text's long-term reliance. As a result, the XLNetpre-trained model is a useful model for tasks involving propensity prediction.

$$\max_{\theta} U_{z\sim Z_T}\left[logp_\theta\left(w_{z>c} \mid W_{z\le c}\right)\right] = U_{z\sim Z_T}\left[\sum_{t=c+1}^{|z|} logp_\theta\left(w_{zt} \mid W_{z\le t}\right)\right] \tag{25}$$

The proneness prediction job involves forecasting change proneness in Android applications by aggregating the supplied input sequence. They supply these aggregated input representations to transformer blocks. Next, an output layer with a softmax activation function is built on top of the XLNet transformer to determine the likelihood of output for every category. Passing through a fully connected layer for fine-tuning, and predicts the data over a given test set. Figure 3 shows the architecture of XLNet.

#### 3.8.1 Hyper-parameter optimization using butterfly optimization algorithm (BOA)

Hyper-parameter optimization is the process of determining the best mix of XLNet's hyper-parameter settings to optimize performance on data in a reasonable quantity. This process is essential to an XLNet's capacity for precise result prediction. Most of this input text used the hyperparameters' default values. The proposed model optimizes the hyperparameter using the butterfly algorithm. Table 2 shows the range of hyperparameter initialization.

The butterfly method is used to optimize these hyperparameters. The butterfly algorithm's step-by-step procedure is explained below.

Step 1: Initialization: Selecting the ideal hyperparameter is the primary goal of this approach. Define the top and lower bounds of the problem, the maximum number of iterations, the dimensionality D of the variables, and the butterfly size N initially. The hyperparameter value is represented by the butterfly in this instance. At first, a random selection is made. The first solution format is seen in the following equation:

$$P_N = \left\{X_1, X_2, \ldots\ldots, X_N\right\} \tag{26}$$

Here, $X_N$ is the $N^t h$ butterfly's position is as follows.

$$X_i = \{n, MaxGen, p, c, a\}_i \tag{27}$$

Step 2: Opposite solution generation**:** opposing solutions are generated once the solution has been initialized. This stage is employed to optimize the algorithm's search capability. Whereas $X_i \in [a, b]$ being a real number, the opposite solution $X_i\prime$ can be calculated as follows;

$$X_i\prime = a + b - X_i \tag{28}$$

Step 3: Calculating fitness: Following initialization, each solution's fitness is assessed using the BOA technique that has been recommended. In this case, the classification accuracy is defined using the fitness function. The solution with the highest fitness value is deemed to be the most effective. The fitness function is calculated with the assistance using,

$$Fitness = Max\left(\frac{TP + TN}{TP + TN + FP + FN}\right) \tag{29}$$

Step 4: Updation using BOA: BOA utilizes 2 distinct techniques known as Fragrance and Movement of Butterflies.

**3.8.1.1 Strategy 1: fragrance (*fr*)** They must first comprehend how a modality such as scent, sound, light, temperature, etc.—is processed by a stimulus to comprehend how fragrance is computed in BOA. Three keywords form the foundation of the entire notion of perceiving and processing the modality, namely, sensory modality (w), stimulus intensity (Q), and power exponent (y). Based on these ideas, the smell in BOA is created in the following way, which changes according to the stimulus's physical intensity:

$$fr = wQ^y \tag{30}$$

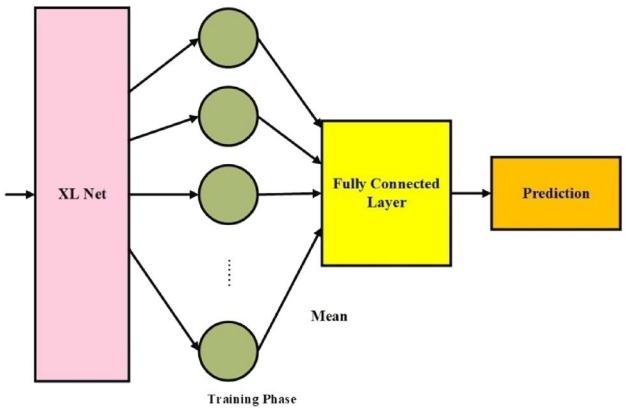



**Fig. 3** Architecture of XLNet

**Table 2** range of hyperparameter initialization

| Hyperparameter | Values |
|---|---|
| Population Size (n) | 20–100 |
| Maximum Iterations (MaxGen) | 50–200 |
| Switch Probability (p) | 0.5–0.9 |
| Sensory Modality (c) | 0.001–0.1 |
| Power Exponent (a) | 0.5–1.0 |

**3.8.1.2 Strategy 2: movement of butterflies** BoA is divided into three stages: the initialization phase, the iteration phase, and the final phase. Each BOA run consists of an initial startup phase, an iterative search phase, and a final phase while the algorithm is ultimately terminated upon finding the best answer. At initiation, the algorithm defines the target function and its solution area. The BOA parameters also have values applied to them. Once the settings are chosen, the technique creates an initial population with butterflies for optimization. A fixed-size memory is set aside to hold the butterflies' data because the overall number of butterflies stays constant during the BOA simulation. Butterfly places are produced at random inside the search space, and their fitness and scent values are computed and recorded. Following this, the algorithm completes the initialization step and begins the iteration phase, which uses the produced fake butterflies to search.

In the method's second phase, also known as the iteration phase, the algorithm runs through several iterations. The method consists of two main steps: the global search phase and the local search phase. The butterfly moves toward the most suitable butterfly/solution f* during the global search phase, which may be represented by the equation,

$$d_i^{t+1} = d_i^t + \left(r^2 \times f - d_i^t\right) \times fr_i \tag{31}$$

where $d_i^t$ is the solution vector $d_i$ for $i^{th}$ butterfly in iteration number $t$. Here, $f$ reflects the best answer currently discovered during the current iteration. The fragrance of $i^{th}$ the butterfly is represented by $fr_i$ and $r$ is a random number in [0, 1].

The local search phase can be represented as

$$d_i^{t+1} = d_i^t + \left(r^2 \times d_j^t - d_k^t\right) \times fr_i \tag{32}$$

where $d_j^t, d_k^t$ are the $j^{th}$ and $k^{th}$ butterflies from the solution space. If $d_j^t$ and $d_k^t$ belongs to the same swarm and $r$ is a random number in [0,1] then the equation becomes a local random walk.

Step 5: Termination condition: The process is repeated until optimal hyper-parameter selection is made. The XLNet receives the chosen hyperparameter value. Below Table 3 shows Butterfly Optimization Algorithm.

Good data fit was obtained, and the algorithm was only limited by max_iterations. It indicates that the algorithm has converged well to a solution in the given iterations, and the result is satisfactory. High accuracy indicates that the algorithm has learned well from the data and has made good predictions. The fact that we witness a successful convergence is a good demonstration of the algorithm's ability to optimize the parameters in an efficient manner to yield desired performance statistics.

## 4 Result and discussion

The findings of our study indicate the usefulness of the suggested methodology in strengthening the prediction of software change proneness even on the presence of imbalance in available data. Our findings highlight the challenges associated with data imbalance in predictive modeling, which can significantly impact predictive accuracy and robustness, and demonstrate improved predictive performance with the proposed method through a rigorous evaluation process. We utilized a SMOTE, Self-Attention, RoBERTaEmbeddings, Recursive Feature Elimination and XLNet with Butterfly optimization models for obtaining these achievements. Overheard from a conversation about recommendations, discusses the implications of our findings for software maintenance and evolution, and elaborates on how we can adopt holistic data driven approaches for dealing with data challenges in predictive modeling.

The test is run on a machine equipped with an Intel (R) core (TM) i5 4570 s CPU @ 2.90 GHz, *GB RAM, and the computer name SSM107.smg.local running Windows 64-bit. Acer is the system manufacturer using the PYTHON tool. Our experimental configuration includes two data centers with four hosts and a total RAM of 8 GB. The host has a bandwidth of 2800 Mbps.

### 4.1 Evaluation metrics

It has chosen many metrics to gauge how well change-proneness prediction models are doing. They have selected ROC area, accuracy, precision, recall, and f-measure for our investigation. The confusion matrix was primarily used to determine the true positive, true negative, false positive, and false negative for the majority of the measurements. To evaluate these findings, compute the accuracy, recall, precision, F1-score, FPR, FNR, MCC, and NPV indicators.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{33}$$

$$recall = \frac{TP}{TP + FN} \tag{34}$$

$$precision = \frac{TP}{2TP + FP} \tag{35}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{36}$$

**Table 3** Butterfly Optimization Algorithm

Define Objective Function f(X), where $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \ldots, \boldsymbol{X}_{\boldsymbol{dim}}\}$, dim = number of dimensions
Initialize the first population of n butterflies randomly: $\boldsymbol{X}_{\boldsymbol{i}} = (\boldsymbol{i} = 1, 2, ..., \boldsymbol{n})$
Define parameters: stimulus intensity (/), sensor modality (c), power exponent (a), and switch probability (p)
Calculate the fitness of each butterfly
Find the best butterfly and assume it is the elite (denoted as $\boldsymbol{f}^*$)
Set t = 0
while (t < iteration)
for each butterfly in the population
Calculate fragrance using Eq. (28)
Generate a random number r from [0, 1]
Move towards the best butterfly/solution using Eq. (29) if r < = p, else move randomly using equation ()
If the current butterfly's fitness $\boldsymbol{f}_{\boldsymbol{t}} < \boldsymbol{f}^*$
Update $\boldsymbol{f}^*$ to $\boldsymbol{f}_{\boldsymbol{t}}$
end for
Update the value of the power exponent (a)
Increment t by 1
end while
Return the elite butterfly ($\boldsymbol{f}^*$)

$$FPR = \frac{F_P}{T_N + F_P} \tag{37}$$

$$FNR = \frac{F_N}{T_P + F_N} \tag{38}$$

$$MCC = \frac{T_P T_N - F_P F_N}{\sqrt{(T_P + F_P) - (T_P + F_N)(T_N + F_P)(T_N + F_N)}} \tag{39}$$

$$NPV = \frac{T_N}{T_N + F_N} \tag{40}$$

TP signifies the true positive, FP the false positive, TN the true negative, and FN the false negative.

### 4.2 Dataset description

This work uses a benchmark dataset of 12 Android mobile apps from recent research to assess the performance of our strategy. For a description of these applications, readers may go to. The essential details of these applications were compiled in Table 2, which included the number of lines of code (# LOC), total commit instances (# TC), number of defective instances (# DC), number of clean instances (# CC), and ratio of defective instances (% DR). A new commit instance is considered flawed if it adds faults; otherwise, it is considered clean. These applications' code lines range from 9,506 to 275,637, indicating that their scales are distinct. The mobile app commit instances within the benchmark dataset are distinguished by a range of features across many scopes, including Diffusion, Size, and History [29].

Source code and dataset detailed link:

https://drive.google.com/drive/folders/1NmUwNU2mMWj4GZkW1XoXE6VojfHaDH-c?usp=share_link Table 4 shows the basic information of the 12 apps.

Table 4 describes basic information about the 12 different software projects include lines of code (LOC), the total number of test cases (TC), detected defects (DC), cyclomatic complexity (CC), and defect removal efficiency (% DR). The size of projects varies widely, with LOC between 9,506 and 275,637. Different projects also reported a different number of test cases and defects detected, hinting different testing and QA efforts. An industry-used metric for code complexity called cyclomatic complexity has values that go from 23 to 2,870, where increasing values indicate more complex codebases. We also indicate defect removal efficiency (DRE) from 14.0% to 40.4%, which illustrates the effectiveness of defect detection and removal processes in the projects.

Table 5 The statistical analysis of the dataset is summarized (mean, median, standard deviation, min and max values). By obtaining these statistics, we can better understand the central tendency, variance, and distribution of each variable.

Figure 4 showing accuracy performance of four machine learning models Ada Boost, Logit Boost, Random Forest (RF), and proposed model. Precision is a classification metric that refers to the fraction of positively predicted instances that are actually positive. This is especially relevant where reducing false positives is vital. The accuracy in predicting positive instances (i.e., the AUC) is highest for the proposed model and approaches a value of 0.98, hence demonstrating its apparent ability to predict positive values with minimal false positives. Ada Boost does pretty good with a precision close to 0.91, followed by Logit Boost with a precision above 0.87. Also, RF doesn't do that well, it has a precision around ~ 0.85 so it fails many more instance when classifying as positive than all the other models. This leads to a clear advantage of the proposed model in identifying positives more accurately resulting in it being a potential good fit for applications requiring higher precision of positive predictions. The comparative analysis also highlights the progressive improvement accomplished by each model over the previous one, with the proposed model clearly depicted as the superior model.

Figure 5 presents the recall performance of four machine learning models (Ada Boost, Logit Boost, and RF, in addition to a proposed model). Recall is one of the most used metrics in classification tasks and as you remember it is how many of the true positive instances you have correctly classified by your model. The Fig. 5 shows that the proposed model has a highest recall of almost 0.98 and outperform the other two models. Such is the indication that the proposed model yields less false negatives, and is better able to identify relevant instances. Next, Ada Boost has a recall of around 0.91, which demonstrates the great climate of the Ada Boost but is still outperformed by the proposed method. The recall

**Table 4** The basic information of the 12 apps

| Project | # LOC | # TC | # DC | # CC | % DR |
|---|---|---|---|---|---|
| Firewall | 77,242 | 1024 | 415 | 612 | 40.5% |
| Alfresco | 152,046 | 1003 | 215 | 791 | 21.4% |
| Sync | 275,634 | 207 | 63 | 148 | 30.1% |
| Wallpaper | 35,917 | 585 | 95 | 495 | 16.1% |
| Keyboard | 114,784 | 2975 | 818 | 2153 | 27.5% |
| Apg | 151,209 | 3787 | 1305 | 2477 | 34.4% |
| Secure | 98,764 | 2576 | 852 | 1728 | 33.2% |
| Facebook | 103,807 | 545 | 181 | 369 | 32.7% |
| Kiwix | 32,595 | 1372 | 351 | 1024 | 25.4% |
| Cloud | 115,167 | 3701 | 831 | 2871 | 22.5% |
| Turner | 30,946 | 165 | 24 | 142 | 14.1% |
| Reddit | 9,508 | 223 | 61 | 163 | 27.1% |

**Table 5** statistical analysis of the dataset

| Variable | Mean | Median | Standard Deviation | Min | Max |
|---|---|---|---|---|---|
| LOC | 99,620.25 | 103,802 | 72,548.63 | 9506 | 275,637 |
| TC | 1432.33 | 806.5 | 1361.65 | 164 | 3780 |
| DC | 461.17 | 284 | 417.54 | 23 | 1304 |
| CC | 866.92 | 611 | 694.85 | 141 | 2870 |
| DR | 27.67% | 27.05% | 7.33% | 14.0% | 40.4% |

of Logit Boost is a little more than 0.86, behind it is Random Forest around 0.83, which means it is not as effective when correctly identifying positive cases. The outcomes emphasize the proposed model's progressive and superior performance compared to conventional boosting and ensemble techniques. This lets id of the proposed approach to embodiment wherein increase in recall would lead to that of precision and vice versa, so that under recall preferred issues that is beneficial.

Figure 6 illustrates the F measure of different classification methods, such as AdaBoost, LogitBoost, Random Forest (RF), and our proposed method. The highest F measure shows that the proposed method achieved the best balance between precision and recall than other techniques. This is especially noteworthy since the F measure is key to assessing model performance on datasets that have an imbalanced distribution of classes, or where the cost of false positives and false negatives is just as high. This higher F measure also indicates that the proposed method can discover relevant cases with fewer errors, so it may be a good candidate for practical implementations. Moreover, it would immensely benefit the research community if you provide insights on the scope of the proposed method and what kind of limits it has with respect to diverse datasets or real-world applications.

Figure 7 presents the ROC area for the different classification methods, including the proposed approach, RF, LogitBoost, and AdaBoost. Proposed method achieved the highest ROC area covering most of the curve, indicating its best classify potential over negative and positive classes compare to any other techniques. This metric is essential because a greater area under the ROC curve (AUC) represents a better sensitivity and specificity of an optimized model at varied thresholds. Furthermore, the stability of the performance illustrated at different classification intervals further indicates that the method possesses consistent predictive power, which is a critical factor for the applicability of such a method in practice. Moreover, discussing the significance of the ROC area in practical use cases including the trade-offs in false-positive rates will offer practical insight on the efficacy of the technique and on useful directions for future improvements. The discussion of the experimental results can be greatly improved by further scrutinising the results.

Overall, the experimental results shown in Fig. 8 indicate the accuracy of different classification methods such as AdaBoost, LogitBoost, RF, and the new method. The accuracy for the proposed method is greater than 98%, which is significant improvement against the other techniques among these. This high accuracy indicates the comparatively efficiency of proposed method in classifying the instances correctly with minimum misclassification. To grasp the real consequences of these numbers, we have to look at the implications of the results on other performance measures like precision, recall, and F1-score. Having these metrics beside accuracy would give insight into the overall performance of

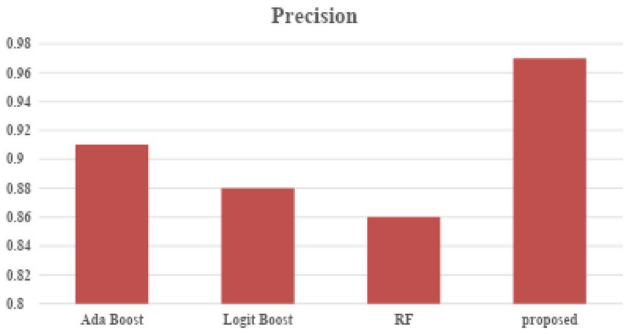



**Fig. 4** Precision compared with the existing method

the model. Finally, it is also important to mention the limitations or edge-cases for which the proposed method could underperform. These insights would help make the analysis more rigorous, as well as more useful to future research. This comprehensive assessment provides thrust to learn further Study of experimental results and thereal use of proposed method in practical applications.

The results shown in Fig. 9 show the metrics such as precision, recall, F measure, and accuracy before and after feature selection. Clearly, feature selection makes the model a lot better in terms of every metric measured. The Feature Selection results show a clear increase in both precision and recall scores. The tangible increase in the F measure indicates the effectiveness of the balance between precision and recall and supports the performance of the proposed method after finding the optimal feature. Moreover, the accuracy metric shows similar characteristic that is indicating that the model is not only being able to correctly classify positives, but are also spending few time being able to correctly classify negatives. It can be further investigated what features were selected and how they lead to an increase in performance. Knowing which predictors contributed the most to the model may inform us relevant studies and future usages. In addition to this, avoiding the expression of skepticism towards the results without disclosing the potential limitations linked with the feature selection strategy, including the possibility of disposing of informative features or the computational expense associated with this would create a more balanced approach to these results.

The ROC curve for our model is shown in Fig. 10. Therefore, the curve is plotted by using a true positive rate and a false positive rate at different thresholds. Our model achieves high performance, where the ROC curve shows that true positive rate (TPR) steadily increases while decreasing false positive rate (FPR) approaches to the upper-left corner indicating best distinguishability between the positive and negative instances. The area under the ROC curve (AUC) quantifies this performance in another way, and our model achieved an impressive [area = 1.00] value.

The fitness enhancement of the Butterfly Optimization Algorithm is presented in Fig. 11. Across the loops, there is a clear progression of fitness improvement, validating the efficacy of the algorithm to optimize the objective function. The reduction of test error with additional data demonstrates the validity of a framework, and the eventual convergence to an optimal solution demonstrates how something works in an iterative manner. The visualization offered some insight into the performance and convergence behavior of the algorithm, and you can see how well it worked at solving the optimization problem. Table 6 describes the efficacy comparison with previous state-of-the-art methods for proneness prediction.




**Fig. 5** Recall compared with the existing method

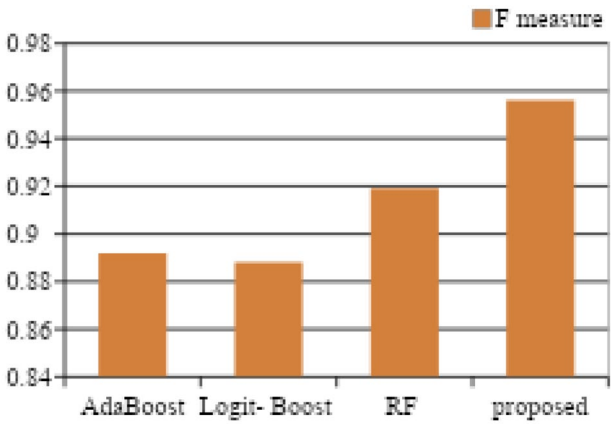



**Fig. 6** F measure compared with the existing method

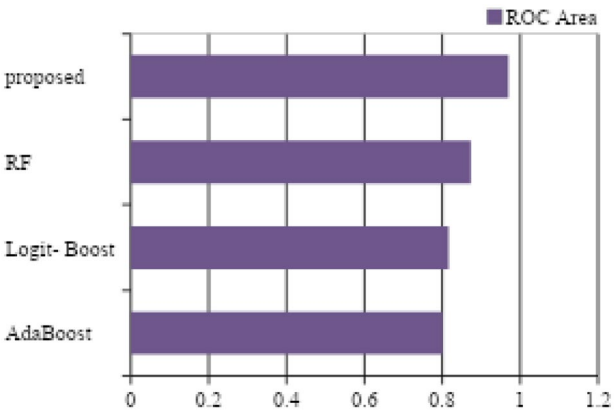



**Fig. 7** ROC Area compared with the existing method

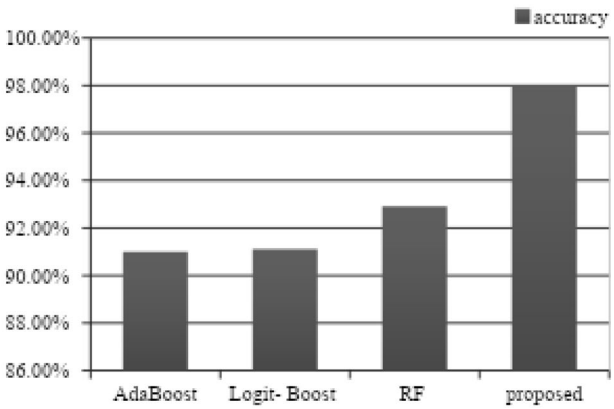



**Fig. 8** Accuracy compared with the existing method

## 5 Conclusion and future scope

We conclude that our model yields a robust way to overcome data imbalance problem while improving predictiveness with regard to identifying proneness to change in Android applications. Our proposed approach offers a comprehensive methodology that combines several cutting-edge techniques in a novel manner, such as SMOTE, Self-Attention, RoBERTa Embeddings, Recursive Feature Elimination, XLNet, and Butterfly optimization, resulting in a superior performance compared to existing models in terms of accuracy, precision, recall, and F-measure. Our hybrid

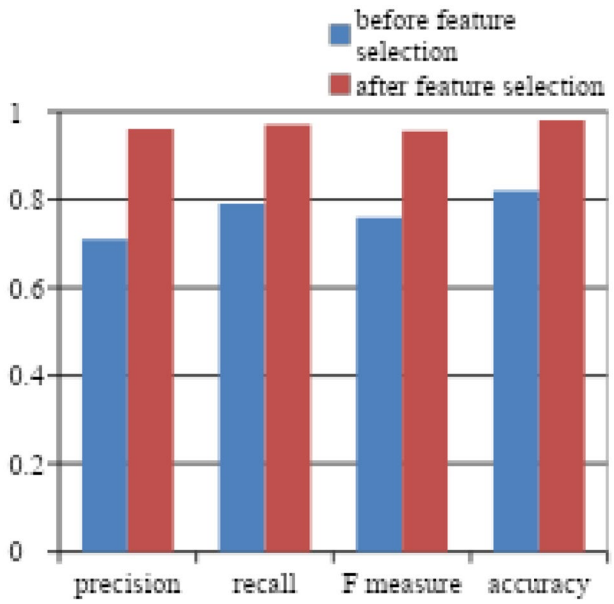



**Fig. 9** The Performance of before feature selection and after feature selection

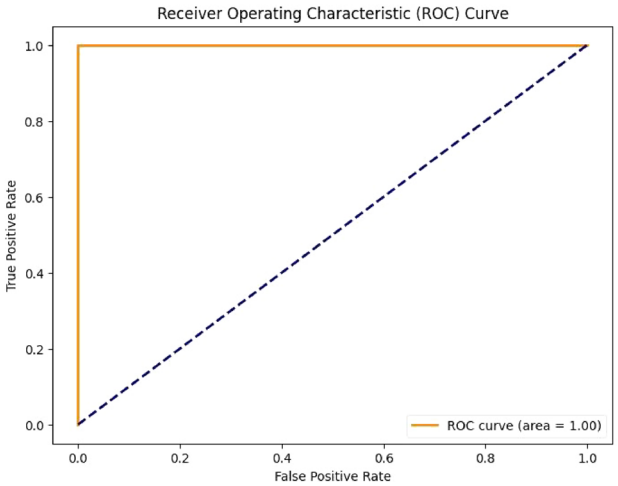



**Fig. 10** ROC Curve

approach yields superior performance in terms of accuracy (98%), precision (96%), recall (97%) and F-measure (95%), as demonstrated by the results. These results demonstrate that capitalizing on advanced approaches improves the prediction of software change proneness, and outline how our methodology can have a big positive impact on the process of software maintenance and evolution in Android applications. In future work, we would like to optimize more the structure of CNN model and find better way to combine the similarity measures, also, we intend to extract more features for change-proneness prediction with more advanced deep learning methods [32, 33].

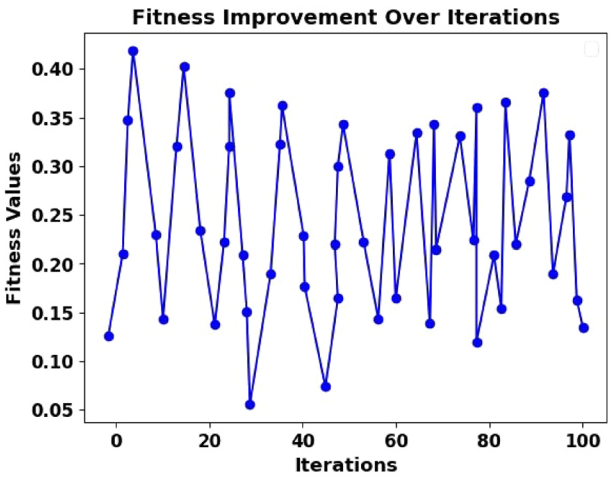



**Fig. 11** Fitness Improvement for BOA

**Table 6** Performance Comparison with Previous State-Of-The-Art Proneness Prediction Methods

| Reference | Author | Dataset | Methods | Metrics | | | |
|---|---|---|---|---|---|---|---|
| | | | | Accuracy (%) | Precision (%) | Recall (%) | F measure (%) |
| [31] | Catolino | Qualitas Corpus | Code Smell | – | 76 | 77 | 80 |
| [21] | Wattanakrieng-krai | line-level-defect-prediction | Model-Agnostic Technique | – | – | 0.62 | – |
| [22] | Ni | PROMISE | Cross-project defect prediction | – | – | 70 | – |
| [23] | Dalla Palma | IaC | Infrastructure-as-code | – | 78 | 67 | 73 |
| Proposed | | benchmark dataset | XLNet | 98 | 96 | 97 | 95 |

**Author contributions** Jyoti Patidar wrote the main manuscript text and Ruchika Malhotra guided and prepared figures and updated experiment. All authors reviewed the manuscript.

**Funding** No funding was received from any organization for conducting the study of the submitted work and preparation of this manuscript.

**Data availability** Data is provided as per request.

## Declarations

**Ethics approval and consent to participate** This article does not contain any studies with human participants and/or animals performed by any of the authors. There is no informed consent for this study.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

## References


1. Pritam N, et al. Assessment of code smell for predicting class change proneness using machine learning. IEEE Access. 2019;7:37414–25. https://doi.org/10.1109/ACCESS.2019.2905133.
2. Rasool G, Arshad Z. A lightweight approach for detection of code smells. Arab J Sci Eng. 2017;42(2):483–506. https://doi.org/10.1007/s13369-016-2238-8.
3. Gatrell M, Counsell S. The effect of refactoring on change and fault-proneness in commercial C# software. Sci Comput Program. 2015;102:44–56. https://doi.org/10.1016/j.scico.2014.12.002.
4. Li Z, Jing X-Y, Zhu X. Progress on approaches to software defect prediction. IET softw. 2018;12(3):161–75. https://doi.org/10.1049/iet-sen.2017.0148.
5. Malhotra R. A systematic review of machine learning techniques for software fault prediction. Appl Soft Comput. 2015;27:504–18. https://doi.org/10.1016/j.asoc.2014.11.023.
6. Paterson D, Campos J, Abreu R, Kapfhammer GM, Fraser G, McMinn P. An empirical study on the use of defect prediction for test case prioritization. in 2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST), Xi'an, China: IEEE, Apr. 2019, pp. 346–357. https://doi.org/10.1109/ICST.2019.00041.
7. Huang Q, Xia X, Lo D. Revisiting supervised and unsupervised models for effort-aware just-in-time defect prediction. Empir Software Eng. 2019;24(5):2823–62. https://doi.org/10.1007/s10664-018-9661-2.
8. Chowdhury S, The good, the bad, and the monstrous: predicting highly change-prone source code methods at their inception, 2024, arXiv: arXiv:2408.05704. https://doi.org/10.48550/arXiv.2408.05704.
9. de Carvalho SR, Farah PR, Vergilio SR. On the use of change history data to enhance class change-proneness prediction models". J Brazil Comput Soc. 2024;30(1):Art. No. 1. https://doi.org/10.5753/jbcs.2024.3782.
10. Godara D. A review of studies on change proneness prediction in object oriented software. Int J Comput Appl. 2014;105:3.
11. Van Koten C, Gray AR. An application of Bayesian network for predicting object-oriented software maintainability. Inf Softw Technol. 2006;48(1):59–67. https://doi.org/10.1016/j.infsof.2005.03.002.
12. Amoui M, Salehie M, Tahvildari L. Temporal software change prediction using neural networks. Int J Soft Eng Knowl Eng. 2009;19(07):995–1014. https://doi.org/10.1142/S0218194009004489.
13. Elish MO, Aljamaan H, Ahmad I. Three empirical studies on predicting software maintainability using ensemble methods. Soft Comput. 2015;19(9):2511–24. https://doi.org/10.1007/s00500-014-1576-2.
14. Melo CS, da Cruz MM, Martins AD, Matos T, da Silva Monteiro Filho JM, de Castro Machado J. a practical guide to support change-proneness prediction:," in Proceedings of the 21st International Conference on Enterprise Information Systems, Heraklion, Crete, Greece: SCITEPRESS - Science and Technology Publications, 2019, pp. 269–276. https://doi.org/10.5220/0007727702690276.
15. Zhu X, He Y, Cheng L, Jia X, Zhu L. Software change-proneness prediction through combination of bagging and resampling methods. J Software Evolu Process. 2018;30(12): e2111. https://doi.org/10.1002/smr.2111.
16. Koru AG, Liu H. Identifying and characterizing change-prone classes in two large-scale open-source products. J Syst Softw. 2007;80(1):63–73.
17. Romano D, Pinzger M. Using source code metrics to predict change-prone Java interfaces. in 2011 27th IEEE International Conference on Software Maintenance (ICSM), Williamsburg, VA, USA: IEEE, Sep. 2011, pp. 303–312. https://doi.org/10.1109/ICSM.2011.6080797.
18. Malhotra R, Khanna M. An empirical study for software change prediction using imbalanced data. Empir Software Eng. 2017;22(6):2806–51. https://doi.org/10.1007/s10664-016-9488-7.
19. Li W, Henry S. Object-oriented metrics that predict maintainability. J Syst Softw. 1993;23(2):111–22. https://doi.org/10.1016/0164-1212(93)90077-B.
20. Sikic L, Kurdija AS, Vladimir K, Silic M. Graph neural network for source code defect prediction. IEEE Access. 2022;10:10402–15. https://doi.org/10.1109/ACCESS.2022.3144598.
21. Wattanakriengkrai S, Thongtanunam P, Tantithamthavorn C, Hata H, Matsumoto K. Predicting defective lines using a model-agnostic technique. IIEEE Trans Software Eng. 2022;48(5):1480–96. https://doi.org/10.1109/TSE.2020.3023177.
22. Ni C, Xia X, Lo D, Chen X, Gu Q. Revisiting supervised and unsupervised methods for effort-aware cross-project defect prediction. IIEEE Trans Software Eng. 2022;48(3):786–802. https://doi.org/10.1109/TSE.2020.3001739.
23. Dalla Palma S, Di Nucci D, Palomba F, Tamburri DA. Within-project defect prediction of infrastructure-as-code using product and process metrics. IIEEE Trans. Software Eng., vol. 48, no. 6, pp. 2086–2104, Jun. 2022, https://doi.org/10.1109/TSE.2021.3051492.
24. Hassouneh Y, Turabieh H, Thaher T, Tumar I, Chantar H, Too J. Boosted whale optimization algorithm with natural selection operators for software fault prediction. IEEE Access. 2021;9:14239–58. https://doi.org/10.1109/ACCESS.2021.3052149.
25. Tumar I, Hassouneh Y, Turabieh H, Thaher T. Enhanced binary moth flame optimization as a feature selection algorithm to predict software fault prediction. IEEE Access. 2020;8:8041–55. https://doi.org/10.1109/ACCESS.2020.2964321.
26. Yuan Z, Chen X, Cui Z, Mu Y. ALTRA: cross-project software defect prediction via active learning and tradaboost. IEEE Access. 2020;8:30037–49. https://doi.org/10.1109/ACCESS.2020.2972644.
27. R. Yedida and T. Menzies, "On the Value of Oversampling for Deep Learning in Software Defect Prediction," Apr. 20, 2021, *arXiv*: arXiv:2008.03835.. http://arxiv.org/abs/2008.03835. Accessed 14 May 2024.
28. Khan F, Kanwal S, Alamri S, Mumtaz B. Hyper-parameter optimization of classifiers, using an artificial immune network and its application to software bug prediction. IEEE Access. 2020;8:20954–64. https://doi.org/10.1109/ACCESS.2020.2968362.
29. Zhao K, Xu Z, Yan M, Tang Y, Fan M, Catolino G. Just-in-time defect prediction for Android apps via imbalanced deep learning model," in Proceedings of the 36th Annual ACM Symposium on Applied Computing, Virtual Event Republic of Korea: ACM, Mar. 2021, pp. 1447–1454. https://doi.org/10.1145/3412841.3442019.
30. Jayaraman AK, Ananthakrishnan G, Trueman TE, Cambria E. Text-based personality prediction using XLNet," in Advances in Computers, vol. 132, Elsevier, 2024, pp. 49–65. https://doi.org/10.1016/bs.adcom.2023.08.002.

31. Catolino G, Palomba F, Fontana FA, De Lucia A, Zaidman A, Ferrucci F. Improving change prediction models with code smell-related information, May 26, 2019, *arXiv*: arXiv:1905.10889. http://arxiv.org/abs/1905.10889. Accessed 14 May 2024
32. Zhu X, Li N, Wang Y. Software change-proneness prediction based on deep learning. J Softw Evol Proc. 2022;34(4): e2434. https://doi.org/10.1002/smr.2434.
33. Liu C, Yang D, Xia X, Yan M, Zhang X. Cross-project change-proneness prediction," in 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan: IEEE, Jul. 2018, pp. 64–73. https://doi.org/10.1109/COMPSAC.2018.00017.

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

Journal of Superconductivity and Novel Magnetism (2025) 38:137
https://doi.org/10.1007/s10948-025-06973-0

RESEARCH

# Evidence of Surface Biasing and Multiferroic Properties in Ferroelectric (BNT)-Ferrimagnetic (LFO) Nanocomposites

**Bineet Kumar[1,2,3] · Lagen Kumar Pradhan[4] · Amar Dev[2] · Amrish K. Panwar[1] · Manoranjan Kar[2]**

Received: 23 May 2024 / Accepted: 12 April 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

**Abstract**
The lead-free composite (1-x)$Bi_{0.5}Na_{0.5}TiO_3$-(x)$LaFeO_3$ with chemical composition [$x$ = 0.00, 0.02, 0.04, 0.06, 0.08 and 1.00] has been prepared via the double step solid-state reaction method. The X-ray diffraction (XRD) patterns of the composites indicate its crystalline nature and indexed with rhombohedral (*R3c*) and orthorhombic (*Pbnm*) crystal symmetries. The Rietveld refinement of XRD patterns has been used to estimate the various crystal structure parameters. The microstructure of the composites indicates the non-homogeneous distribution of grains with different sizes. The dispersive nature of the frequency dependent dielectric properties is observed in the present composites. The temperature-dependent dielectric constant of the composites exhibits the relaxor behavior. The ferroelectric and magnetic properties of the composites describe its multiferroic behavior. The dielectric, ferroelectric, and magnetic properties of the present composites have been correlated with the multiferroic behavior.

**Keywords** Multiferroic · Rietveld refinement · Dielectric constant · Ferroelectric · Ferrimagnetic · Magnetic biasing

## 1 Introduction

Nowadays, various emerging technologies require different functional materials for its future growth and development. As per the material aspects, the physical and chemical properties of materials strongly depend upon the crystal structure [1]. Among all, the perovskite crystal structure ($ABO_3$: A and B are the cations with different radii, O: oxygen ion) has been given more attention due to their multifunctionality behavior [1–3]. There are large numbers of applications such as electronic devices, gas sensor, solid oxide fuel cell, and actuators. However, the environmental concern on lead-based compounds demands to developed lead-free perovskite materials for sustainable growth and development [4]. From last two decades, bismuth sodium titanate ($Bi_{0.5}Na_{0.5}TiO_3$) is one of the most promising lead-free perovskite compounds which have exceptional dielectric and ferroelectric properties [5]. Due to the presence of lone pair electrons on Bi similar to Pb, this provides the modified electrical polarization. Also, it has been reported that BNT exhibits relaxor ferroelectric properties with depolarization temperature ($T_d$) around 463 K; two different diffuse temperature peaks near ~553 K and ~800 K correspond to rhombohedral to tetragonal and tetragonal to cubic phase transition, respectively [5]. Despite this, BNT ceramic has several significant properties, but it has a few limitations which are hindrances for broader technological applications. These are low depolarization temperature along with high value of remanent polarization (38 kV/cm) [6]. To overcome these limitations along with the improvement of material properties, appropriate amount of substitutions/doping of foreign elements or solid solutions with other perovskite compounds ($BiFeO_3$, $BaTiO_3$, $SrTiO_3$, $NaNbO_3$, etc.) have been reported [7]. As per the earlier reports, the properties of BNT can be enhanced with the substitution of both donor and acceptor elements by removing the mobile oxygen vacancies. M. Lal et al. reported the structural, dielectric, and impedance properties of $Bi_{0.5}Na_{0.5}TiO_3$-based composites [8]. In addition, the rare earth orthoferrites ($ReFeO_3$) are taken in for consideration

✉ Manoranjan Kar
mano@iitp.ac.in

1 Department of Applied Physics, Delhi Technological University, Delhi 110042, India

2 Department of Physics, Indian Institute of Technology Patna, Bihar 801106, India

3 Department of Physics, Sheodeni Sao College, Kaler (Arwal), Magadh University, Bihar 824127, India

4 Department of Physics, Deogarh College, Sambalpur University, Sambalpur, Odisha 768108, India

Published online: 02 May 2025

due to their unique electrical, magnetic, and magnetoelectric properties [9, 10]. Due to the presence of two magnetic sublattices, $Fe^{3+}$ and $Re^{3+}$ lead to super-exchange interaction and show interesting magnetic properties. There are several multiferroic composites reported based on $ReFeO_3$ such as $NdFeO_3$–$PbTiO_3$, $LaFeO_3$–$PbTiO_3$, $BiFeO_3$–$SrTiO_3$, $NdFeO_3$–$SrTiO_3$, and $BiFeO_3$–$BaTiO_3$ [11–13]. Other than orthoferrite, electrical and magnetic properties of spinal ferrite (cobalt ferrite) doped with KNNS-BKT have been studied by P. Thakur et al. They observed the magnetoelectric coupling in the synthesized ceramics [14]. Also, the structural, morphological, electrical, and magnetic properties of $LaMn_{1-x}Fe_xO_3$ ceramics are reported by P. Thakur et al. [15]. It is interesting to note that the nanocrystalline (< 30 nm) BNT exhibits weak ferromagnetism. Hence, in this article, the aim is to understand the biasing of magnetic nanoparticle at the surface of nanocrystalline BNT. To understand the magnetic interaction between BNT and $LaFeO_3$ in BNT + LFO nanocomposite, the BNT with crystallite size above 30 nm has been prepared. The percentage of LFO is kept < 10% compared to BNT in BNT-LFO composite to obtain a nanocomposite as well as to avoid the interaction between LFO particles in the composite. Hence, only surface interaction between BNT and LFO can be observed. In this regard, the composites of [(1-x) BNT + (x) $LaFeO_3$ with $x = 0.00$, 0.02, 0.04, 0.06, 0.08, 1.00)] have been prepared, and their magnetic and dielectric properties are explored. The magnetic interaction between BNT and LFO has been clearly observed from the $\Delta H$ versus M plot. Hence, the present article opens a new window to realize the magnetism due to surface interaction in the nanocomposite.

## 2 Experimental Section

### 2.1 Synthesis of Multiferroic Composite

The polycrystalline samples of (1-x) $Bi_{0.5}Na_{0.5}TiO_3$ -(x) $LaFeO_3$ with $x = 0.00$, 0.02, 0.04, 0.06, and 0.08 composite were synthesized in the double step solid-state reaction method. In first step, lanthanum ferrite (LFO) was prepared through high energy planetary ball mill method in ethanol medium by taking respective amount (in molar ratio) of metal oxides of lanthanum oxide $La_2O_3$ (≥ 99.0% assay, Alfa Aesar) and iron oxide $Fe_2O_3$ (≥ 99.0% assay, Alfa Aesar). During the synthesis process, the ball mill was kept at 250 RPM (rotation per minutes) for 12 h, and the ball (stainless steel ball with 2 mm diameter) to powder ratio was taken at 20:1. In this stage, the main purpose to use the planetary ball mill is to prepare homogeneous mixture of all the metal oxides and carbonates. Furthermore, the mixtures were heated at 140 °C for 4 h to evaporate the ethanol from it and followed by the calcination process at 1100 °C. Similarly, required amount of molar ratio of metal oxides and carbonates of bismuth oxide $Bi_2O_3$ (≥ 99.0% assay, Alfa Aesar), sodium carbonate $Na_2CO_3$ (≥ 99.5% assay, Alfa Aesar), and titanum oxide $TiO_2$ (≥ 99.0% assay, Merck) were mixed to prepare bismuth sodium titanate (BNT). The mixture was heated at 1150 °C. The LFO and BNT powders were well grounded and followed by mixing of the both powders with different weight percentage by using the ball mill method to synthesize the multiferroic composite. The final mixtures (LFO + BNT) were uniaxially formed discs (pellet) of 10 mm diameter with the thickness around 1 mm by using the hydraulic press (Technosearch, India). The pellets (two pellets of each composition) were sintered at 1150 °C for 5 h with 2°/min ramping in a furnace with closed alumina boat covered by the powder of respective composition to prevent the evaporation of bismuth and sodium at high temperature. After that, the parallel surfaces of the remaining pellets were painted by the conducting silver paste to make electrode for electrical characterization.

### 2.2 Characterization

To determine the crystal structure of multiferroic composite, the X-ray diffraction (XRD) technique as well as Raman spectroscopy is being used. The powder X-ray diffraction measurements in the present work have been carried out at room temperature by using the Rigaku (TTRX-III) X-ray diffractometer with rotating anode. The characterization of the present materials has been carried out at 5 kW. The present instrument works in $\theta - 2\theta$ geometry which is well known as Bragg–Brentano geometry/focusing geometry in $2\theta$ range from 20 to 80° at a scan rate 2°/min with 0.02° scan step. Here, sample to detector distance and the source to sample distance remain equal. In the present measurements, $Cu\text{-}K_\alpha$ radiation of wavelength, $\lambda = 1.5418$ Å, has been used as an X-ray radiation source. Un-polarized Raman spectra has been measured in the backscattering geometry by using the micro-Raman spectrometer (Seki Technotron Corp., Japan) with the 514.5 nm to analyze the room temperature Raman spectra (100–1300 $cm^{-1}$) of the solid solutions. FESEM (field emission scanning electron microscopy, ZEISS, Germany) is being used to determine microstructure of the solid solutions. In dielectric study, different parameters such as dielectric constant and dielectric loss are being measured in a wide frequency range (100 Hz–1 MHz) at different temperatures between 40 and 400 °C. The frequency dependent electrical properties of materials can be described via the complex permittivity ($\varepsilon^*$), complex impedance ($Z^*$), electrical modulus ($M^*$), admittance ($Y$), ac conductivity, and dielectric loss or dissipation factor ($\tan\delta$). The P-E loop tracer with amplifier is being used to study the ferroelectric (hysteresis loop) properties of the composite, which have been recorded by using Radiant Ferroelectric Tester (precision

10 kV HV). The magnetic properties have been studied by recording the M-H loops with the help of Vibrating Sample Magnetometer (VSM) (Versalab, Quantum design).

## 3 Result and Discussion

### 3.1 Structural Analysis

The room temperature (RT) X-ray diffractograms in $2\theta$ range of 20–80° for the (1-x) BNT + (x) LFO composites are shown in Fig. 1. The presence of sharp peaks in XRD patterns confirms the formation of highly crystalline phase, and the samples are of polycrystalline nature. The allowed peaks are indexed with the crystal symmetries rhombohedral (*R3c*) and orthorhombic (*Pbnm*), corresponding to BNT and LFO, respectively, by using the International Centre for Diffraction Data (JCPDS: 360,340 for BNT & JCPDS: 37–1493 for LFO) [16, 17]. In addition, the XRD patterns of BNT and LFO do not show any secondary phases within the detection limit of XRD technique which is confirmed by the Rietveld refinement as discussed later. The Rietveld refinement of composite XRD patterns is carried out by assuming rhombohedral (*R3c*) and orthorhombic (*Pbnm*) as the composites have prepared by taking BNT and LFO. Also, it is noticed that the Bragg's peaks of BNT and LFO are closely located each other due to the neighboring space groups. Hence, one of the XRD peak ($x = 0.06$) is deconvoluted near 46.75° by using the Gaussian–Lorentzian function to distinguish the individual phases, as shown in Fig. 1b. Furthermore, the crystallite size of BNT has been estimated by using the Debye–Scherrer's formula, as mentioned below [18].

$$D = \frac{k\lambda}{\beta \cos\theta} \tag{1}$$

Here, $D$ is the average crystallite size, $k$ is the crystalline shape parameter, and its value is equal to 0.9 for assumption of spherical shape.$\beta$ is the full width half maxima, $\lambda$ is the wave length of X-ray, and $\theta$ is the Bragg's angle. The average crystallite size (in nm) of BNT is estimated to be approximately 38, 38, 45, 41, and 48 for $x = 0.00$, 0.02, 0.04, 0.06, and 0.08 in (1-x)BNT + (x)LFO, respectively. It indicates the random variation of average crystallite size of the BNT in BNT-LFO composite.

Furthermore, XRD patterns of the composites have been simulated by using the Rietveld refinement technique with the double phase model (*R3c* + *Pbnm*) to estimate the various crystal structure parameters [19]. The refined XRD patterns are shown in Fig. 2. The Fullprof suit program has been used for Rietveld refinement. During the Rietveld refinement, some of the parameters were kept fixed such as zero correction parameters, scale factor, and atomic fractional coordinates (x, y, z) of different elements. The pseudo-Voigt function has considered for the refinement of XRD patterns [20]. In addition, the different angles among crystal axes in BNT are taken as $\alpha = \beta = 90°$, $\gamma = 120°$ whereas $\alpha = \beta = \gamma = 90°$ for LFO. The refined XRD patterns are indexed with the *R3c* crystal symmetry. This has been considered due to the similar Bragg's positions of BNT as well as LFO. The various refined parameters are enlisted in Table 1. The lattice parameters for lanthanum ferrite are little higher compared to the literature values. It could be due to limitation of analysis of XRD patterns for the presence of small percentage of crystal symmetry in the composite. It is clearly observed that the refined lattice constants of both *R3c* and *Pbnm* crystal structure of all the composites are found to be similar. Also, the reliability factors are in range of accepted values as per the literatures. The Rietveld refinement of XRD patterns along with crystal structure parameters indicates the formation of composites with the presence of both crystal phases without any secondary phases within the limitation of XRD technique. In addition, the different phase percentages (*R3c* & *Pbnm*) in the composites are calculated from the Rietveld refinement method. The various calculated values are found to be as 97.39, 93.27, 87.99, and 75.11 for *R3c* and 2.61, 6.23, 12.01, and 24.89 for *Pbnm* with respect to $x = 0.02$, 0.04, 0.06, and 0.08 in (1-x)BNT + (x)LFO composite.

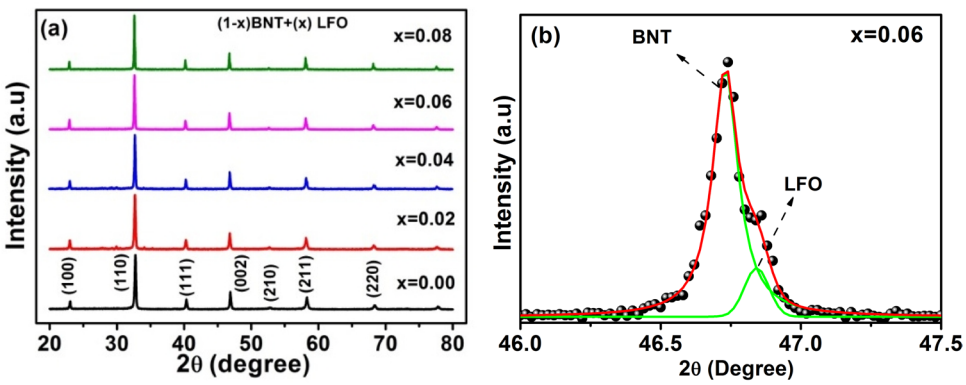



**Fig. 1** **a** The XRD patterns of (1-x)BNT + (x)LFO with $x = 0.00$, 0.02, 0.04, 0.06, and 0.08 and **b** enlarged version of peak around 46.75° of $x = 0.06$

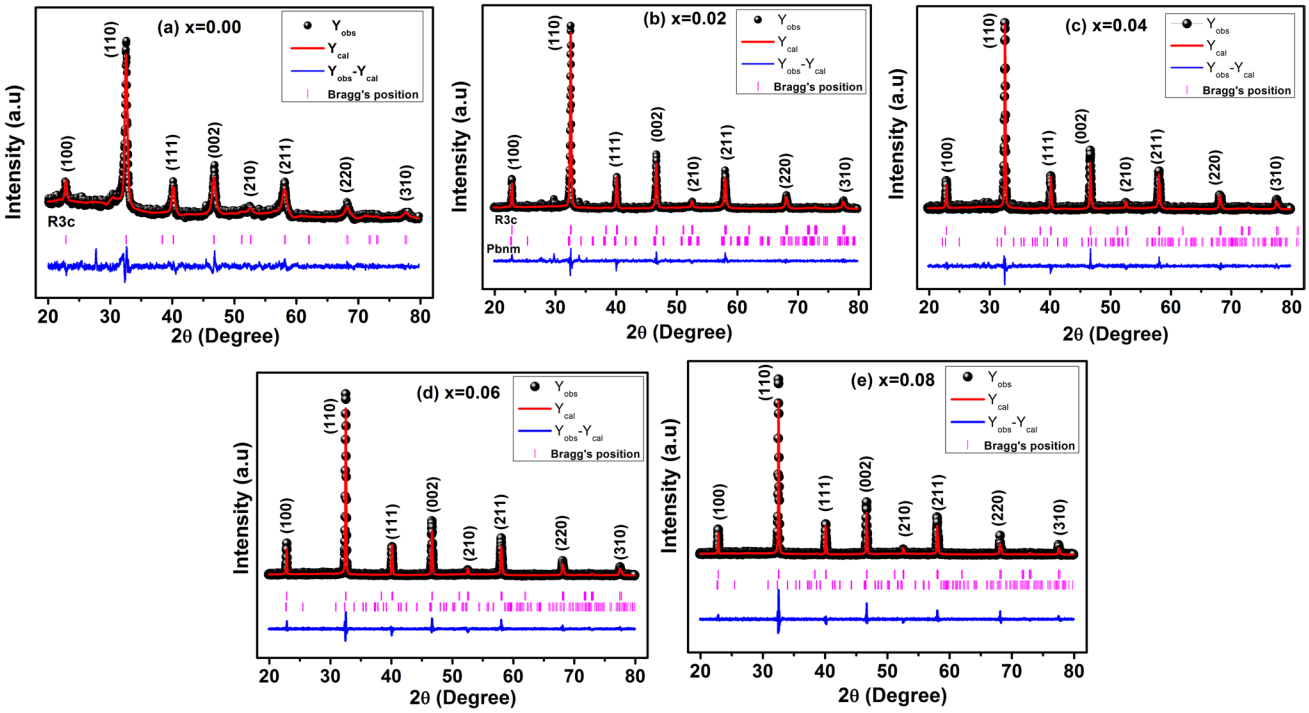



**Fig. 2** The Rietveld refinement of XRD patterns of (1-x)BNT+(x)LFO with x = 0.00,0.02,0.04,0.06 &0.08

**Table 1** Various structure parameters calculated from the Rietveld refinement of XRD patterns of (1-x)BNT+(x)LFO with x = 0.00,0.02,0.04,0.06 &0.08

| Composition (1-x)BNT-(x)LFO) | $\chi^2$ | *Rp* | *Rwp* | *Rexp* | Lattice parameters *a* = *b*, *c* (Å) *R3c* | Lattice parameters *a*, *b*, *c* (Å) *Pbnm* |
|---|---|---|---|---|---|---|
| $x = 0.00$ | 1.77 | 7.71 | 9.97 | 7.50 | 5.450(1), 13.468(7) | – |
| $x = 0.02$ | 2.28 | 13.8 | 18.6 | 11.49 | 5.511(3), 13.470(5) | 5.537(8), 5.569(2), 7.817(0) |
| $x = 0.04$ | 2.52 | 16.3 | 16.5 | 10.95 | 5.505(8), 13.472(5) | 5.864(5), 5.687(6), 7.839(1) |
| $x = 0.06$ | 2.44 | 13.5 | 17.8 | 11.41 | 5.507(6), 13.471(4) | 5.788(2), 5.289(1), 7.841(6) |
| $x = 0.08$ | 2.28 | 19.0 | 15.1 | 11.10 | 5.501(5), 13.491(4) | 5.799(1), 5.277(2), 7.843(8) |

## 3.2 Microstructural Analysis

In addition to crystal structure, the microstructure of the ceramic influences on various physical properties such as dielectric behavior and ferroelectric properties. Therefore, microstructures of the composites have been explored by the FE-SEM technique. Typical FE-SEM micrographs of the (1-x) BNT + (x) LFO with $x = 0.02$ and 0.08 nanocomposites are shown in Fig. 3 to investigate its microstructure. The FESEM micrographs show the random distribution of grains with different sizes. It can be correlated with the present synthesis process, i.e., solid-state reaction (ball mill method) of two different types of compounds (LFO & BNT) having different crystal structures. The overall particle size of the present composites is found to be in the order of micro-meter range which is good for the electrical properties prospective.

As the microstructures are not much different between $x = 0.02$ and $x = 0.08$, it is expected that the variation of dielectric constant will be due to variation of filler percentage which is discussed later.

## 3.3 Frequency Dispersion of Dielectric Constant

The variation of dielectric constant with frequency at different temperatures (50 °C, 100 °C, 150 °C, 200 °C, 250 °C, 300 °C, & 400 °C) for the composites (1-x) BNT + (x) LFO ($x = 0.00$, 0.02, 0.04, 0.06, and 0.08) is shown in the Fig. 4. The dielectric constant of ceramic strongly depends upon the presence of various types of polarizations such as dipolar polarization, ionic polarization, electronic polarization, and interfacial polarization [21]. As per the earlier experimental reports, dipolar and interfacial polarizations are more prominent at

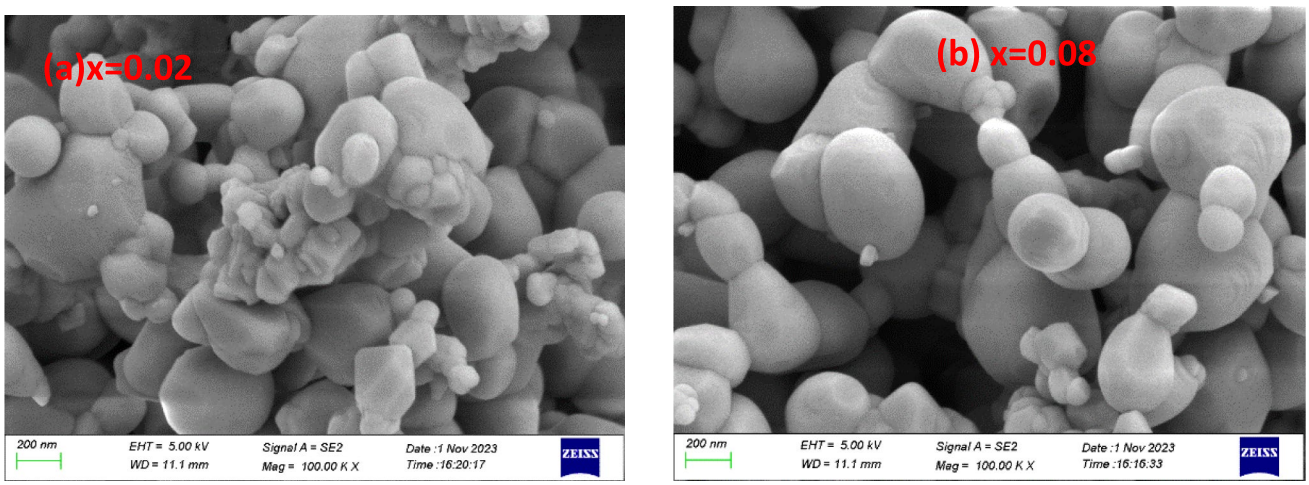



**Fig. 3** The microstructures of (1-x)BNT + (x)LFO composites

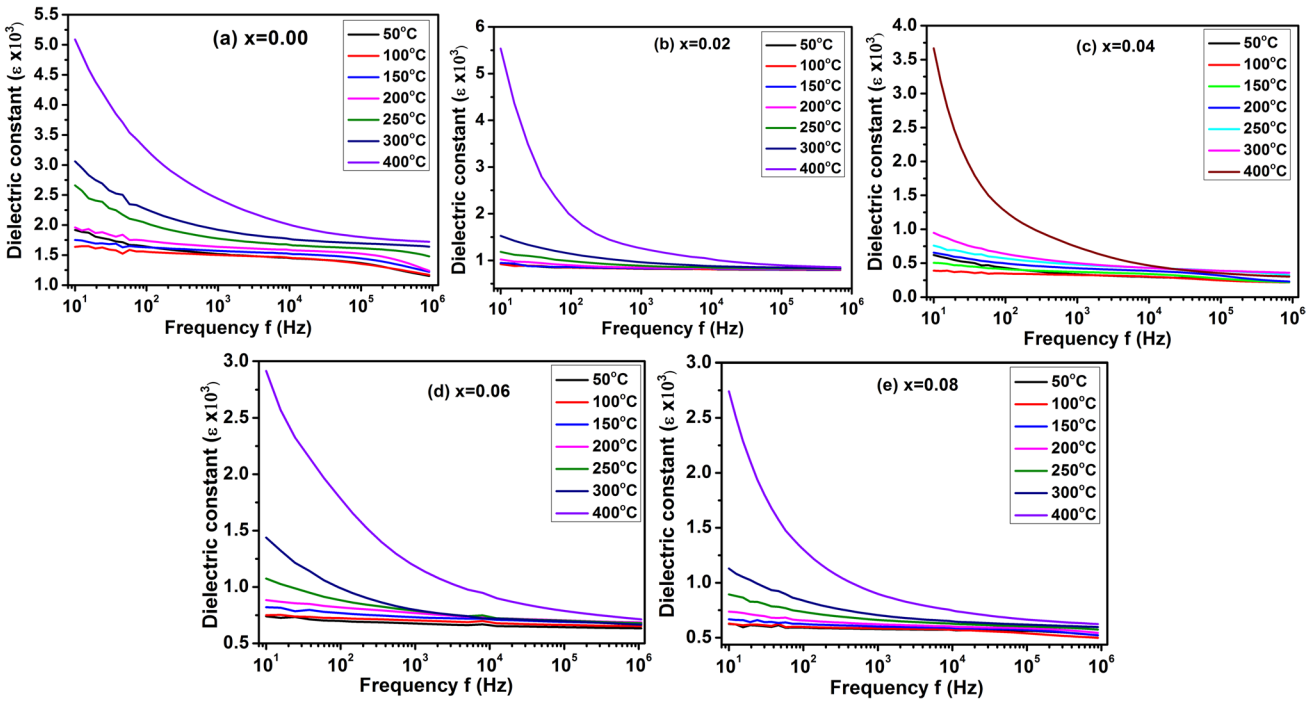



**Fig. 4** The frequency dependent dielectric constant of (1-x)BNT+(x)LFO with x = 0.00,0.02,0.04,0.06 & 0.08

the low frequencies (< ~ $10^3$ Hz). Dipolar polarization in the material occurs due to the orientation of microscopic dipoles. Similarly, interfacial polarization arises due to presence of charges in the grain boundaries or material interfaces [21]. Therefore, the contribution from dipolar and interfacial polarization on dielectric reduces rapidly with the increase in the frequency which leads to a decrease in the overall dielectric constant of the ceramic. In the present composite, two ceramics with different crystal structures (BNT and LFO) present with different proportions such as (1-x)BNT + (x)LFO with $x = 0.00$, 0.02, 0.04, 0.06, and 0.08. As per the literature, bismuth sodium titanate (BNT) exhibits good piezoelectric properties with higher dielectric value, and lanthanum ferrite (LFO) shows moderate value of both dielectric and magnetic properties. From the present experimental results, it has been observed that dielectric constant value of the nanocomposite is gradually decreasing with the increase of LFO mole fraction in place of BNT which is similar to other reported nanocomposite. In all the nanocomposites, the dielectric constant decreases with increasing in frequencies and attains nearly

constant value at higher frequency (Fig. 4). This happens due to the major contribution of ionic and electronic polarizations to overall dielectric constant. In addition, the variation of dielectric constant with frequency can be correlated with the Maxwell–Wagner-type interfacial polarization theory along with Koop phenomenological model [22]. From 10 to $10^3$ Hz, there is rapid reduction of dielectric constant which can be related to dipole relaxation process. It is due to the free charge contribution, which is a common phenomenon in the dielectric materials. However, dielectric constant attains almost a saturation value at higher frequency (~ $10^3$–$10^6$ Hz) due to the hopping mechanism. In the present composite, different charge carriers are hopping in the respective cations, i.e., electrons hopping and holes hopping [23]. It is well known that dipole oscillations cannot follow the ac electric field after certain higher frequency which leads to reduce the dielectric constant. In addition, the dielectric constant of the nanocomposite significantly increases with the rise of the temperature, as shown in the Fig. 4. It happens due to the thermally stimulated polarization process with the application of the electric field which enhances the dielectric constant of the nanocomposite. In the present observation, the dispersion behavior of dielectric constant with frequency of nanocomposite is higher at lower frequency range as compared to bismuth sodium titanate. It is due to the presence of interfacial charge carrier in nanocomposite which response decreases rapidly with the respect to electronic dipole at higher frequency. Hence, the dielectric constant reduces with further increase of the frequency [24]. In addition, the variation of dielectric constant with frequency is similar in all compositions. But the dielectric constant is almost comparable at all temperatures for the composition $x = 0.02$ near the frequency ~ $10^6$ Hz (Fig. 4b). This behavior is similar for all the samples, but difference of between two temperatures is somewhat less for $x = 0.02$. It is due to extrinsic property of the sample which affects the dielectric properties of the ceramics but not due to the intrinsic property.

Furthermore, variation of dielectric loss tangent with frequency plots at different temperatures is shown in Fig. 5. It is observed that the loss tangent value is increasing with the temperature for all the nanocomposites which is similar to other reported nanocomposites in the literature. In addition, the loss tangent of the nanocomposites increases with the increase in the mole fraction of LFO in BNT. It is due to higher leakage current behavior of LFO which is also clearly observed in ferroelectric properties as described below. Also, the dispersion behavior of loss tangent with frequency is quite similar to the dielectric constant, as shown in Fig. 4. However, the present nanocomposite shows slight higher value of loss tangent value as compared to bismuth sodium titanate (BNT) due to the presence of lanthanum ferrite (LFO). This behavior mainly depends upon the microstructure of the different phases in the composite along with the density of the pellet and mole percentage of individual phases in the composite. Hence, variations are observed with the composition, but it is within the expected errors for tangent loss measurement.

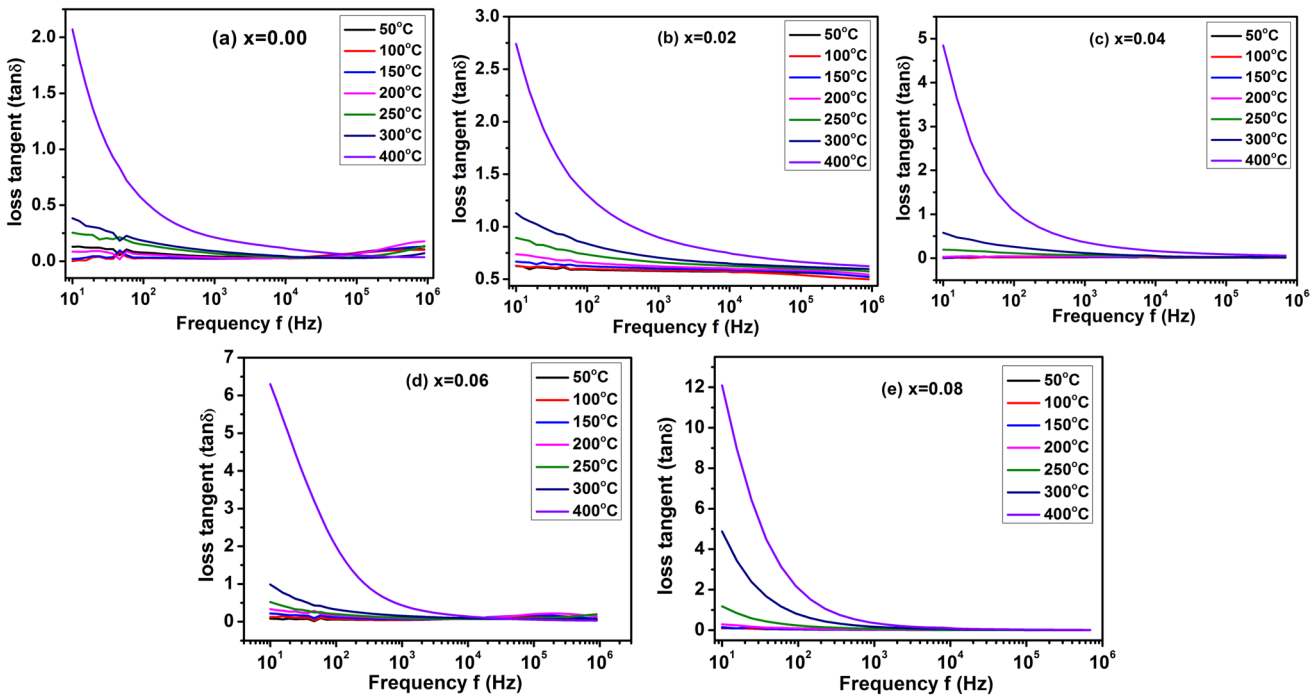



**Fig. 5** The frequency dependent loss tangent of (1-x)BNT+(x)LFO with x = 0.00,0.02,0.04,0.06 & 0.08

Furthermore, temperature-dependent dielectric constant ($\varepsilon$) plots at different frequencies (1 kHz, 10 kHz, 50 kHz, 100 kHz, 500 kHz, and 1 MHz) for the present composites are shown in Fig. 6a–e. The presence of multiple dielectric peaks has been observed in $\varepsilon \sim T$ curve of the bismuth sodium titanate along with dispersion with frequency, as shown in Fig. 6a, which indicates the behavior of the relaxor ferroelectric [25]. The peak corresponding to lower temperature is attributed to the ferroelectric (FE) to relaxor antiferroelectric (RAFE) phase transition, and the higher one related to the RAFE to paraelectric (PE) phase transition [26]. It is clearly observed that the FE-RAFE transition displays the frequency dependent dispersive behavior similar to the relaxor ferroelectric. The FE-RAFE (depolarization temperature: $T_d$) and RAFE-PE (maximum temperature: $T_m$) phase transition temperatures are observed, respectively, around 250 °C and 320 °C at 10 kHz. However, the intensity of distinct dielectric peaks reduces with the addition of LFO in BNT. Basically, the relaxor behavior of BNT is due the presence of polar nanoregions (PNRs). The addition of LFO in BNT is probably affecting the polar nanoregions of BNT during the formation of composites. It is also observed that the relative permittivity of the composite increases with the addition of LFO up to $x = 0.06$. This result can be correlated with the interfacial polarization due to presence of multi phases in the composite.

### 3.4 Ferroelectric Properties

The ferroelectric hysteresis (*P-E*) loops of the composites are shown in Fig. 7a–c. The different characteristic parameters such as maximum polarization ($P_{max}$), remanent polarization ($P_r$), coercive field ($E_c$), and loop area ($A$) in the P-E loop represent the ferroelectric properties of nanocomposites. In the present work, *P-E* loop of the nanocomposites is measured in the range of applied electric field 40 kV/cm. The various characteristic parameters are calculated from ferroelectric hysteresis loops as enlisted in the Table 2. It is observed that the ferroelectric loops are not well saturated within 40 kV/cm. Therefore, the maximum polarization ($P_{max}$) is taken into consideration and found that it decreases from 3.14 to 0.47 μC/cm$^2$ with the increase of mole fraction of LFO in BNT from $x = 0.00$ to $x = 0.08$. Similarly, the remanent polarization ($P_r$) increases for the composition $x = 0.00$ to $x = 0.04$, i.e., 0.56 to 0.91 μC/cm$^2$. After that, it decreases for the composition $x = 0.06$ and 0.08 due to the conductive nature of the LFO. In addition, the area of the hysteresis loop increases with LFO. It is assumed that the observation is due to increase in electrical conductivity which is arising by LFO incorporation in BNT matrix. Hence, it leads to an increase in the leakage current of the composite considerably. Similar result has been reported by other researchers [27]. Hence, the electric dipoles present in the composites are unable to hold the electric field as the material has free electron

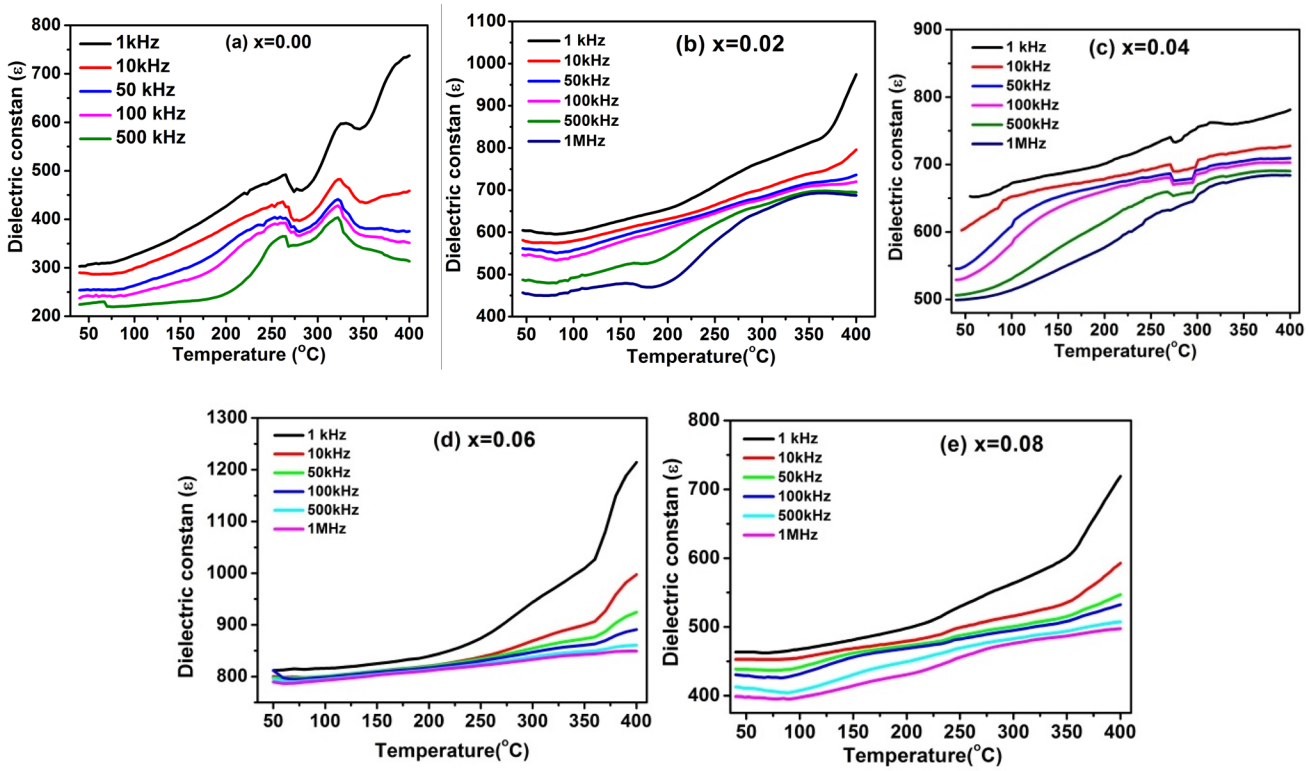



**Fig. 6** Temperature variation of dielectric constant of (1-x)BNT+(x)LFO with x = 0.00,0.02,0.04,0.06 & 0.08

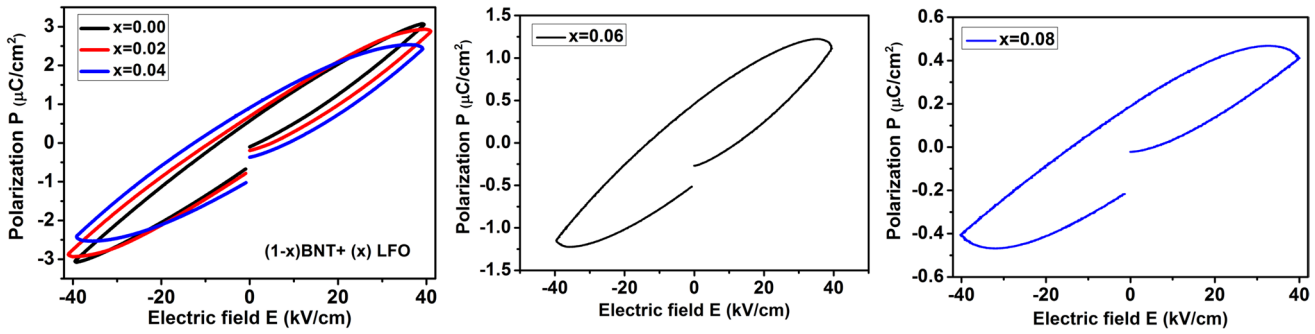



**Fig. 7** The ferroelectric hysteresis loops of (1-x)BNT+(x)LFO with x = 0.00,0.02,0.04,0.06 & 0.08

**Table 2** Various characteristic parameters estimated from the ferroelectric hysteresis loops

| Composition (1-x)BNT+(x)LFO) | Maximum polarization ($P_{max}$) μC/cm$^2$ | Remanent polarization ($P_r$) μC/cm$^2$ | Coercive field ($E_r$) kV/cm |
|---|---|---|---|
| $x = 0.00$ | 3.14(1) | 0.56(1) | 6.94(0) |
| $x = 0.02$ | 2.89(7) | 0.70(1) | 9.06(3) |
| $x = 0.04$ | 2.43(4) | 0.91(7) | 12.67(0) |
| $x = 0.06$ | 1.24(0) | 0.47(0) | 13.42(4) |
| $x = 0.08$ | 0.47(3) | 0.19(0) | 13.87(1) |

transport. The increase of conductivity due to addition of LFO in BNT matrix could be due to interaction of BNT and LFO at their interfaces in the BNT-LFO nanocomposite. Hence, the ferroelectric properties of nanocomposites are decreased compared to that of BNT. In addition, the remanent polarization and coercive field of composite are also increasing with respect to increase of LFO [28, 29]. Overall, the selection of good candidate for multiferroic behavior of the composite needs optimum ferroelectric as well as magnetic behavior. Hence, the magnetic behavior of the nanocomposites is explored and discussed in this article.

### 3.5 Magnetic Properties of (1-x) BNT + (x) LFO Nanocomposites

The room temperature magnetic hysteresis (*M-H*) loops of the composites [(1-x) BNT + (x) LFO] with $x = 0.00$, 0.02, 0.04, 0.06, and 0.08 are shown in Fig. 8. The maximum magnetic field of 30 kOe is applied for recording the M-H curves. The various magnetic characteristic parameters such as coercivity ($H_c$), remanent polarization ($M_r$), and maximum magnetization ($M_{max}$) are enlisted in Table 3 except $x = 0.00$ (i.e., bismuth sodium titanate). BNT exhibits weak ferromagnetic with diamagnetic behavior. From the information mentioned in Table 3, it is clearly observed that the magnetic behavior of the composite is gradually increasing with the

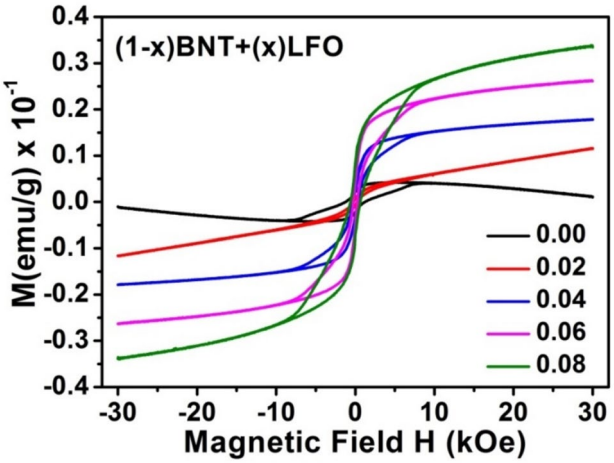



**Fig. 8** Magnetic hysteresis loops of (1-x)BNT+(x)LFO with x = 0.00,0.02,0.04,0.06 & 0.08

increase of $LaFeO_3$ mole fraction in $Bi_{0.5}Na_{0.5}TiO_3$. As per the literature, LFO exhibits antiferromagnetism along with weak ferromagnetic behavior [30]. The magnetic behavior of LFO depends on the bond angle between Fe–O-Fe. The bond angle between Fe–O-Fe is ~ 180° for ideal perovskite structure which shows perfect antiferromagnetic ordering. However, LFO exhibits orthorhombic crystal structure (i.e., distorted structure: orthorhombic) which leads to reduce the bond angle. Due to this, the magnetic spin associated with the linkage (Fe–O-Fe) deviates from the proper anti-alignment and develops weak ferromagnetic behavior (mostly canted spin).

The inter-particle interaction in the magnetic samples can be understood by the $\Delta H$ versus $M$ plot. It is an easy method compared to other tedious methods exits in the literature [31–33]. A ferromagnetic material exhibits a not distorted M-H loops. Hence, it exhibits a smooth bell shape which indicates a long range ferromagnetic ordering rather having inter-particle biasing. However, a composite shows kinks in the bell shape in the $\Delta H$ versus $M$

**Table 3** Various magnetic parameters of the composites calculated from M-H loops. Errors are given in the bracket

| (1-x) BNT + (x) LFO composite | Coercivity ($H_c$) (kOe) | | Remanent magnetization ($M_r$) (emu/g) | | Maximum magnetization ($M_{max}$) (emu/g) |
|---|---|---|---|---|---|
| | $H_{c+}$ | $H_{c-}$ | $M_{r+}$ | $M_{r-}$ | |
| $x = 0.02$ | 251.49(4) | 240.21(2) | 0.007(0) | 0.007(6) | 0.115(2) |
| $x = 0.04$ | 272.61(5) | 282.66(2) | 0.026(7) | 0.029(5) | 0.178(2) |
| $x = 0.06$ | 296.82(8) | 242.71(5) | 0.046(7) | 0.049(3) | 0.262(4) |
| $x = 0.08$ | 614.27(1) | 621.27(0) | 0.086(2) | 0.071(0) | 0.337(1) |

plot. The magnitude of kink gives idea about the strongness of magnetic interaction between particles in the composite or weak magnetic materials. Also, it depends on saturation magnetic field of different phases in the composite, i.e., if the saturation magnetic fields of different materials in a composite are different, the biasing of magnetic materials on each other will be different at the different applied magnetic fields. These assumptions explain clearly the magnetic behavior of the present composites. Let us consider the $\Delta H$ versus $M$ plot in Fig. 9, it exhibits clear distinct kinks which suggest the interaction between magnetic particles. Also, it is expected that the interaction at the surfaces of BNT and LFO. Mostly, it happens due to the presence of weak ferromagnetism in the sample which has been raised due to surface anisotropy and uncompensated spin of nanocrystalline LFO. It is interesting to note that the behavior is continuing with small variation of LFO (< 10%) in the (1-x) BNT + (x) LFO nanocomposites (Fig. 9a–d). In this percentage, it is expected that the LFO particles are far away to have self-magnetic biasing to each other. Hence, it is assumed that the surface of LFO is interacting with the BNT and creating a biasing at the surface even though the BNT size is greater than its critical crystallite size for weak ferromagnetic property. In a summary, a magnetic biasing behavior has been observed in the nanocomposite of (1-x) BNT + (x) LFO. Hence, both the ferroic orders are observable in the (1-x) BNT + (x) LFO nanocomposite. In the present composite, the ferroelectric ordering is primarily responsible by BNT, and weak ferromagnetism is primarily due to LFO and magnetic biasing at the surface between BNT and LFO which is distributed in 0–3 structure.

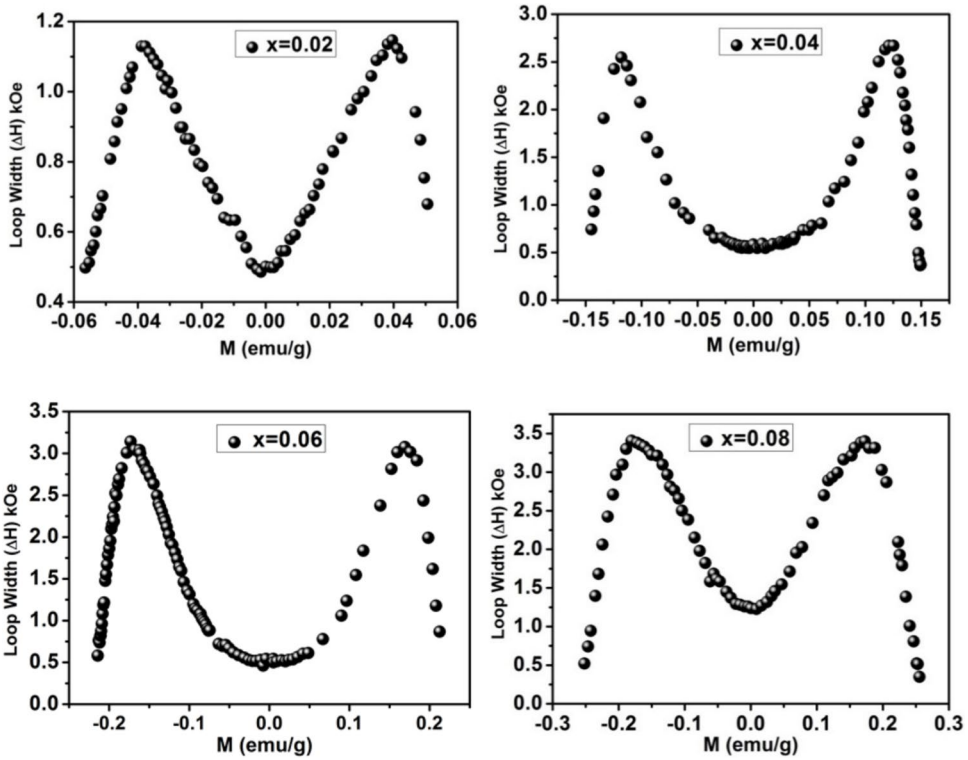



**Fig. 9** Magnetic hysteresis loop width vs magnetization (ΔH~M) of (1-x)BNT+(x)LFO with x = 0.00,0.02,0.04,0.06 & 0.08

## 4 Conclusion

Lead-free multiferroic composites (1-x)$Bi_{0.5}Na_{0.5}TiO_3$-(x) $LaFeO_3$ with $x = 0.00, 0.02, 0.04, 0.06$, and 0.08 are successfully synthesized by the two-step solid-state reaction method. From the XRD pattern analysis, it is confirmed the absence of any other secondary phases except rhombohedral (i.e., $Bi_{0.5}Na_{0.5}TiO_3$) and orthorhombic (i.e., $LaFeO_3$) crystal symmetries. The addition of lanthanum ferrite (LFO) in bismuth sodium titanate (BNT) reduces the overall dielectric and ferroelectric properties due to increase of iterant electrons at the interfaces which lead to increase the leakage current in the composite compare to BNT. The temperature-dependent dielectric constant exhibits the relaxor ferroelectric nature. Also, the magnetic study of the composites reveals the presence of weak ferromagnetic behavior. Magnetic biasing has been evidenced from the $\Delta H$ versus $M$ plot which is derived from M-H loops. Hence, magnetic interaction at the grain surfaces is assumed from the $\Delta H$ versus $M$ plots. The present report provides the basic understanding of physical properties of (1-x) $Bi_{0.5}Na_{0.5}TiO_3$ -(x) $LaFeO_3$ composite, and this composite can be considered a multiferroic composite. In summary, the present results open a window to modify the multiferroic properties of the lead-free composite for its practical applications.

**Acknowledgements** We extend our gratitude to the Department of Physics, Indian institute of Technology Patna for overall supporting of the current research work.

**Author Contributions** B. Kumar: Conceptualization, investigation, data analysis and writing of original manuscript, L. K. Pradhan & A. Dev: Data analysis, editing & software operation, A. K. Panwar & M. Kar: Supervision and manuscript correction.

**Data Availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing interests** The authors declare no competing interests.

## References

1. De, A., Ranjan, R.: Large structural heterogeneity in submicrometer $BaTiO_3$ revealed via $Eu^{+3}$photoluminescence study. J. Appl. Phys. **128**, 124104 (2020)
2. Thakur, P., Sharma, N., Pathak, D., Sharma, P., Kishore, K., Dhar, S.: Lal, M: State of art review on smart perovskites material: properties and applications. Emargent Mater. **7**, 667–694 (2024)
3. Thakur, P., Kumari, S., Chaudhary, S., Sharma, N., Lal, M.: Multifunctional tuning of structural, dielectric and magnetic properties of Ti-doped $BaMnO_3$ ceramics. Emergent Mater. **7**, 1733–1752 (2024)
4. Adhikary, G.D., Ranjan, R.: Abrupt change in domain switching behavior within tetragonal phase regime of (x)Na1/2Bi1/2TiO3-(1–x)K1/2Bi1/2TiO3. J. Appl. Phys. **128**, 204102 (2020)
5. Aksel, E., Forrester, J.S., Kowalski, B., Jones, J. L., Thomas, P.A.: Phase transition sequence in sodium bismuth titanate observed using high-resolution x-ray diffraction. Appl. Phys. Lett. **99,** (2011). https://doi.org/10.1063/1.3664393
6. Suchanicz, J., Roleder, K., Kania, A., Handerek, J.: Electrostrictive strain and pyroeffect in the region of phase coexistence in $Na_{0.5}Bi_{0.5}TiO_3$. Ferroelectrics **77**, (1988). https://doi.org/10.1080/00150198808223232
7. Pasternak, M.P., Xu, W.M., Rozenberg, G.K., Taylor, R.D. :Electronic, magnetic and structural properties of the $RFeO_3$ antiferromagnetic-perovskites at very high pressures. Mater. Res. Soc. Symp. Proc. **718**, (2002). https://doi.org/10.1557/PROC-718-D2.7
8. Lal, M., Thakur, P., Thakur, P., Sharma, N., Sharma, P., and Shukla, A. K.: Rietveld refined structural , dielectric and impedance properties of lead free 1-x $(K_{0.4}Na_{0.6}Na_{0.96}Sb_{0.04}O_3)$-x$(Bi_{0.5}K_{0.5}TiO_3)$ $(0 \leq x \leq 0.07)$ . J. Mater. Sci. Mater. Electron. **34**, 1188 (2023). https://doi.org/10.1007/s10854-023-10580-x
9. Wang, Z.Q., Lan, Y.S., Zeng, Z.Y., Chen, X.R., Chen, Q.F.: Magnetic structures and optical properties of rare-earth orthoferrites $RFeO_3$ (R = Ho, Er, Tm and Lu). Solid State Commun. **288**, (2019). https://doi.org/10.1016/j.ssc.2018.11.004
10. Behera, A.K., Mohanty, N.K., Satpathy, S. K., Behera, B., Nayak, P.: Investigation of complex impedance and modulus properties of Nd doped $0.5BiFeO_3$–$0.5PbTiO_3$ multiferroic composites. Cent. Eur. J. Phys. **12**, 851-861 (2014)
11. Murtaza, T., Khan, M.S., Ali, J., Hussain, T., Asokan, K.: Structural, electrical and magnetic properties of multiferroic NdFeO3–SrTiO3 composites. J. Mater. Sci. Mater. Electron. **29**, (2018). https://doi.org/10.1007/s10854-017-8123-8
12. Singh, A., Gupta, A., Chatterjee, R.: Enhanced magnetoelectric coefficient (α) in the modified $BiFeO_3$ -PbTi $O_3$ system with large la substitution. Appl. Phys. Lett. **93**, 022902 (2008)
13. Zhang, Q., Saito, F.: Effect of $Fe_2O_3$ crystallite size on its mechanochemical reaction with $La_2O_3$ to form $LaFeO_3$. J. Mater. Sci. **36**, 2287–2290 (2001)
14. Thakur, P., Gupta, K., Thakur, P., Kumar, A.S., Sudarsanam, V.: Sharma, P and Lal, Improvement in the structural, dielectric and magnetic properties of CFO-doped KNNS-BKT ceramics. J. Mater. Sci. Mater. Electron. **34**, 311 (2023)
15. Thakur, P., et al.: Tailoring of structural, morphological, electrical and magnetic properties of $LaMn_{1-x}Fe_xO_3$ ceramics. RSC Adv. **14**, 23592 (2024)
16. Razmara, Z.: Lanthanum (III) complex as ferromagnetic supraprecursor for preparation of $La_2O_3$ nanoparticles by thermal decomposition method. Res. Chem. Intermed. **45**, 2887–2901 (2019)
17. P Pookmaneea, P., Rujijanagulb, G., Anantab, S., Heimannc, R.B., Phanichphant, S. Effect of sintering temperature on microstructure of hydrothermally prepared bismuth sodium titanate ceramics. J. Eur. Ceram. Soc. **24**, 517–520 (2004)
18. Zubair, M.A., Mozahid, F.A., Takeda, H., Hossain, A.K.M.A.: Effect of processing temperature on structural, optical and frequency dependent electrical responses of solid-state sintered bismuth sodium titanate. Mater. Sci. Eng. B **274**, 115474 (2021)
19. Pradhan, L.K., Pandey, R., Kumar, S., Supriya, S., Kar, M.: Octahedral distortion due to oxygen vacancy reduction in $La^{3+}$ modified BNT-BTO solid solutions near morphotropic phase boundary. J. Phys. D Appl. Phys. **51**, 375301 (2018)
20. Stinton, G.W., Evans, J.S.O.: Parametric Rietveld refinement. J. Appl. Cryst. **40**, 87–95 (2007)
21. Saafan, S.A., Assar, S.T.: Dielectric behavior of nano-structured and bulk Li Ni Zn ferrite samples. J. Magn. Magn. Mater. **324**, 2989–3001 (2012)
22. Mondal, R.A., Murty, B.S., Murthy, V.R.K.: Maxwell-Wagner polarization in grain boundary segregated NiCuZn ferrite. Curr. Appl. Phys. **14**(12), 1727–1733 (2014)

23. Siddique, M.N., Ahmed, A., Tripathi, P.: Dielectric relaxation and hopping conduction mechanism in $Ni_{1-x}Sr_xO$ nanostructures. Mater. Chem. Phys. **239**, 121959 (2020)
24. Paswan, S.K., Pradhan, L.K., Kumar, P., Kumari, S., Kar, M., Kumar, L.: Electrical transport properties of nanocrystalline and bulk nickel ferrite using complex impedance spectroscopy: a comparative study. Phys. Scr. **97**, 095812 (2022)
25. Gao, F. et al.: Energy-storage properties of $0.89Bi_{0.5}Na_{0.5}TiO_3$–$0.06BaTiO_3$–$0.05K_{0.5}Na_{0.5}NbO_3$ lead-free anti-ferroelectric ceramics. J. Am. Ceram. Soc. **94**, 12, 4382–4386 (2011)
26. Liu, X. et al.: Ultrahigh energy density and improved discharged efficiency in bismuth sodium titanate based relaxor ferroelectrics with A-site vacancy. J. Materiomics **4**, 202–207 (2018)
27. Adhlakha, N., Yadav, K.L.: Study of structural, dielectric and magnetic behaviour of $Ni_{0.75}Zn_{0.25}Fe_2O_4$–$Ba(Ti_{0.85}Zr_{0.15})O_3$ composites. Smart Mater. Struct. **21**, 115021 (2012)
28. Zheng, M., Hou, Y., Ge, H., Zhu, M., Yan, H.: Effect of sintering temperature on internal-bias field and electric properties of 0.2PZN–0.8PZT ceramics. Phys. Status Solidi A **210**, 2 (2013)
29. Bochenek, D., Chrobak, A., Ziółkowski, G.: Electric and magnetic properties of the multiferroic composites made based on $Pb(Fe_{1/2}Nb_{1/2})_{1-x}Mn_xO_3$ and the nickel-zinc ferrite. Materials **16**, 3785 (2023)
30. Rearick, T. M., Catchen, G. L., Adams, J.M.: Combined magnetic-dipole and electric-quadrupole hyperfine interactions in rare-earth ortho-ferrite ceramics. Phys. Rev. B **48**, 224, 59(1993). https://doi.org/10.1103/physrevb.48.224
31. Kumar, R., Singh, R.K., Zope, M.K., Kar, M.: Tuning of magnetic property by lattice strain in lead substituted cobalt ferrite. Mater. Sci. Eng. B **220**, 73–81 (2017)
32. Pandey, R., Pradhan, L.K., Kar, M.: Structural, magnetic, and electrical properties of (1–x) $Bi_{0.85}La_{0.15}FeO_3$-(x) $CoFe_2O_4$ multiferroic composites. J. Phys. Chem. Solids **115**, 42–48 (2018). https://doi.org/10.1016/j.jpcs.2017.12.001
33. Kumar, S., Guha, S., Supriya, S., Pradhan, L.K., Kar, M.: Correlation between crystal structure parameters with magnetic and dielectric parameters of Cu-doped barium hexaferrite. J. Magn. Magn. Mater. **499**, 166213 (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.

Check for updates

*Concurrency and Computation: Practice and Experience*

WILEY

REVIEW ARTICLE

# Exploring Blockchain-Based Patient-Centric Healthcare: A Comprehensive Review

Deepak Kumar Mishra[1,2] | Pawan Singh Mehra[1]

[1]Department of CSE, Delhi Technological University, Delhi, India | [2]Department of Computer Science, College of Vocational Studies, University of Delhi, Delhi, India

**Correspondence:** Pawan Singh Mehra (pawansinghmehra@gmail.com)

**Received:** 16 June 2024 | **Revised:** 12 April 2025 | **Accepted:** 26 April 2025

**Funding:** The authors received no specific funding for this work.

**Keywords:** Blockchain | consent management | healthcare | interoperability | patient-centricity | privacy | review | security | smart contracts

## ABSTRACT

Blockchain, an immutable, decentralized, and distributed ledger technology, has the potential to revolutionize the sharing and access of confidential data, particularly in the healthcare sector, where patient-centric systems benefit greatly from its robust security features and ability to ensure data integrity. Recognizing the potential, this paper reviews current research on blockchain-based healthcare architectures that focus on patient-centric solutions, exploring the present state of these architectures and highlighting the growing interest in blockchain applications for enhancing healthcare. The review identifies key trends, challenges, and opportunities within this domain, analyzing the evolution of research over recent years, the types of studies conducted, their geographical distribution, and publication channels. It delves into the prevalent use cases, primary challenges addressed, and contributions of various studies, covering common blockchain platforms, types of blockchains, and the implementation of smart contracts. Guided by four critical questions aimed at improving security and privacy, streamlining patient consent management, fostering interoperability, and ensuring compliance with healthcare data regulations, this research underscores the importance of patient data security through advanced encryption techniques, access controls, and decentralized storage solutions. It also highlights the benefits of patient-controlled data sharing, empowering individuals with greater control over their health information. Despite progress, the review identifies areas requiring further exploration, such as advanced security threats, nuanced and dynamic consent models, comprehensive interoperability solutions, and alignment with evolving healthcare regulations. Addressing these gaps necessitates research into scalable privacy-preserving blockchain architectures that prioritize dynamic consent management and interoperability standards, thus making regulatory compliance possible in an ever-changing healthcare environment.

## 1 | Introduction

Healthcare today has evolved into a patient-centric approach to facilitate well-being and empower people to take control of their health. Patients should be at the center of their healthcare journey, and their medical data should be maintained safely and discreetly, while being readily accessible when needed. The traditional healthcare systems, in contrast, consist of centralized institutions like hospitals that control patients' private medical information [1]. In the world of traditional healthcare, patients' information is recorded and maintained as Electronic Health Records (EHRs) [2] in centralized databases holding crucial health information, ranging from medical history to current prescriptions or test results [3]. While there were significant improvements in maintaining and accessing health data, when moving from paper-based to electronic health records, the essential framework

© 2025 John Wiley & Sons Ltd.

remained the same: Central authorities still had control over patient data [4–6]. This centralized approach raises significant concerns, notably in relation to security, privacy, and confidentiality of health data [7].

Patient information is vulnerable to breaches and illegal access when it is contained within centralized systems. A growing number of cyberattacks and data breaches expose patients to the dangers of medical identity theft, insurance fraud, and even interference with treatment procedures [8]. Moreover, patients frequently seek treatment from a variety of professionals as healthcare becomes more dynamic and multifaceted [9]. As a result, patient records are fragmented and dispersed across several systems, making it difficult for patients to access a whole medical history to guide treatment decisions [10]. The manual and time-consuming processes of data sharing, with patients signing consent forms for each transfer, only add to the complexity and inefficiency of the system [9]. A wide range of health-related data is being collected from many sources with the introduction of wearable technology, including smartwatches, smartphones, and wireless sensors [11]. The potential of these data streams is to transform the healthcare industry as it enables better-informed decision-making, personal care, and improved patient outcomes [12]. To achieve this aim, however, calls for a dependable and effective data sharing architecture. It is what sets apart the sensitive nature of healthcare information from other forms of data. Medical reports can go as far as genetic information or past mental health, or medical disorders, and are very private in nature. If such an individual's details were accessed without their permission, they could suffer social stigma, emotional disturbance, or be prone to breaches of their privacy [13]. These weaknesses found in traditional healthcare systems underscore the need for a paradigm shift that focuses on patient focus, safety, and efficient data exchange.

Blockchain technology is a decentralized ledger system, and its potential resides in its capacity to provide transparent, immutable, and tamper-proof records [14]. It presents a way in which patients are in charge of their health data, giving them access to share it with trustworthy parties as they see suitable. Beyond managing patient data, blockchain technology in healthcare has significant benefits [15]. Given the complexity and gravity of the problems the healthcare sector is facing, the need for blockchain-based patient-centric designs is only becoming more critical. Table 1 summarizes the key differences between blockchain-based and traditional healthcare architectures.

To develop patient-centric architectures, this review paper explores the applications, advantages, and difficulties of implementing blockchain technology in healthcare. It focuses on how blockchain can give patients ownership of their data, eliminate single points of failure, improve cybersecurity, and promote trust among all parties involved in healthcare. With the knowledge and resources necessary to take control of their healthcare journey, patients may at last move from the margins to the centre. At the brink of a new era in healthcare, one that holds the potential for a more connected, secure, and patient-focused future, blockchain emerges as the facilitator.

**TABLE 1** | Comparison of blockchain-based and traditional architectures.

| Comparison criteria | Blockchain-based architectures | Centralized architectures |
|---|---|---|
| User control | Distributed and decentralized control | Centralized control |
| Trust model | Distributed among network participants | Placed in a central authority |
| Possible data operations | Read and write operations with data | Read, write and modification operations possible with data |
| Privacy | Traceability and user authentication ensure privacy | Users cannot track their data |
| Security | Data encryption and Consensus algorithms ensure security | Prone to malicious attacks |
| Transparency | Immutable and shared ledger ensure transparency | Limited transparency due to authorized access |

## 1.1 | Motivation

The rapid digitization of healthcare data has exposed vulnerabilities in data security and patient privacy. Traditional systems struggle to ensure data integrity, secure sharing, and patient consent management [16, 17]. Furthermore, achieving interoperability amidst diverse healthcare providers remains a challenge [18]. Blockchain technology offers a promising solution by providing decentralized, tamper-resistant storage, patient-controlled access, and data traceability. This research seeks to address these critical gaps by systematically analysing the landscape of patient-centric blockchain-based healthcare architectures. By identifying trends, challenges, and emerging solutions, this study aims to contribute to the development of robust and compliant healthcare systems that prioritize data security, patient consent, and interoperability.

## 1.2 | Objectives and Contributions

To enhance the security and privacy of healthcare data in blockchain-based architectures, this research paper analyses patient-centric blockchain-based healthcare systems. The following are the unique objectives and contributions of this paper:

- Provides a systematic literature review to identify, classify, and analyze the patient-centric blockchain-based healthcare architectures to discover the main research gaps, trends, and challenges, and proposes areas for future research with potential solutions.
- Compiles a selection of 41 representative papers from a collection of 217 papers to ensure a comprehensive analysis and identifies the trend over the publication years, publication types, their geographical distribution, publication channels, source of publication, main challenges addressed and their contributions, catered use cases, most prevalent



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

blockchain platforms, types of blockchains and smart contract programming languages used in the implementation of existing architectures.

- Critically discusses the four research questions centered on enhancing security and privacy, streamlining patient consent management, addressing interoperability, and aligning architectures with healthcare data regulations while proposing future research directions for patient-centric blockchain-based healthcare architectures.

The contributions of this paper can have a significant impact on shaping the landscape of patient-centric blockchain-based healthcare architectures. By offering a comprehensive literature review, this study acts as a guide for researchers and practitioners, facilitating informed decision-making. The proposed future research directions pave the way for advanced security measures, dynamic consent mechanisms, and legal frameworks, promising a ripple effect of innovation in the design of secure, privacy-respecting, and regulation-compliant healthcare systems.

The rest of this paper is structured as follows: In Section 2, a comprehensive introduction to blockchain technology is presented. Section 3 presents an overview of the existing literature in this field. The research methodology is elaborated upon in Section 4, followed by the presentation of the outcomes of the systematic literature review in Section 5. Section 6 conducts an in-depth analysis of the findings, addressing the research questions and suggesting potential directions for future research. Finally, Section 7 concludes the paper.

## 2 | Overview of Blockchain

The underlying ideas behind blockchain technology are briefly explained in this section. It discusses the overview of blockchain, explains the various kinds of blockchains, and investigates the idea of smart contracts in relation to blockchain networks.

### 2.1 | Background

Blockchain technology represents a decentralized and distributed database system that maintains a secure and unalterable record of transactions. Satoshi Nakamoto first introduced it in 2008 as a peer-to-peer electronic cash system [19]. Blockchain consists of a series of blocks, each of which ensures the chronological order of transactions by including a timestamp and a link to preceding blocks [20]. Figure 1 illustrates a series of interconnected blocks in a blockchain. Each block consists of a header and a body. The body contains a Merkle tree with hashes of all the data nodes. The header includes a timestamp, nonce, hash pointer to the previous block, Merkle root, transaction counter, and version information, forming the foundation of the blockchain's structure. Due to its capacity to enable direct interactions between interacting entities, eliminating the need for an intermediary, this technology has attracted considerable attention [12]. The core attributes of blockchain—decentralization, immutability, and cryptographic consensus—offer diverse potential applications across sectors such as finance, education, voting, and healthcare.

Blockchain guarantees secure transaction storage and tracking of assets because it is a shared and immutable ledger that is distributed and decentralized [21]. With this system, transactions are confirmed and approved by authorized nodes, removing the need for trust in third-party intermediaries [22]. Blockchain is a compelling option for healthcare data management because its tamper-proof nature protects data integrity [21, 23]. The architecture of blockchain relies on a distributed ledger and a peer-to-peer network [24]. All transactions are recorded in the distributed ledger, and each node keeps an identical local copy of this shared ledger, preventing malicious nodes from altering the data [24]. Before being uploaded to the network, transactions go through consensus algorithms for verification and authentication, assuring the reliability of the data. A transaction becomes immutable once it is recorded, making any tampering readily evident [21].

To establish trust within the network, blockchain leverages consensus algorithms like Proof of Work (PoW) and Proof of Stake (PoS) [25]. PoW necessitates validators to solve complex problems to generate new blocks, earning the right to validate transactions. On the other hand, PoS grants validation rights based on the amount and duration of coins held by nodes. These robust algorithms ensure a high level of trust and security, making blockchain an ideal solution for applications that require trust among interacting entities [25]. Amidst this transformative

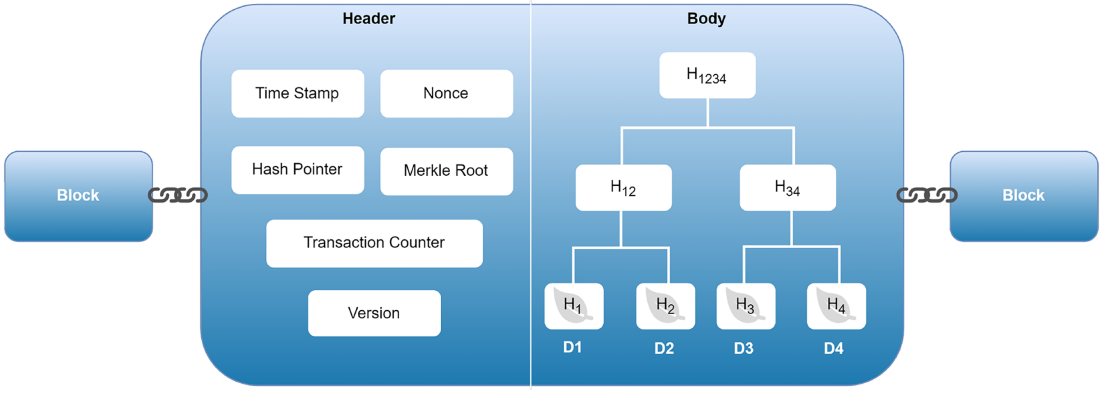



**FIGURE 1** | Structure of the blockchain.



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

landscape, dynamic consent emerges as an innovative data management approach, continuously empowering individuals with control over their data for research and healthcare purposes [26, 27]. Here, blockchain shines as a beacon of hope, providing a resilient and privacy-preserving solution [28]. Blockchain's inherent features, such as enhanced accountability and data security, fortify the dynamic consent model. Patients gain the power to dynamically update their consent preferences, safeguarding the integrity and confidentiality of their personal data [28, 29].

Embracing blockchain empowers the healthcare industry to elevate data management practices, enrich patient care, and ensure data security and privacy. As research and innovations continue to unfold, blockchain stands poised as a transformative force, shaping the future of patient-centric healthcare systems.

## 2.2 | Blockchain Types

Various methodologies exist for establishing a blockchain network, each tailored to specific use cases and their unique requirements. For instance, when considering cryptocurrencies such as Bitcoin, a public network is utilized, wherein any node is granted the privilege to participate in the verification of transactions. In stark contrast, a financial institution like a bank necessitates a private network to uphold strict access controls, ensuring that only authorized users possess the right to access confidential information [30]. This fundamental dichotomy has led to the classification of blockchain networks into two main categories: Public and private. In public blockchains, access is granted to all participants, thereby allowing anyone within the network to validate transactions. Consequently, power is dispersed equitably among nodes, and no single entity wields ultimate control over the network's operations. Conversely, private blockchains impose restrictions on access by exclusively authorizing a selected number of nodes, as designated by the network administrator. These authorized nodes are entrusted with the capability to participate in the blockchain network [31]. Moreover, these classifications can be further delineated based on the nature of who can generate new blocks in the blockchain network, culminating in the stratification of blockchain into four distinct types [28, 32]. Table 2 summarizes the comprehensive differentiation of blockchain networks.

- Private permissionless: This category encompasses blockchain networks that grant entry only to a limited number of authorized nodes. Within this network paradigm, all participating nodes are empowered to engage in read and write operations, thereby collaboratively contributing to the generation of new blocks. Each node is deemed of equal significance. An illustrious example of a private permissionless blockchain is Holochain [33].
- Private permissioned: In this classification, only a select number of authorized nodes are afforded entry into the blockchain network. While all participants are granted the ability to engage in read operations, the privilege of conducting write operations and initiating the generation of new blocks is reserved for a specified subset of nodes. Noteworthy instances of private permissioned networks include Hyperledger Fabric [34] and Hyperledger Besu [35].
- Public permissionless: This particular type of blockchain network extends an open invitation to any node seeking to join and participate in read or write operations. However, building trust among participants necessitates the implementation of costly consensus protocols. Preeminent examples of public permissionless blockchains comprise renowned platforms such as Ethereum, Bitcoin, and Witnet [36].
- Public permissioned: Analogous to public permissionless blockchains, public permissioned blockchain networks embrace an unrestricted influx of nodes and allow them to engage in read operations and access the network's contents. Nevertheless, the crucial distinction lies in the authorization of only select participants or nodes to execute write operations. In view of the less intensive consensus protocols employed, establishing trust amongst the participants is rendered more cost-effective. Eminent illustrations of public permissioned blockchains encompass ESIO [37] and Ripple [38].

In conclusion, the variety of blockchain network establishment methodologies caters to diverse use cases and their respective requisites. By classifying blockchain networks into distinct categories, based on public or private characteristics, and further differentiating them into permissioned and permissionless structures, the blockchain landscape offers a comprehensive framework to cater to an extensive array of applications across various industries.

## 2.3 | Smart Contracts

In the realm of blockchain-based patient-centric healthcare architectures, smart contracts have emerged as a powerful and transformative tool. Unlike traditional contracts, smart contracts offer a multitude of benefits, including increased automation, enhanced transparency, heightened security, cost reduction, and the elimination of intermediaries. Smart contracts are self-executing agreements with predefined terms and conditions

**TABLE 2** | Comprehensive differentiation of blockchain networks.

| Aspect | Private permissionless | Private permissioned | Public permissionless | Public permissioned |
|---|---|---|---|---|
| Participants | Limited authorized | Limited authorized | Anyone | Anyone |
| Read operations | Anyone | Anyone | Anyone | Anyone |
| Write operations | Anyone | Limited authorized | Anyone | Limited authorized |
| Platforms | Holochain [33] | Hyperledger Fabric [34], Hyperledger Besu [35] | Ethereum, Witnet [36] | ESIO [37], Ripple [38] |



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

directly embedded in code [39]. When certain conditions are met, they execute automatic operations and are deployed and stored on the blockchain. Contractual obligations are fulfilled with the help of this automation, eliminating the need for intermediaries or centralized control [40].

Smart contracts have several benefits when integrated into healthcare systems. To streamline processes and get rid of human intervention, smart contracts automate predefined actions. As a result, the pace of healthcare activities increases while administrative burden drops, which is ultimately good for patients as well as healthcare practitioners [12]. Smart contracts' immutability implies that the terms and conditions in them are open and unchangeable. This creates trust among all concerned parties, thereby encouraging patient centricity where transparency is everything. Patient data, along with sensitive healthcare information, is handled securely thanks to cryptographic algorithms employed in the design process of smart contracts. Unauthorized access plus data breaches are mitigated by this robust security mechanism [41]. When intermediaries are eliminated from the equation, transactional costs of health care operations can be significantly reduced. In this respect, smart contracts enable direct peer-to-peer interactions and could therefore save both patients and providers much money.

To make blockchain-based healthcare architectures implementing and deploying smart contracts require some competence in technology. Generally, programming languages such as Solidity are used to write smart contracts on platforms like Ethereum. The development process has to be "right" so that the contract can be executed well and remain stable. As soon as a contract is deployed on the blockchain, it becomes immutable. To avoid any mistakes or flaws that may have unforeseen consequences, code for a given contract must be meticulously vetted and tested. Seamless integration is necessary when introducing smart contracts into existing healthcare systems. For ease of communication between various components of the healthcare infrastructure and a smart contract, as well as ease of its interaction with them, interoperability hindrances should be overcome [42]. However, despite their numerous advantages, issues concerning legal status still exist among proponents of this concept. These remain to be non-binding, universally enforceable ones [39]. Therefore, to operate optimally within the existing regulations, it is essential to integrate a legal framework component that will facilitate the adoption of smart contracts in the healthcare sector.

A blockchain-based, patient-centric healthcare architecture utilizes smart contracts that are capable of revolutionizing healthcare delivery. As technology progresses, efforts should be directed towards further research and development to improve smart contract functionalities. The creation of a strong environment that maximizes the utilization of smart contracts in healthcare necessitates the participation of technical professionals, lawyers, and other stakeholders in the health sector. To sum up, it can be said that innovative progress in blockchain-based patient-centric healthcare architectures is led by smart contracts. While technical challenges and legal considerations exist, the future outlook for smart contracts in healthcare is promising, paving the way for a more efficient, secure, and patient-centric healthcare landscape.

## 3 | Related Work

This section summarizes the relevant existing literature by highlighting the key findings, contributions, and limitations of each study in a concise and organized manner, along with a detailed analysis.

### 3.1 | Existing Literature Summary

In a study by researchers [43], a decentralized framework called MedRec was introduced, utilizing the Ethereum blockchain to effectively manage and exchange Electronic Medical Records (EMRs). This approach ensures secure storage, authentication, and accountability for medical data sharing. However, it relies on private databases for data storage, which introduces vulnerabilities to potential cyberattacks and permits internal access to sensitive information.

In separate studies [44] and [45], authors aimed to maintain privacy, data integrity, and patient anonymity by devising a blockchain-based system for medical data access and sharing, facilitated through an immutable distributed ledger. Nonetheless, it's worth noting that scalability concerns were identified within their proposed system. Lodha et al. [22] designed a healthcare system for secure medical data sharing among various stakeholders on a blockchain storage system without any trusted third party. To prevent repeated extensive check-ups, the proposed system does, however, require the patient to upload their entire medical history. The requirement of necessary hardware and a stable internet connection also serves as a roadblock.

The authors in Reference [46] developed a framework for the earlier diagnosis of diabetes utilizing machine learning algorithms and secure blockchain-based patient EHR management to protect privacy. The data is collected from wearable devices, stored on the IPFS (Interplanetary File System), sent to the EHRs manager, and given as input to ML classification models. After processing, the predictive results are generated and stored on the blockchain after the patient's consent. In Reference [47], a system was designed to share and manage personal health data using blockchain and cloud storage, eliminating the dependence on a centralized third-party for sensitive data sharing. However, the proposed system is yet to be integrated with the existing healthcare systems. In Reference [48], A decentralized blockchain-based system for managing patient-centered electronic health records with IPFS storage and a control mechanism for limited access was proposed by Reen et al. The system faces limitations like a maximum replication factor of 2, privacy infringement, and the on-chain existence of a single entity of a hospital.

Dubovitskaya et al. [9] developed a proposed framework called ACTION-EHR that uses a permissioned blockchain to give radiation oncology patients secure access to their medical record sharing. This approach allows medical practitioners and researchers to access patient data more quickly and efficiently while reducing costs. But there is a risk of having a single point of failure of the system, and in case of emergency situations, there is a requirement for a "break glass" mechanism, which serves as a limitation of this paper. In Reference [49], the authors proposed



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

HapiChain—a blockchain-based patient-centric framework for teleconsultation service in complementary to telemonitoring, for effective and efficient communication. The system needs to be tested against security attacks to ensure security. In Reference [50], the authors proposed an efficient blockchain-based patient-centric framework making clinical data available to patients and enabling clear, traceable, secure, and efficient data exchange but the proposed architecture has a costly maintenance and deployment and does not guarantee that it will be applicable to every health care data.

Mani et al. [51] proposed a patient-centric approach to store health records on an IPFS-based distributed ledger system using Hyperledger Fabric technology and enabled decentralized storing and sharing of health information while guaranteeing scalability, privacy, security, and confidentiality issues. A large number of resources are required for the implementation of multi-blockchain systems and an extended framework integrated with Non-Fungible Tokens. Chelladurai et al. [52] proposed a blockchain-based patient-centric approach for EHR management when health records are fragmented among healthcare entities.

Jabarulla et al. [53] devised a decentralized approach for managing patient-centric images using blockchain technology. Their system aimed to enhance data security and privacy while eliminating the need for a centralized authority. However, this system presented a potential drawback in terms of private key management due to the manual encryption and decryption of medical images. According to Fatokun et al. [54], an EHR system that is based on the Ethereum blockchain and smart contracts was introduced in this paper. The authors' approach aimed at preventing any reliance on third-party systems as a way of ensuring patient-centricity, interoperability, and data security. Furthermore, they noticed that the said system had some issues connected with the scalability of blockchain and communication overheads related to it, such as bandwidth requirements for block mining and overall network communication. An innovative framework called PCH was introduced by the authors in Reference [55], which is patient-centric healthcare. Within this architecture, blockchain, cloud computing, and IoT are combined to enhance semantic interoperability among healthcare systems. It should be realized, though, that implementation and maintenance costs would be enormous. This approach also does not entirely overcome legal complexities associated with sharing patients' data across different healthcare platforms.

SynCare is a newfangled blockchain and cryptographic remote patient monitoring platform introduced by Pighini et al. in their work [56]. In this way, the innovation helps to make possible the storage of patient-focused data, sharing and retrieval of it, while strictly protecting privacy for sensitive details. Nevertheless, one disadvantage is that the process encrypts and decrypts information when shifting between functions, thereby leading to a long loading time. Taylor et al. [57] proposed VigilRx—a drug dispensing system which is designed with a focus on patients and interoperability. This therefore gives patients more control over their data, as well as reducing bottlenecks of information through better transfer efficiency of data. It also seeks to promote compatibility, curb fraud, and improve the precision and availability of data at all times. It mainly relies on blockchain technology so that it can achieve scalability, but not security in network terms. Due to this heavy dependence on it, there are challenges such as blockchain congestion or susceptibility to security breaches.

George et al. [58] presented the MediTrans system for safe interoperable patient-centric data access management to address the concerns of patient-centricity, transfer of large-scale data sets, and lack of trust. The system is dependent on the willingness of stakeholders to join the network. Mudaliar et al. [59] presented a blockchain-based and machine learning-based electronic health record management system to track and predict the risk of arrhythmia and heart illnesses, respectively, but the technology carries a potential data theft risk.

In Reference [60], the authors proposed a patient-centric blockchain-based health record access framework using Hyperledger Fabric to enhance patient control, protect privacy, and data integrity. Patients can impose necessary conditions before granting access to their medical data. But they proposed a conceptual framework that has not been implemented in a real-world setting and integrated with the existing healthcare models. Rai in Reference [61] proposed a patient-controlled blockchain-based electronic health records management system to ensure patient ownership of their medical data, elimination of security risks, data integrity, and confidentiality. But the smart contracts of the system are based on time and notifications.

To conclude, the analysis of blockchain-based healthcare architectures demonstrates advancements and challenges in patient-centricity, data security, privacy, and interoperability. Many healthcare architectures empower patients, but with data security and scalability trade-offs. These studies highlight blockchain's potential in revolutionizing healthcare but underscore the need for robust solutions. Table 3 summarizes the key findings, contributions, and limitations of each study, emphasizing the ongoing research required to ensure security, scalability, and patient-centered care in the evolving landscape of blockchain-based healthcare architectures.

### 3.2 | Analysis of Existing Literature

The analysis of various blockchain-based healthcare architectures reveals a diverse landscape of approaches and trade-offs. Addressing data security and patient privacy are common challenges across the studies, with most proposing permissioned blockchains like Ethereum and Hyperledger Fabric. While several studies emphasize user control and traceability, some exhibit limitations in addressing these aspects comprehensively. Interoperability [62] and scalability [63] considerations vary, with a mix of on-chain and off-chain data storage solutions. Notably, consensus mechanisms predominantly include Proof of Work and Proof of Authority. A majority of studies lack detailed security analysis and performance evaluation. Thus, while these architectures demonstrate potential in enhancing healthcare systems, the existing literature suggests the need for more comprehensive solutions that balance security, privacy, user control over data, interoperability, and scalability while addressing specific healthcare domain requirements. Table 4 presents an analysis of the existing literature.



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**TABLE 3** | Summary of existing literature.

| Author | Year | Findings/contributions | Limitations |
|---|---|---|---|
| Azaria et al. [43] | 2016 | • MedRec: Decentralized EMR management using blockchain<br>• Ensures data sharing, accountability, authentication, and confidentiality for sensitive medical information.<br>• Working prototype presented with detailed analysis and discussion | • Scalability challenges persist.<br>• Lack of detailed security analysis.<br>• System relies on provider nodes adhering to external regulations like HIPAA for data copying. |
| Chen et al. [47] | 2018 | • Blockchain-based storage for safe sharing of medical data<br>• Designed block structure, transaction types, and main functions<br>• Managed personal medical data via blockchain and cloud storage | • Not yet integrated with existing healthcare systems.<br>• Regulatory and legal challenges not addressed |
| Reen et al. [48] | 2019 | • An Electronic Health Records (EHR) privacy and security solution on a permissioned Ethereum blockchain.<br>• Combined symmetric and asymmetric key cryptography for access control and secure data storage. | • The use of biometric authentication for identity verification may not be feasible for all patients.<br>• On-chain existence of a single entity of the hospital. |
| Dubovitskaya et al. [9] | 2020 | • ACTION-EHR, a secure patient-centric EHR data sharing framework on a permissioned blockchain, enabling patients to manage records.<br>• Ensures privacy and data security via encryption and digital signatures.<br>• Tested with deidentified patient data using Hyperledger Fabric. | • Single point of failure if the system deployment lacks redundancy (e.g., with a single orderer and single CA). |
| Kordestani et al. [49] | 2020 | • HapiChain, a blockchain-based architecture for patient-centered telemedicine that enhances security and reliability.<br>• The system integrates telemonitoring and teleconsultation services with three layers: Interface, DApp, and blockchain. | • Does not provide any detailed analysis of performance.<br>• Need to address the attacks and verify the security of HapiChain against those attacks. |
| Kanagi et al. [50] | 2020 | • A patient-centric blockchain framework for efficient clinical data sharing.<br>• Demonstrates rapid sharing of health data, addressed interoperability issues, and reduced response times in Taiwan's healthcare. | • Costly maintenance and deployment.<br>• Does not promise its generalizability to any health care data. |
| Mani et al. [51] | 2021 | • Patient-centric healthcare data management (PCHDM) framework based on blockchain, ensuring security and privacy of health records.<br>• It includes a secure Health record chain network architecture, a permissioned network in Hyperledger fabric, and performance testing. | • Requires significant computational resources and energy consumption, leading to scalability issues.<br>• May not provide absolute protection against all security and privacy threats. |
| Chelladurai et al. [52] | 2021 | • A patient-centric EHR storage and integrity management system based on blockchain to ensure secure transactions in e-Health systems.<br>• Emphasis on responsible handling of private health data, ensuring integrity and confidentiality. | • Encountered challenges such as regulatory compliance, interoperability, and scalability. |
| Jabarulla et al. [53] | 2021 | • An IPFS and Ethereum blockchain-based proof-of-concept architecture for a patient-centric image management (PCIM) system.<br>• Enables hospitals, patients, and image requestors to share medical data in a secure and effective manner. | • Scaling challenges with growing medical images and participants.<br>• Increased transactions and data volume on blockchain might impact system performance. |
| Fatokun et al. [54] | 2021 | • A patient-centric Electronic Healthcare Record (EHR) system constructed using Ethereum blockchain and smart contracts.<br>• Aims to fulfill security, privacy, interoperability, and performance needs. | • Scalability challenge due to increasing data size.<br>• Mining new blocks and broadcasting communications to all network nodes add extra burden. |
| Gohar et al. [55] | 2022 | • A Patient-Centric Healthcare Framework (PCH), that combines Blockchain, Cloud, and IoT technologies.<br>• Enhances healthcare systems interoperability, security, and efficiency. | • Cost of implementation and maintenance is high.<br>• Does not address the legal and regulatory challenges for sharing medical data. |
| Pighini et al. [56] | 2022 | • SynCare, a patient-centered ecosystem developed by LifeCharger.<br>• Aims to keep patient monitoring and health data recording secure.<br>• Utilizes public Ethereum blockchain for validating data sharing consents and implements two-level encryption to enhance data security. | • Data encryption and decryption occur each time the user accesses the calendar function.<br>• Transition delay between functions due to increasing data volume. |
| Taylor et al. [57] | 2022 | • VigilRx, a patient-centric prescription management system focusing on interoperability.<br>• Gives patients more control, avoids information blocking, and improves the effectiveness of data transport. | • Heavy blockchain dependence.<br>• Limits effectiveness due to scalability.<br>• Vulnerable to congestion & security breaches. |
| George et al. [58] | 2022 | • MediTrans, a secure and interoperable blockchain-based system for managing patient-centric data access.<br>• Patients can grant data access to healthcare providers for a set duration.<br>• Employs CP-ABE to ensure patient data security and privacy. | • The proposed system is in the prototype phase.<br>• Requires additional testing and development to implement on a broader scale. |
| Mudaliar et al. [59] | 2022 | • A blockchain, IoT, and AI-based patient-centric solution to enhance healthcare management and improve data privacy.<br>• Incorporates machine learning for arrhythmia and heart disease risk prediction. | • The accuracy of the ML algorithms used was compromised due to the high feature dependence in the dataset. |
| Abutaleb et al. [60] | 2023 | • A framework that empowers patients for health records access control and authorization.<br>• Involves experiments for the optimal choice of blockchain operating system.<br>• Performance analysis shows acceptable speed for the operating system and smart contracts. | • Limited cost analysis & legal challenges discussed.<br>• Implementation might require substantial healthcare infrastructure changes. |
| Rai [61] | 2023 | • A decentralized Patient-Controlled Blockchain Enabled Electronic Health Records (PcBEHR) to allow patients to have control over their medical data that is immutable and traceable, using IPFS for storage. | • The proposed solution is still in the theoretical stage.<br>• Requires further research and development before it can be implemented in real-world healthcare systems. |



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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 4** | Analysis of existing literature.

| Study | Existing challenges | | | | | | Use case | Type of blockchain | Technology | Data storage | Consensus algorithm | Security analysis | Performance evaluation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | a | b | c | d | e | f | | | | | | | |
| Azaria et al. [29] | ✓ | ✓ | ✓ | ✓ | ✓ | X | Electronic Medical Records | Permissioned | Ethereum | Off-chain | Proof of Work | X | X |
| Chen et al. [47] | ✓ | ✓ | ✓ | ✓ | X | X | Healthcare Data Sharing | Permissioned | — | — | — | X | X |
| Dubovitskaya et al. [4] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Radiation Oncology Data | Permissioned | Hyperledger Fabric | Mixed | — | ✓ | X |
| Kordestani et al. [49] | ✓ | ✓ | ✓ | ✓ | X | ✓ | Telemedicine and Telehealth | — | Ethereum | Off-Chain (IPFS) | Proof of Work | X | X |
| Kanagi et al. [50] | ✓ | ✓ | ✓ | ✓ | ✓ | X | Clinical Data Sharing | Permissioned | Ethereum | Mixed | Proof of Work | ✓ | X |
| Chelladurai et al. [52] | ✓ | ✓ | ✓ | X | ✓ | X | Electronic Health Record Management | — | — | On-Chain | Proof of Work | ✓ | ✓ |
| Jabarulla et al. [53] | ✓ | ✓ | ✓ | ✓ | X | X | Medical Image Management | Permissioned | Ethereum | On-Chain | Proof of Work | ✓ | ✓ |
| Mani et al. [51] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Healthcare Data Sharing | Permissioned | Hyperledger Fabric | On-chain | Byzantine Fault Tolerance | ✓ | ✓ |
| Fatokun et al. [54] | ✓ | ✓ | ✓ | X | ✓ | X | Electronic Health Record Management | Consortium | Hyperledger Fabric | On-Chain | — | ✓ | ✓ |
| Gohar et al. [55] | ✓ | ✓ | ✓ | ✓ | ✓ | X | Healthcare Data Sharing | Permissioned | Hyperledger Fabric | Mixed | — | ✓ | X |
| Pighini et al. [56] | ✓ | ✓ | ✓ | X | X | X | Remote Patient Monitoring | Permissionless | Ethereum | Off-Chain (Cloud) | Proof of Work | ✓ | X |
| Taylor et al. [57] | X | ✓ | ✓ | ✓ | ✓ | ✓ | Prescription Management System | Permissionless | Ethereum | On-chain | Proof of Work | ✓ | ✓ |
| George et al. [58] | ✓ | ✓ | ✓ | ✓ | ✓ | X | Data Access Management | Permissioned | Ethereum | Off-Chain (Cloud) | Proof of Authority | X | ✓ |
| Abutaleb et al. [60] | ✓ | ✓ | ✓ | X | ✓ | X | Healthcare Data Sharing | Permissioned | Hyperledger Fabric | — | — | ✓ | ✓ |
| Rai [61] | ✓ | ✓ | ✓ | ✓ | ✓ | X | Electronic Health Records | Permissionless | Ethereum | On-Chain | Proof of Work | X | ✓ |

*Note:* a: Data Security, b: Patient Privacy, c: User Control over Data, d: Traceability, e: Interoperability, and f: Scalability.



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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 | Research Methodology

The chosen research methodology for this study involves a systematic literature review. The aim is to perform a detailed study and understand the contemporary research on patient-centric systems utilizing blockchain technology by presenting a quantitative portrayal of the existing research evidence in this domain [64].

### 4.1 | Article Selection

The first step in the systematic literature review is the selection of all the appropriate research papers based on the topic of research. Hence, the systematic literature review on "Blockchain-based Patient-centric Healthcare Architectures" embarked on a rigorous article selection process to identify relevant and high-quality research papers that can further be used to develop research questions and objectives. This section gives a systematic explanation of the article selection procedure, covering the search strategy, inclusion and exclusion criteria, screening, and outcomes. The entire process of selection of relevant articles is depicted in Figure 2.

#### 4.1.1 | Search Strategy

To conduct a systematic literature search, a well-defined search protocol was created. To ensure the retrieval of all publications pertaining to the convergence of blockchain technology and patient-centric healthcare, the search phrase was carefully developed. Pilot searches were performed to identify suitable keywords and terms. Ultimately, the search query was designed to capture papers with abstracts mentioning "patient centric", "patient-centric", "patient-controlled", "blockchain", "smart contracts", "distributed ledger technologies (DLT)" or "DLT". The

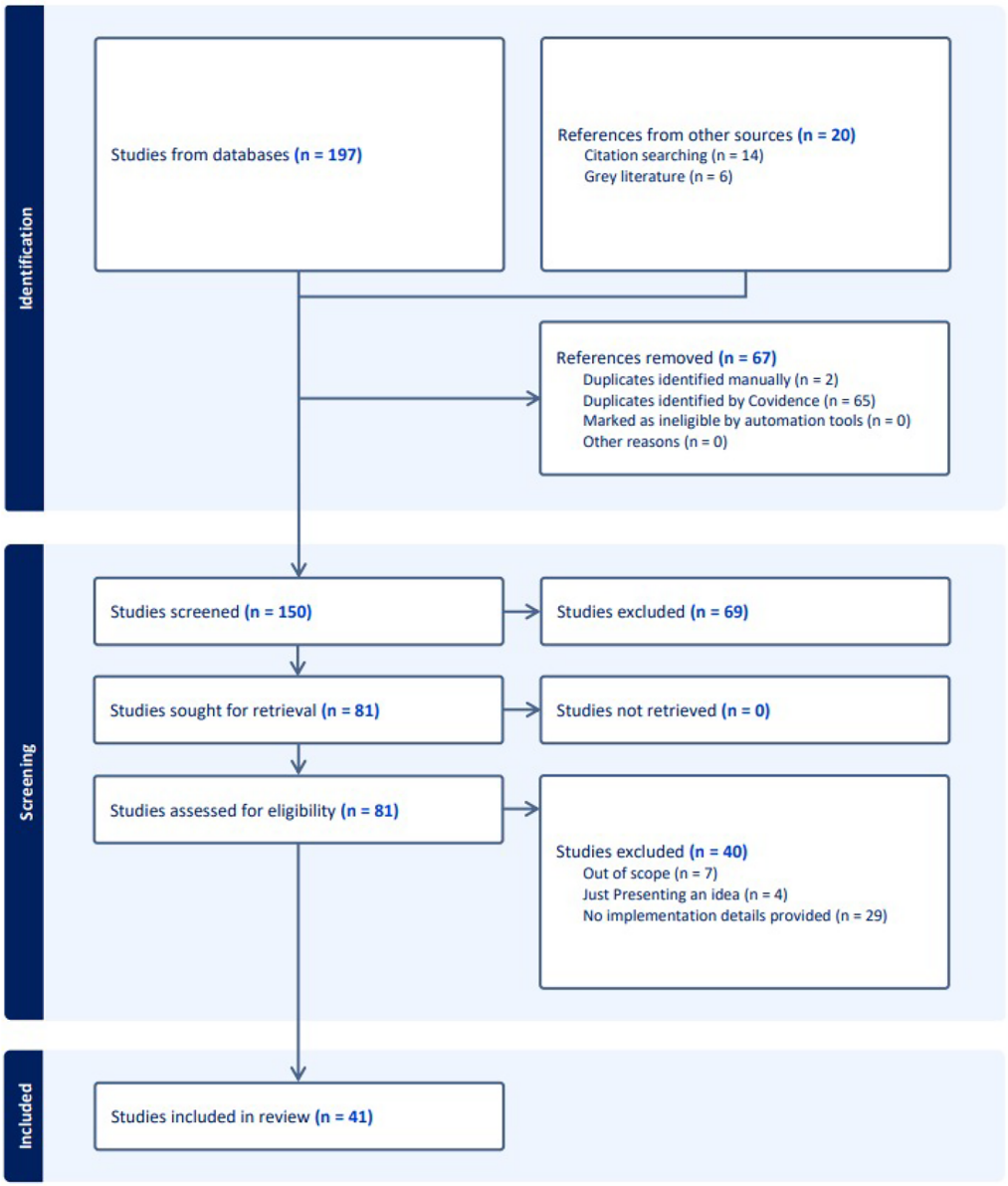



**FIGURE 2** | PRISMA search methodology.



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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** | Search query to find relevant papers.

| SEARCH QUERY: |
|---|
| [[Abstract: "Patient centric"] OR [Abstract: "Patient-centric"] OR |
| [Abstract: "Patient-controlled"] OR [Abstract: "Patient controlled"]] AND |
| [[Abstract: Blockchain] OR [Abstract: "Smart contracts"] OR |
| [Abstract: "Distributed ledger technologies"] OR [Abstract: dlt]] |
| AND [E-Publication Date: (01/01/2018 TO 31/12/2023)] |

**TABLE 6** | Inclusion criteria.

| **Inclusion criteria** | |
|---|---|
| 1. | Original peer-reviewed publication |
| 2. | Publication on the topic of blockchain-based patient-centric systems |
| 3. | Publication must describe a blockchain-based conceptual model, framework, architecture, prototype, proof-of-concept, implementations, or pilot relevant to patient-centric applications |
| 4. | Publications including sufficient explanation of the research findings |

**TABLE 7** | Exclusion criteria.

| **Exclusion criteria** | |
|---|---|
| 1. | Publications not in English |
| 2. | Review articles |

search period was limited from January 2018 to December 2023 to focus on recent and relevant publications. Table 5 illustrates the search query.

Clear inclusion and exclusion criteria were established to ensure the selection of appropriate papers. To be included, articles needed to be original, peer-reviewed publications discussing blockchain-based patient-centric systems. The publications had to describe conceptual models, frameworks, architectures, prototypes, proof-of-concepts, implementations, or pilots relevant to patient-centric applications. Moreover, the included articles should provide a sufficient explanation of their research findings. Exclusion criteria encompassed non-relevant publications, non-English articles, and review articles. Tables 6 and 7 depict the inclusion and exclusion criteria, respectively.

#### 4.1.2 | Identification of Relevant Studies

The search yielded a total of 197 studies from various scientific databases. The primary papers were searched using ScienceDirect, Elsevier, Plos One, Google Scholar, MDPI, Springer, SpringerLink, JMIR (Journal of Medical Internet Research) Publications, Scopus, and IEEE Xplore. In addition to the database searches, 20 potentially relevant papers through citation searching and references from other sources, as well as exploring grey literature, were identified, resulting in a combined pool of 217 potential studies. The generated articles were transferred to Covidence, an online application tool utilized for systematic review management. After removing 67 duplicate references, the remaining 150 studies proceeded to the screening phase.

#### 4.1.3 | Screening Process

During the screening phase, the articles were carefully assessed based on their titles, abstracts, and full-text availability. The screening phase involved assessing titles and abstracts to exclude studies that were clearly unrelated to blockchain-based patient-centric healthcare architectures. Studies that mentioned "blockchain" in contexts not aligned with technical aspects of computer science were also excluded. In ambiguous cases, where relevancy could not be determined solely based on titles and abstracts, the papers were passed to the next screening phase for further evaluation. Papers without full-text availability, papers not in English, duplicates, and papers presenting unrelated topics were also excluded.

At this stage, 69 studies were excluded as they did not align with the research topic or were out of scope. The remaining 81 studies were sought for retrieval, and all 81 were successfully retrieved for further evaluation. The 81 retrieved studies underwent a meticulous eligibility assessment based on the inclusion and exclusion criteria. As a result, 40 studies were disregarded for a variety of reasons, including being out of scope, offering abstract concepts without specifics about how they would be implemented, or being unrelated to the goals of the research.

After screening the final collection of 41 carefully selected research papers provides a structure for the in-depth study by providing significant contributions to the field and highlighting the potential of blockchain technology to transform patient-centric healthcare architectures.

### 4.2 | Data Extraction and Formulated Research Questions

The data extraction process for our systematic literature review on "Blockchain-Based Patient-Centric Healthcare Architectures" involved the gathering of key information from the selected research papers to address the research questions effectively. To achieve this, a meticulously drafted data extraction form was designed, capturing relevant data items to facilitate a comprehensive analysis of the papers. The form contained a variety of data items, as listed in Table 8. The formulated research questions on blockchain-based patient-centric healthcare systems are described below:

- RQ1: How can the security and privacy of healthcare data be enhanced in blockchain-based storage and sharing mechanisms to protect patient information?
  The existing healthcare data sharing mechanisms based on blockchain may not adequately address security and privacy concerns. Patient data requires utmost protection



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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 8** | Data extraction items.

| # | Data item | Description |
|---|---|---|
| 1 | Study identifier | Unique identifier for each research paper (e.g., #1, #2, etc.) |
| 2 | Title | Title of the article |
| 3 | Authors | Name of the authors |
| 4 | Country | Country of origin of the author(s) |
| 5 | Year | Publication year of the research paper |
| 6 | Publication channel | Name of publication place |
| 7 | Publication type | Type of Publication (Conference/Journal) |
| 8 | Publication source | Source of Publication (Academia/Industry) |
| 9 | Use Case | The specific use case addressed in the paper |
| 10 | Main Challenge Addressed | The primary challenge or problem discussed in the paper |
| 11 | Contribution | The contribution of the paper to the research domain |
| 12 | Limitations/Disadvantages | Limitations or disadvantages mentioned in the paper |
| 13 | Contribution Type | Type of contribution (e.g., architecture, framework, system) |
| 14 | Blockchain Type | Type of blockchain (e.g., permissioned, permissionless, consortium) |
| 15 | Platform/Framework | The blockchain platform or framework utilized in the study |
| 16 | Smart Contracts Language | The programming language used for smart contracts in the study |

from unauthorized access, data breaches, and tampering. Research must focus on improving encryption techniques and access control mechanisms to fortify the security and privacy of healthcare data within the blockchain ecosystem.

- RQ2: What methods can be developed to effectively identify and manage patient consent for data sharing within blockchain-based healthcare architectures?
  In blockchain-based healthcare architectures, advanced methods are urgently needed to accurately identify and manage patients' consent for data sharing. Existing consent mechanisms may not be sufficient to handle the complex and dynamic nature of patient data within a decentralized environment. Developing effective consent management techniques will strengthen patient control over their data and promote trust in the system.
- RQ3: What strategies can be implemented to achieve patient-centric interoperability in healthcare systems, ensuring seamless data exchange and communication among various stakeholders?
  The lack of patient-centric interoperability in healthcare systems hampers seamless data exchange and communication between different stakeholders such as healthcare providers, insurers, and patients. Identifying strategies to establish patient-centric interoperability will facilitate the integration of diverse healthcare entities and foster collaborative patient care. Without addressing this question, patient data may remain fragmented, obstructing the implementation of a truly patient-centric healthcare ecosystem.
- RQ4: How can blockchain-based patient-centric healthcare architectures be designed to ensure compliance with healthcare data regulations and provide patients with clear data ownership?
  The existing research lacks a blockchain-based patient-centric healthcare architecture that fully adheres to healthcare data regulations while ensuring clear data ownership for patients. Current solutions may not provide the necessary data control and compliance mechanisms, potentially hindering widespread adoption in healthcare systems. Addressing this research question will pave the way for a robust and legally compliant patient-centric healthcare blockchain system.

## 5 | Results

This section discusses the results from analyzing patient-centric Blockchain-based Healthcare Architectures. It covers the analysis of publication trends over recent years, distribution by geographic origin, different publication channels, source distribution, contribution types, use cases of the research conducted, main challenges addressed, blockchain platforms used, and smart contract implementations.

### 5.1 | Publication Years

The analysis of publication years reveals that the research papers selected for the systematic literature review predominantly belong to recent years, indicating the emerging nature of Blockchain in patient-centric healthcare architectures as a research area, as demonstrated in Figure 3. The earliest publication identified was in 2018. However, the majority of the selected papers were published in subsequent years. Specifically, in the year 2019, four papers (9.7'5%) were published, while in 2020, six papers (18.75%) were identified. The number of publications increased further in 2021, with nine papers (21.95%) published. In the year 2022, it further increased to ten papers (24.39%). Finally, in the current year, 2023, eleven papers (26.83%) were included in the review. The steady growth in the number of



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

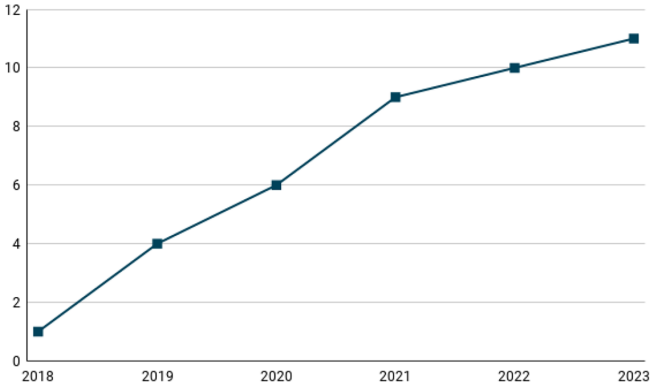



**FIGURE 3** | Timeline for papers extracted for year-wise publication.

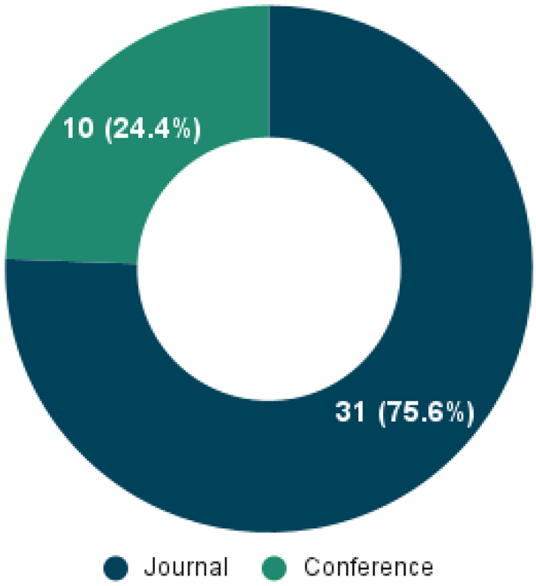



**FIGURE 4** | Publication type of the selected research papers.

publications over the years indicates a growing interest and ongoing research in the field of patient-centric Blockchain-based Healthcare Architectures.

### 5.2 | Publication Type

Regarding the types of publications, 24.39% of the selected papers were presented at conferences, and the remaining 75.61% were published in journals as represented in Figure 4. This observation indicates that the predominant avenue for the dissemination of discoveries pertaining to Patient-centric Blockchain-based Healthcare Architectures is journals, which the research community prefers. The high percentage of journal articles demonstrates the demand for thorough, peer-reviewed research in this evolving field.

### 5.3 | Geographical Distribution

The fact that authors of the research papers included in the systematic literature review came from various countries highlights the interest in patient-centric Blockchain-based Healthcare Architectures around the world. The majority of papers (39.02%) originated from India, demonstrating the substantial contribution of academics and industry experts from the Indian subcontinent to this topic. The USA and China both contributed 9.75% and 12.19% of the papers, respectively, making them important contributors as well. The remaining 39.02% of the papers that were selected came from Saudi Arabia, Ireland, Italy, Taiwan, Sweden, Australia, Bangladesh, Egypt, France, Colombia, Switzerland, Malaysia, and the United Arab Emirates. Figure 5 shows the varied geographical distribution.

### 5.4 | Publication Channel

The selected research papers were published in various reputable publication channels, with the most prominent being IEEE, accounting for 41.46% of the papers. Following closely, the second most common channel was MDPI, publishing 21.95% of the papers. Additionally, ELSEVIER and SPRINGER accounted for 7.3% of the papers each. Other channels, including WILEY, JMIR, GEORG THIEME VERLAG KG, OXFORD UNIV PRESS, ICMJE, NATURE PORTFOLIO, IEEE Access, Frontiers, and ACM, collectively contributed to the remaining 22% of the selected papers. Figure 6 depicts the distribution of publication channels.

### 5.5 | Publication Source

Figure 7 illustrates the distribution of publication sources for the selected research papers on blockchain-based patient-centric healthcare architectures. Out of the total 41 papers, the majority (40 papers) were published by academia, indicating a strong academic interest and involvement in this research domain. In contrast, only one paper was identified as originating from the industry. This skewed distribution highlights the dominant academic contribution in advancing the understanding and development of patient-centric blockchain solutions in the healthcare sector. However, it is essential to acknowledge that industry-driven insights and implementations may play a crucial role in the practical adoption and real-world applications of blockchain technology in healthcare.

### 5.6 | Contributions

The systematic literature review identified three main areas of contribution in the field of Patient-centric Blockchain-based Healthcare Architectures as depicted in Figure 8. The most common contribution type present in 41.5% of the papers was related to Patient-centric Data Management and Control. Security and Privacy Enhancements ranked as the second most addressed contribution area, present in 34.1% of the papers. Interoperability and Efficient Data Sharing were observed as the contribution in the remaining 24.4% of the papers. The study indicates that researchers are actively focusing on enhancing patient data management and security aspects while promoting data sharing and interoperability in the context of Blockchain-based healthcare solutions.



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

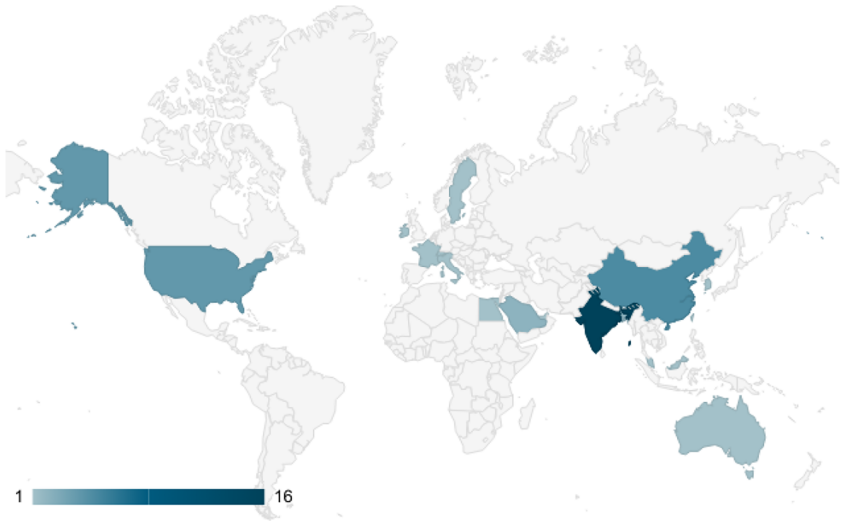



**FIGURE 5** | Geographical distribution of the selected research papers.

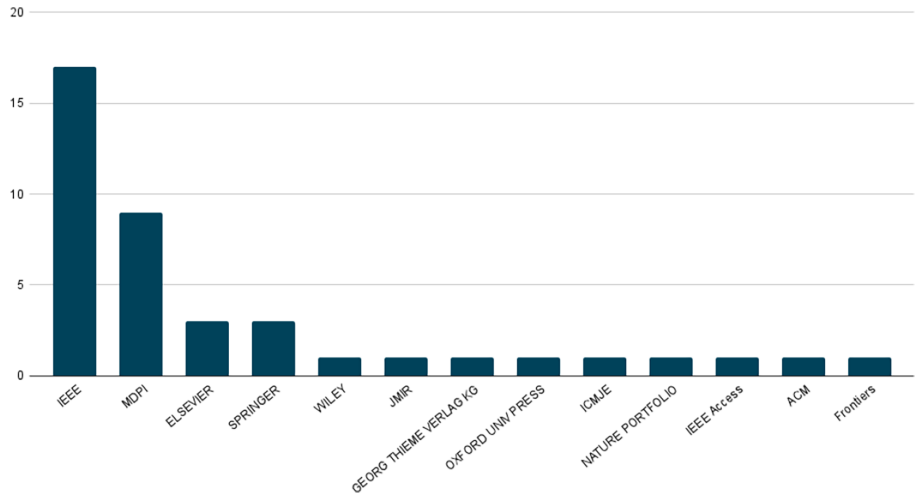



**FIGURE 6** | Publication channel of the selected research papers.

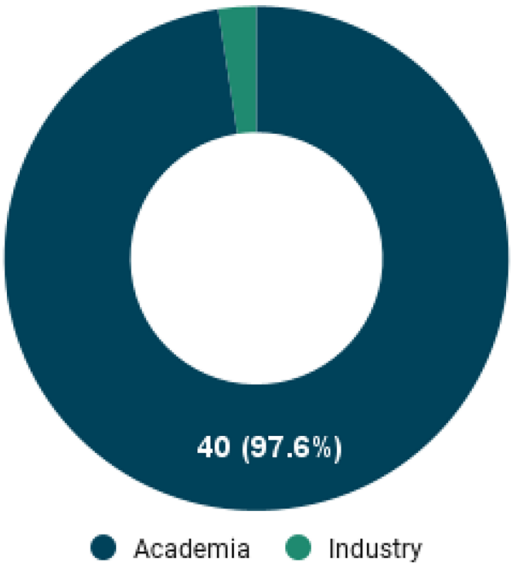



**FIGURE 7** | Publication source of the selected research papers.

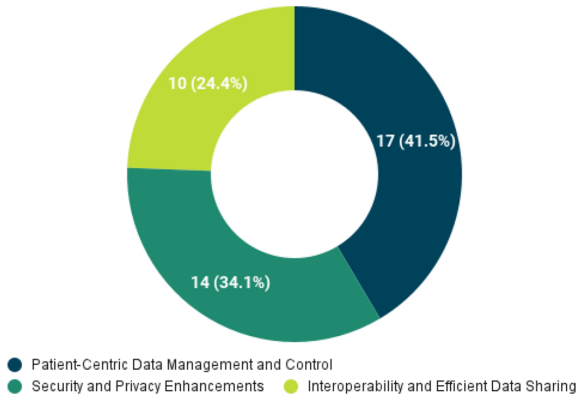



**FIGURE 8** | Contribution of the selected research papers.

## 5.7 | Contribution Type

In terms of contribution types, the selected research papers predominantly proposed Frameworks, accounting for 39% of the papers as demonstrated in Figure 9. System designs were



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

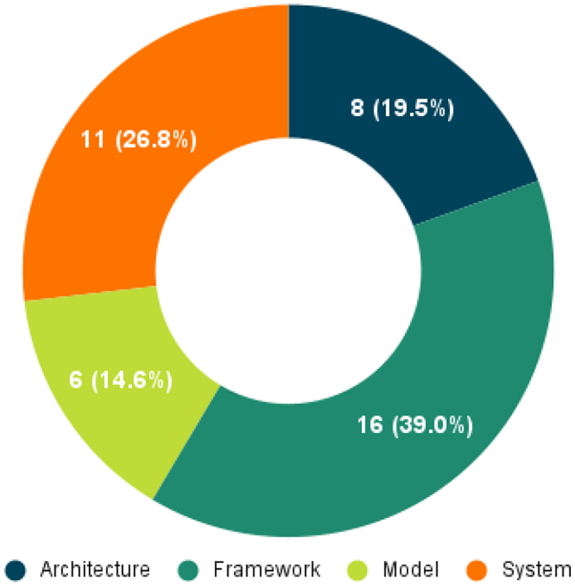



**FIGURE 9** | Contribution type of the selected research papers.

also prevalent, comprising 26.8% of the papers. Models and Architectures were less common but still significant, contributing to 14.6% and 19.5% of the papers, respectively. This diversity in contribution types demonstrates the comprehensive and multidisciplinary nature of research efforts in Patient-centric Blockchain-based Healthcare Architectures.

### 5.8 | Use Cases

Figure 10 represents the various use cases identified in the systematic literature review where Blockchain technology was applied in healthcare settings. The most common use case, present in 43.9% of the papers, was related to Healthcare Data Sharing and Interoperability. Electronic Health Record (EHR) Management was the second most prevalent use case, appearing in 29.3% of the papers. Patient Monitoring and Remote Healthcare, along with some other Miscellaneous use cases, were relatively less common but still relevant, with each being addressed in 9.8% of the papers. Telemedicine and Telehealth accounted for the smallest proportion, with 7.3% of the papers exploring this use case.

### 5.9 | Main Challenges Addressed

The selected research papers extensively addressed various challenges related to Patient-centric Blockchain-based Healthcare architectures, and the results are demonstrated in Figure 11. The most commonly addressed challenge was Data Interoperability and Management, present in 43.9% of the papers. Data privacy and protection were also major concerns, highlighted in 34.1% of the papers. Security and Trust ranked as the third most addressed challenge, accounting for 22% of the papers. This indicates that researchers recognize the importance of overcoming interoperability issues, ensuring data security and trustworthiness, and safeguarding patient data privacy in blockchain-based healthcare systems.

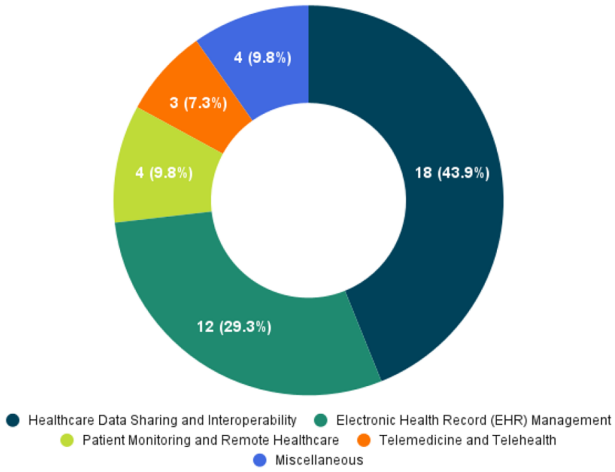



**FIGURE 10** | Use cases identified in the selected research papers.

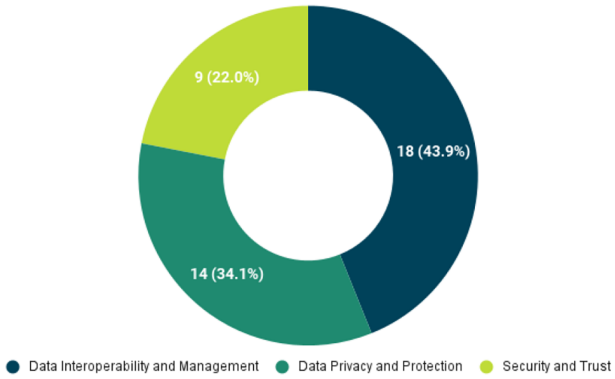



**FIGURE 11** | Main challenges addressed in the selected research papers.

### 5.10 | Blockchain Platforms

The analysis of the blockchain platforms used in the selected research papers showed that Ethereum/Ganache/Remix and Hyperledger Fabric were the most popular options, as shown in Figure 12. In comparison, 48.8 percent of the papers used Ethereum/Ganache/Remix, whereas 31.7 percent of the papers used Hyperledger Fabric. Other blockchain platforms, such as Hyperledger Indy, appeared in a smaller number of papers (7.3% of the papers). This reveals a preference for implementing Patient-centric Blockchain-based Healthcare Architectures on mature and well-known blockchain platforms. The remaining 7.3% of the selected papers utilized other blockchain platforms for their implementation.

### 5.11 | Blockchain Type

The majority of the selected papers (68.3%) implemented patient-centric blockchain-based healthcare architectures using permissioned/private blockchains. This particular blockchain type offers enhanced privacy and control, making it appropriate for use in healthcare applications. The less desirable choice, permissionless/public blockchains, was mentioned in only 7.3% of the articles. Consortium blockchains, offering a balance between



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

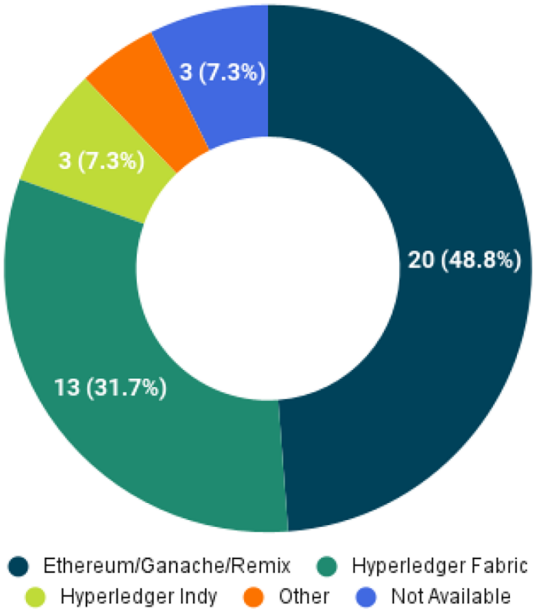



**FIGURE 12** | Blockchain platforms used in the selected research papers.

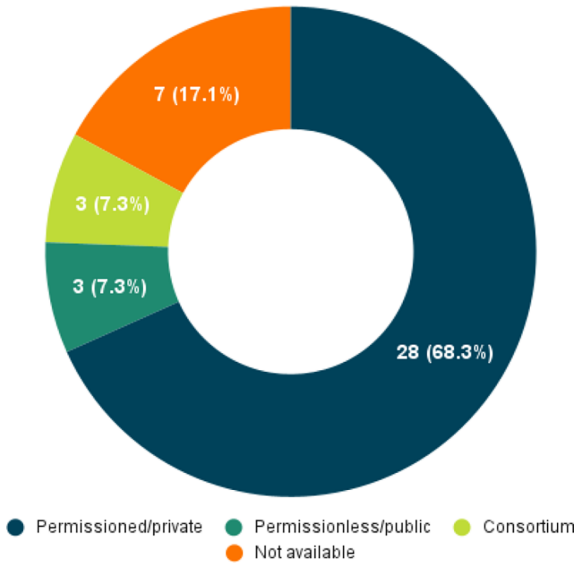



**FIGURE 13** | Types of blockchain used in the selected research papers.

centralization and decentralization, were employed in 7.3% of the papers. A few papers (17.1%) did not specify the type of blockchain used. The type of blockchain used for implementation is demonstrated in Figure 13.

### 5.12 | Smart Contracts

Smart Contracts, an essential feature of blockchain technology, were addressed in the selected papers with varying frequency. The majority of the papers (43.9%) did not provide details regarding the smart contract platforms used. Solidity, a popular smart contract language associated with Ethereum, was employed in

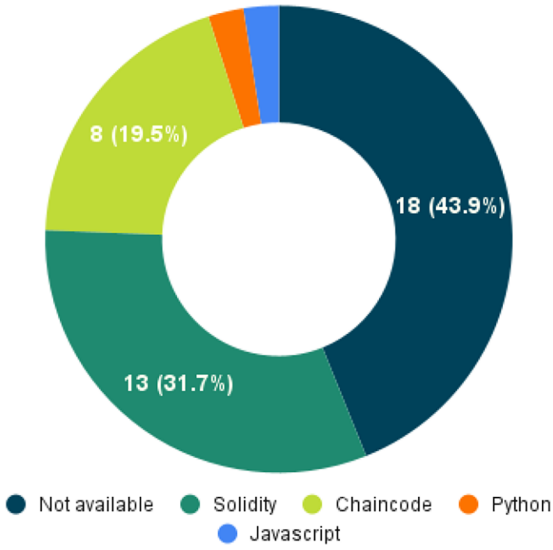



**FIGURE 14** | Smart contracts employed in the selected research papers.

31.7% of the papers that explicitly mentioned the smart contract platform. Chaincode, typically used in Hyperledger Fabric, was featured in 19.5% of the papers. Python and JavaScript were the least common choices for implementing smart contracts, appearing in only 2.4% of the papers each. Figure 14 shows the distribution of smart contracts employed.

These results of the systematic literature review showcase the emerging nature of patient-centric Blockchain-based Healthcare Architectures as a research area, with a surge in publications in recent years. Journals have been the primary dissemination channel, underscoring the demand for peer-reviewed research. Geographically diverse authorship emphasizes global interest, while dominant academic contributions highlight the pursuit of innovative solutions. The identified contributions encompass patient-centric data management, security enhancements, interoperability, and efficient data sharing. Implementation trends reveal a preference for established blockchain platforms and permissioned blockchains for privacy. The varied use cases, challenges addressed, and smart contract platforms employed collectively signify the multidimensional nature of patient-centric blockchain applications.

## 6 | Discussion

This section comprehensively analyzes the results and their implications, addressing the four main research questions. It interprets the findings in the context of existing literature and highlights their significance. The research gaps are identified, paving the way for future research. Figure 15 summarizes the challenges addressed and the future research directions in blockchain-based patient-centric healthcare architectures.

### 6.1 | RQ1

The investigation into enhancing the security and privacy of healthcare data within blockchain-based storage and sharing



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

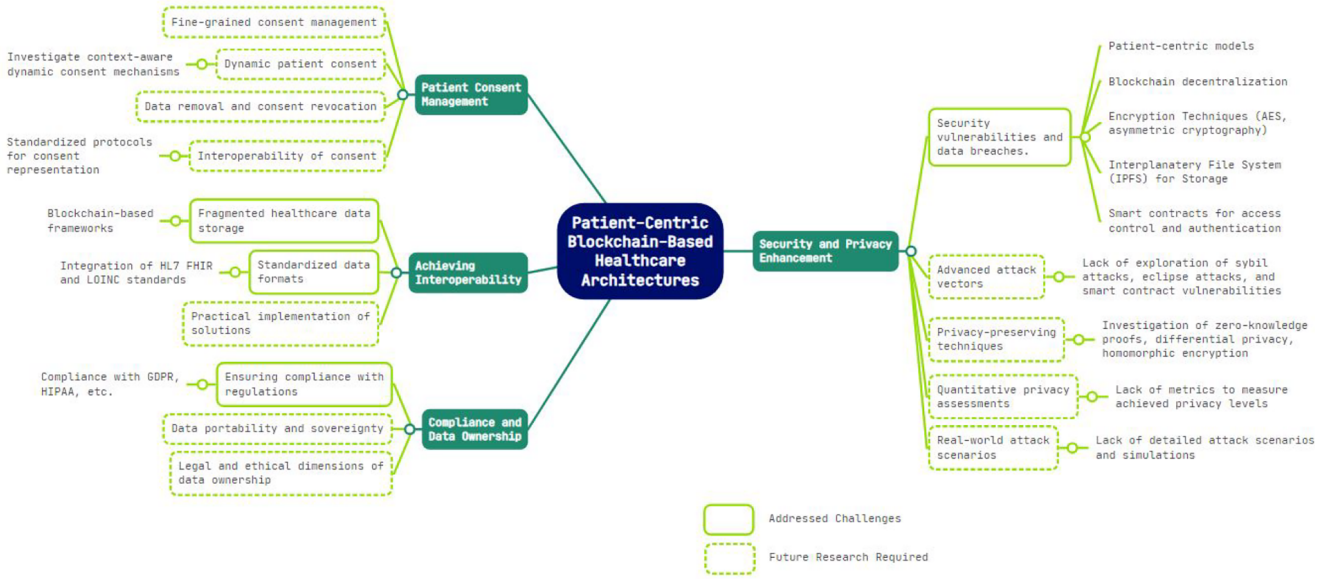



**FIGURE 15** | Summary of addressed challenges and future research directions.

mechanisms has yielded comprehensive insights. The analysis of the selected research papers reveals a strong focus on addressing security and privacy challenges in healthcare data management. These challenges are significantly underscored by the vulnerability of patient information to external and internal attacks, and the growing concerns regarding data breaches [9, 42, 49, 50, 52, 53, 56, 57, 59, 61, 65–77]. Several proposed solutions emphasize the need for patient-centric models [53, 55–60, 65, 66, 68–71, 75–78] ensuring that patients have complete control over their health data [42, 49, 53, 55, 56, 58–60, 69, 71]. These models often utilize blockchain technology to ensure decentralization, transparency, and immutability of patient records. Encryption techniques, such as AES and asymmetric cryptography [79], are widely adopted to secure sensitive medical data [42, 48, 49, 52, 55, 56, 58, 66–69, 75, 77, 78, 80, 81]. Off-chain solutions, such as Inter-Planetary File System (IPFS), are integrated for decentralized and secure data storage [42, 48–50, 52, 55, 57–59, 61, 65–68, 70, 71, 77, 80, 81]. Smart contracts are employed for access control, authentication, and data sharing [9, 42, 48, 49, 51–53, 55, 57–59, 61, 65, 68–72, 75–78]. Additionally, research acknowledges the importance of trust-building mechanisms, audit trails, and user-friendly interfaces to ensure accountability and ease of use. While the majority of studies focus on secure data sharing and storage, some explore additional aspects such as patient identity management, telemedicine, interoperability, and remote patient monitoring. In conclusion, the synthesis of these research papers underscores the significance of patient-centric, blockchain-enabled healthcare data management to enhance security and privacy. Utilizing encryption, decentralized storage, access control, and trust-building mechanisms, these studies collectively contribute to a holistic approach that empowers patients while safeguarding their sensitive information.

The examination of the research landscape reveals several crucial dimensions that require consideration when addressing the security and privacy of healthcare data within blockchain-based storage and sharing mechanisms. While the surveyed studies offer valuable insights, certain areas remain under-explored and merit attention to comprehensively enhance security and privacy safeguards [82].

- The existing research acknowledges the potential vulnerability of blockchain systems to various security attacks, such as 51% attacks, selfish mining attacks, transaction data malleability, and deanonymization [56]. However, a deeper exploration of advanced attack vectors, including but not limited to sybil attacks, eclipse attacks, and sophisticated smart contract vulnerabilities, would contribute to a more comprehensive understanding of potential threats.
- While encryption is widely adopted for data security, specific privacy-preserving techniques, such as zero-knowledge proofs, differential privacy, and homomorphic encryption, are often overlooked. Investigating these techniques and their applicability to healthcare data sharing would strengthen the overall privacy framework.
- While studies recognize the importance of privacy, few offer quantitative assessments of achieved privacy levels. Incorporating robust metrics to quantify the degree of privacy protection achieved by different blockchain-based healthcare systems would provide a more objective basis for comparison.
- Many research papers lack concrete exploration of real-world attack scenarios that could compromise patient data security and privacy. Incorporating detailed attack scenarios and simulations can help in assessing the resilience of proposed systems and aid in devising effective countermeasures.

By addressing these gaps and incorporating insights from emerging attack vectors, cutting-edge privacy techniques, and real-world scenarios, the security and privacy of healthcare data within blockchain-based systems can be significantly enhanced. This comprehensive approach ensures a more robust and resilient framework for safeguarding patient information.



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

### 6.2 | RQ2

The analysis of the collected research papers reveals a diverse range of approaches and strategies to effectively identify and manage patients' consent for data sharing within blockchain-based healthcare architectures, where patients play a pivotal role in granting and managing consent for data sharing. While patient consent is not explicitly discussed in all papers, it can be inferred that patient-centric control and privacy are central themes in many proposed solutions. The surveyed research papers propose a variety of methods to address this important aspect.

Numerous studies propose systems that empower patients to retain full control over their health data. Through blockchain technology, patients can grant or revoke access rights to healthcare providers, ensuring that data sharing occurs only with explicit patient consent [9, 42, 48, 49, 51, 54, 57–59, 67, 69–71, 74, 75, 77, 78, 81]. Several research papers highlight the role of smart contracts in enforcing patient consent and access control. Smart contracts enable the creation of permissioned data access, ensuring that healthcare providers can only access patient data upon obtaining explicit consent [42, 55, 59, 61, 66, 68–70, 75, 76, 81]. Some papers discuss patient identity management within blockchain-based architectures, implying that patients control the access and sharing of their personal health information, which would likely involve granting consent [72, 75, 76, 78]. Certain systems incorporate electronic consent mechanisms, ensuring that patients provide explicit consent for data access and sharing. These consent records are securely recorded on the blockchain, establishing a traceable and auditable consent history [50, 76, 78, 81].

The current research on blockchain-based healthcare architectures predominantly emphasizes patient-centric control and privacy as essential components. Patient consent is integral to these systems, where patients are empowered to manage and grant access to their health data through various mechanisms, including smart contracts, encryption, electronic consent, and traceable records on the blockchain. These approaches collectively contribute to enhancing patient autonomy and data privacy in the context of data sharing within healthcare systems. In the existing research, patient-controlled consent mechanisms in blockchain-based healthcare architectures include traceable consent history, which ensures transparency and accountability in data sharing, electronic consent records, which securely record approvals on the blockchain, and patient-managed access, which lets people grant or revoke data access. Certain unexplored areas in the current research landscape are as follows:

- While several research papers emphasize patient control over data sharing, there's a need for more nuanced granularity in consent management. Future research could focus on the development of consent models that allow patients to specify not only which healthcare entities have access to their data but also the specific types of data each entity can access. This would provide patients with fine-grained control over their information, enhancing privacy and customization [83].
- Dynamic patient consent is an evolving concept that aligns with the dynamic nature of healthcare interactions. However, current research mainly focuses on static or initial consent. Future investigations could explore context-aware dynamic consent mechanisms. These would enable patients to grant or revoke consent based on changing circumstances, such as the purpose of data access, time frame, or specific healthcare scenarios.
- Current research primarily covers data access control; however, limited attention is given to data removal and consent revocation. Future studies could delve into the mechanisms that enable patients to not only restrict data access but also request the removal of their data from blockchain systems while adhering to data integrity requirements.
- The interoperability of consent across different blockchain-based healthcare platforms remains an underexplored area. Standardized protocols can be investigated for consent representation and sharing, ensuring seamless interoperability between various systems. This would facilitate consistent consent management and data sharing practices.

### 6.3 | RQ3

The investigation into strategies like standardization of data formats and protocols and integration of standards across different domains for achieving patient-centric interoperability in healthcare systems has revealed several key insights. The analyzed research papers collectively underscore the significance of interoperability as a fundamental challenge in modern healthcare systems. Interoperability concerns are often driven by the fragmented nature of healthcare data storage, the lack of standardized data formats, and the need for secure and seamless communication between different stakeholders, including patients, healthcare providers, and institutions.

A recurrent theme in the selected papers is the utilization of blockchain technology to establish patient-centric interoperability. Blockchain offers decentralized and tamper-resistant storage of health records, granting patients control over their data. Several systems, such as MediTrans [58], MedHypChain [69], and VigilRx [57], propose blockchain-based frameworks that ensure interoperability while allowing the secure sharing of patient data among healthcare entities. Additionally, the implementation of patient-centric architectures, such as patient-centric EHR systems [52, 54], enables the unified storage of patient data from various sources, promoting interoperability and facilitating expedited data exchange. The importance of data standardization is highlighted by various studies. Patient data often resides in heterogeneous systems using diverse formats. The following are the prevalent methods to promote interoperability in the studied papers:

- Interoperability can be supported by the integration of established standards and frameworks such as the HL7, Fast Healthcare Interoperability Resources (FHIR) [9, 42, 72, 77, 80] and Logical Observation Identifiers Names and Codes (LOINC) [50]. The adoption of such standards helps bridge the gap between different healthcare providers and systems,



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

ensuring that data is exchanged and interpreted consistently [9, 77].

- Many solutions leverage decentralization to ensure the availability and accessibility of data. Interplanetary File Storage (IPFS) is used in various papers [51, 53], to enable distributed storage of medical data, enhancing interoperability.
- In addressing the challenge of interoperability, one approach relies on blockchain relays, essentially smart contracts running on one blockchain that can verify events on another blockchain. Another method involves creating trusted gateways to relay details of cross-chain transactions, particularly relevant for private blockchains. Additionally, the concept of a "blockchain of blockchains" has gained attention to achieve interoperability between different blockchain networks [84].

Although blockchain-based solutions offer promising avenues for enhancing interoperability, there is a need for further exploration of their practical implementation and evaluation of their effectiveness. Additionally, many of the reviewed papers focus on specific aspects of interoperability, warranting comprehensive investigations that encompass the broader healthcare ecosystem. Furthermore, while standards like FHIR contribute to interoperability, deeper analyses are required to assess their impact across diverse healthcare settings. Exploring the integration of multiple standards and protocols could enhance the interoperability of healthcare systems across various domains [9, 77].

In conclusion, patient-centric interoperability in healthcare systems necessitates a combination of blockchain technology, standardized frameworks, and patient-controlled access to health data. While the current research offers valuable insights, further exploration and validation are required to establish robust and scalable solutions that address the multifaceted challenges of achieving seamless data exchange and communication among various healthcare stakeholders.

### 6.4 | RQ4

Designing a patient-centric healthcare architecture based on blockchain technology is a crucial step towards ensuring compliance with healthcare data regulations and providing clear data ownership to patients. Most of the proposed systems emphasize patient data privacy and ownership through blockchain-based solutions [48, 51, 53, 54, 56, 61, 66]. These systems enable patients to control access to their health data, share it securely, and define access rights [60, 67, 77]. Patients are empowered to grant or revoke access, ensuring that only authorized parties can view their medical records [52, 68, 69]. The following observations were made by analysing the current research:

- Numerous systems explicitly aim to ensure compliance with healthcare data regulations, such as General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) [9, 50, 51, 56, 58, 66, 70, 76]. Blockchain's tamper-proof nature and access controls help meet regulatory requirements [59, 71, 80]. Permissioned blockchain networks ensure data integrity and enforce privacy, contributing to compliance [65, 75, 78].
- However, future research could delve deeper into the alignment of blockchain-based healthcare systems with evolving and stringent healthcare data regulations such as GDPR, HIPAA, and other region-specific laws. Ensuring a robust and adaptable regulatory framework within blockchain solutions is crucial to address potential regulatory challenges.
- While blockchain offers data ownership to patients, ensuring data portability and sovereignty across different healthcare providers and platforms remains a complex issue. Mechanisms facilitating seamless health data transfer amidst diverse healthcare ecosystems need further exploration.
- Additionally, the legal and ethical dimensions of patient-controlled data ownership require further attention [85]. This could include examining legal frameworks for disputes, liability, and data breach scenarios within decentralized healthcare systems.

### 6.5 | Future Research Directions

Existing research has revealed several critical gaps and challenges in blockchain-based healthcare data architectures. Future research directions should focus on the following key areas to advance the field and improve the security, privacy, and efficacy of these systems.

- A thorough examination of advanced attack vectors such as sybil attacks, eclipse attacks, and sophisticated smart contract vulnerabilities is required. Comprehensive assessments of potential threats other than the commonly recognized ones, such as 51 percent attacks and deanonymization, will result in a more robust security framework.
- The application of specific privacy-preserving techniques such as zero-knowledge proofs, differential privacy, and homomorphic encryption to blockchain-based healthcare data sharing requires further investigation. Researchers should investigate the feasibility and efficacy of these techniques for improving patient privacy.
- It is critical to develop fine-grained consent models that allow patients to specify not only which entities can access their data but also the types of data each entity can access. Dynamic consent mechanisms that are aligned with evolving healthcare interactions, as well as mechanisms for data removal and consent revocation, should be investigated.
- For improved interoperability, standardization of consent representation and sharing protocols across various blockchain-based healthcare platforms should be pursued. This will ensure that consent management practices are consistent.
- Mechanisms for seamless health data transfer across diverse healthcare ecosystems, as well as an examination of the legal and ethical dimensions of patient-controlled data ownership, necessitate additional research. In decentralized healthcare systems, legal frameworks for dispute resolution and liability should be thoroughly examined.

Incorporating these research directions will result in the creation of more robust and secure patient-centric blockchain-based healthcare systems.



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

## 7 | Conclusion

In conclusion, this research paper presents a comprehensive synthesis of the current landscape of patient-centric blockchain-based healthcare architectures. The systematic literature review highlights the increasing interest and ongoing research in this evolving field. The analysis reveals that most selected research papers focus on enhancing security and privacy aspects of patient data within blockchain systems, emphasizing patient-controlled models and encryption techniques. Patient consent management is a key theme, with various approaches utilizing blockchain and smart contracts to ensure patients' explicit control over data sharing. Interoperability is identified as a fundamental challenge in modern healthcare, and blockchain is proposed as a solution to achieve patient-centric data exchange. Compliance with healthcare data regulations is addressed through blockchain's tamper-proof nature and access controls.

While the analyzed research papers contribute valuable insights, several avenues for future research are identified. These include deeper exploration of advanced security threats, incorporation of privacy-preserving techniques, quantitative assessment of achieved privacy levels, investigation of real-world attack scenarios, and exploration of dynamic consent mechanisms. Additionally, research can further delve into nuanced consent models, data removal and revocation mechanisms, and standardized protocols for consent representation and sharing. Achieving seamless data exchange and communication among various healthcare stakeholders also requires comprehensive solutions that encompass the broader healthcare ecosystem. Moreover, the alignment of blockchain-based healthcare systems with evolving and stringent healthcare data regulations needs further scrutiny. By addressing these gaps and challenges, the field of patient-centric blockchain-based healthcare architectures can achieve more robust, secure, and patient-centered solutions that empower individuals while safeguarding their sensitive health data.

### Data Availability Statement

The authors have nothing to report.

### References


1. D. C. Nguyen, P. N. Pathirana, M. Ding, and A. Seneviratne, "A Cooperative Architecture of Data Offloading and Sharing for Smart Healthcare With Blockchain," in *Proceedings of the IEEE International Conference on Blockchain and Cryptocurrency, ICBC 2021* (IEEE, 2021), 5.

2. I. Keshta and A. Odeh, "Security and Privacy of Electronic Health Records: Concerns and Challenges," *Egyptian Informatics Journal* 22, no. 7 (2021): 177–183.

3. S. Alam, S. Bhatia, M. Shuaib, et al., "An Overview of Blockchain and IoT Integration for Secure and Reliable Health Records Monitoring," *Sustainability* 15 (2023): 5660.

4. P. Uppamma and S. Bhattacharya, "Diabetic Retinopathy Detection: A Blockchain and African Vulture Optimization Algorithm-Based Deep Learning Framework," *Electronics (Switzerland)* 12, no. 3 (2023): 742, https://doi.org/10.3390/electronics12030742.

5. L. B. Harman, C. A. Flite, and K. Bond, "Electronic Health Records: Privacy, Confidentiality, and Security," *Virtual Mentor* 14 (2012): 712–719.

6. S. O. Zandieh, K. Yoon-Flannery, G. J. Kuperman, D. J. Langsam, D. Hyman, and R. Kaushal, "Challenges to EHR Implementation in Electronic- Versus Paper-Based Office Practices," *Journal of General Internal Medicine* 23, no. 6 (2008): 755–761, https://doi.org/10.1007/s11606-008-0573-5.

7. B. K. Rai, S. Fatima, and K. Satyarth, "Patient-Centric Multichain Healthcare Record," *International Journal of E-Health and Medical Communications* 13, no. 4 (2022): 1–14, https://doi.org/10.4018/IJEHMC.309439.

8. S. M. Cleveland and M. Haddara, "IoT for Diabetics: A User Perspective," in *Intelligent Computing. Lecture Notes in Networks and Systems*, vol. 285, ed. K. Arai (Springer, 2021), https://doi.org/10.1007/978-3-030-80129-8_13.

9. A. Dubovitskaya, F. Baig, Z. Xu, et al., "ACTION-EHR: Patient-Centric Blockchain-Based Electronic Health Record Data Management for Cancer Care," *Journal of Medical Internet Research* 22, no. 8 (2020): e13598.

10. VDD, "Medical Internet of Things and Big Data in Healthcare," *Hir* 22, no. 3 (2016): 156–163.

11. A. I. Stoumpos, F. Kitsios, and M. A. Talias, "Digital Transformation in Healthcare: Technology Acceptance and Its Applications," *International Journal of Environmental Research and Public Health* 2 (2023): 20.

12. N. Duong-Trung, H. X. Son, H. T. Le, and T. T. Phan, "On Components of a Patient-Centered Healthcare System Using Smart Contract," in *ACM International Conference Proceeding Series* (ACM, 2020), 31–35.

13. X. Yue, H. Wang, D. Jin, M. Li, and W. Jiang, "Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain With Novel Privacy Risk Control," *Journal of Medical Systems* 10 (2016): 40.

14. "Benefits of Blockchain - IBM Blockchain," (2023), https://www.ibm.com/topics/benefits-of-blockchain.

15. N. R. Pradhan, S. S. Rout, and A. P. Singh, "Blockchain Based Smart Healthcare System for Chronic -Illness Patient Monitoring," in *Proceedings of the 3rd International Conference on Energy, Power and Environment: Towards Clean Energy Technologies, ICEPE 2020* (IEEE, 2021), 3.

16. A. A. Zaidan, B. B. Zaidan, A. Al-Haiqi, M. L. M. Kiah, M. Hussain, and M. Abdulnabi, "Evaluation and Selection of Open-Source EMR Software Packages Based on Integrated AHP and TOPSIS," *Journal of Biomedical Informatics* 53, no. 2 (2015): 390–404.

17. K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, "Big Data Security and Privacy in Healthcare: A Review," *Procedia Computer Science* 1, no. 113 (2017): 73–80.

18. M. Ayaz, M. F. Pasha, M. Y. Alzahrani, R. Budiarto, and D. Stiawan, "The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities," *JMIR Medical Informatics* 9, no. 7 (2021): e21929.

19. S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," www.bitcoin.org.

20. H. Wu, Y. Shang, L. Wang, L. Shi, K. Jiang, and J. Dong, "A Patient-Centric Interoperable Framework for Health Information Exchange via Blockchain," in *ACM International Conference Proceeding Series* (ACM, 2019), 76–80.

21. I. Bashir, *Mastering Blockchain*, vol. 161 (Packt, 2017), https://www.packtpub.com/product/mastering-blockchain/9781787125445.

22. G. Lodha, M. Pillai, A. Solanki, S. Sahasrabudhe, and A. Jarali, "Healthcare System Using Blockchain," in *Proceedings - 5th International Conference on Intelligent Computing and Control Systems* (ICICCS 2021, 2021), 274–281.




15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

23. B. Zou, F. Nie, L. Xiao, T. Zhang, and C. Zhu, "MeDA: Using Blockchain for Patient-Controlled Medical Data Auditing in Institutions," in *Proceedings - 22nd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGrid 2022* (IEEE, 2022), 297–306.

24. Cherednichenko, "SDesigning a Blockchain Architecture: Types, Use Cases, and Challenges," 2023.

25. N. Szabo, "Formalizing and Securing Relationships on Public Networks," *First Monday* 2, no. 9 (1997): 1–12.

26. E. W. Karlson, N. T. Boutin, A. G. Hoffnagle, and N. L. Allen, "Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. Journal of," *Personalized Medicine* 6, no. 1 (2016): 2.

27. "Replacing Paper Informed Consent With Electronic Informed Consent for Research in Academic Medical Centers: A Scoping Review," (2023), https://pubmed.ncbi.nlm.nih.gov/32477626/.

28. F. Albalwy, A. Brass, and A. Davies, "A Blockchain-Based Dynamic Consent Architecture to Support Clinical Genomic Data Sharing (ConsentChain): Proof-of-Concept Study," *JMIR Medical Informatics* 9, no. 11 (2021): e27816, https://doi.org/10.2196/27816.

29. Synopsys, "What Is Blockchain and How Does It Work?," (2023), https://www.synopsys.com/glossary/what-is-blockchain.html.

30. Edureka, "Different Types of Blockchain and Why We Need Them," 2023.

31. R. Stevanovic, "Blockchain From Scratch: Permissioned and Permissionless Blockchains," 2023.

32. M. T. Oliveira, C. R. Carrara, N. C. Fernandes, et al., "Towards a Performance Evaluation of Private Blockchain Frameworks Using a Realistic Workload," in *Proceedings of the 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)* (2019), 180–187, https://doi.org/10.1109/ICIN.2019.8685888.

33. "Holochain - A Framework for Distributed Applications," (2020), https://holochain.org/.

34. "Hyperledger Fabric - Hyperledger Foundation," (2023), https://www.hyperledger.org/use/fabric.

35. Hyperledger Foundation, "Announcing Hyperledger Besu," (2019), https://www.hyperledger.org/blog/2019/08/29/announcing-hyperledger-besu.

36. A. S. dePedro, D. Levi, and L. I. Cuende, "Witnet: A Decentralized Oracle Network Protocol," 2017.

37. "EOS.IO Technical White Paper," (2023), https://steemit.com/eos/@eosio/eos-io-technical-white-paper.

38. R. Deviations, "The Ripple Protocol Consensus Algorithm," (2023), https://reasonabledeviations.com/notes/papers/ripple_consensus_protocol/.

39. S. Nzuva, "Smart Contracts Implementation, Applications, Benefits, and Limitations," 2019.

40. H. Wu, Y. Shang, L. Wang, L. Shi, K. Jiang, and J. Dong, "A Patient-Centric Interoperable Framework for Health Information Exchange via Blockchain," in *Proceedings of the 2019 2nd International Conference on Blockchain Technology and Applications* (ACM, 2019), 76–80.

41. M. Vinodhini and M. Prakash, "A Patient-Centric Doctor Referral Model Based on Hyperledger Chaincode," in *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)* (2021), 1–5, https://doi.org/10.1109/CSITSS54238.2021.9683481.

42. Y. Zhuang, L. R. Sheets, Y. W. Chen, Z. Y. Shae, J. J. P. Tsai, and C. R. Shyu, "A Patient-Centric Health Information Exchange Framework Using Blockchain Technology," *IEEE Journal of Biomedical and Health Informatics* 24, no. 8 (2020): 2169–2176.

43. A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "MedRec: Using Blockchain for Medical Data Access and Permission Management," in *2016 IEEE 2nd International Conference on Open and Big Data (OBD)* (2016), 25–30, https://doi.org/10.1109/OBD.2016.11.

44. S. Alexaki, G. Alexandris, V. Katos, and E. N. Petroulakis, "Blockchain-Based Electronic Patient Records for Regulated Circular Healthcare Jurisdictions," in *Proceedings of the IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, CAMAD* (IEEE, 2018).

45. M. T. D. Oliveira, "Towards a Blockchain-Based Secure Electronic Medical Record for Healthcare Applications," in *Proceedings of the IEEE International Conference on Communications* (IEEE, 2019).

46. M. Chen, T. Malook, A. U. Rehman, et al., "Blockchain-Enabled Healthcare System for Detection of Diabetes," *Journal of Information Security and Applications* 58 (2021): 102771, https://doi.org/10.1016/j.jisa.2021.102771.

47. Y. Chen, S. Ding, Z. Xu, H. Zheng, and S. Yang, "Blockchain-Based Medical Records Secure Storage and Medical Service Framework," *Journal of Medical Systems* 43, no. 1 (2018): 5, https://doi.org/10.1007/s10916-018-1121-4.

48. G. S. Reen, M. Mohandas, and S. Venkatesan, "Decentralized Patient Centric e-Health Record Management System Using Blockchain and IPFS," in *Proceedings of the 2019 IEEE Conference on Information and Communication Technology, CICT 2019* (IEEE, 2019).

49. H. Kordestani, K. Barkaoui, and W. Zahran, "HapiChain: A Blockchain-Based Framework for Patient-Centric Telemedicine," in *Proceedings of the 2020 IEEE 8th International Conference on Serious Games and Applications for Health, SeGAH 2020* (IEEE, 2020).

50. K. Kanagi, C. C. Y. Ku, L. K. Lin, and W. H. Hsieh, "Efficient Clinical Data Sharing Framework Based on Blockchain Technology," *Methods of Information in Medicine* 59, no. 6 (2020): 193–204.

51. V. Mani, P. Manickam, Y. Alotaibi, S. Alghamdi, and O. I. Khalaf, "Hyperledger Healthchain: Patient-Centric Ipfs-Based Storage of Health Records," *Electronics (Switzerland)* 10, no. 23 (2021): 1–14.

52. M. U. Chelladurai, D. S. Pandian, and D. K. Ramasamy, "A Blockchain Based Patient Centric Electronic Health Record Storage and Integrity Management for e-Health Systems," *Health Policy and Technology* 10, no. 4 (2021): 1–10.

53. M. Y. Jabarulla and H. N. Lee, "Blockchain-Based Distributed Patient-Centric Image Management System," *Applied Sciences (Switzerland)* 11, no. 1 (2021): 1–20.

54. T. Fatokun, A. Nag, and S. Sharma, "Towards a Blockchain Assisted Patient Owned System for Electronic Health Records," *Electronics (Switzerland)* 10, no. 5 (2021): 1–14.

55. A. N. Gohar, S. A. Abdelmawgoud, and M. S. Farhan, "A Patient-Centric Healthcare Framework Reference Architecture for Better Semantic Interoperability Based on Blockchain, Cloud, and IoT," *IEEE Access* 10 (2022): 92137–92157.

56. C. Pighini, "SynCare: An Innovative Remote Patient Monitoring System Secured by Cryptography and Blockchain," in *IFIP Advances in Information and Communication Technology* (Springer Science and Business Media Deutschland GmbH, 2022), 73–89.

57. A. Taylor, A. Kugler, P. B. Marella, and G. G. Dagher, "VigilRx: A Scalable and Interoperable Prescription Management System Using Blockchain," *IEEE Access* 10 (2022): 25973–25986.

58. M. George and A. M. Chacko, "MediTrans–Patient-Centric Interoperability Through Blockchain," *International Journal of Network Management* 32, no. 3 (2022): 1–18.



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

59. A. Mudaliar, "Blockchain-Based User-Centric Electronic Health Record Management System," in *2022 IEEE International Conference on Blockchain and Distributed Systems Security, ICBDS* (Institute of Electrical and Electronics Engineers Inc, 2022).

60. R. A. Abutaleb, S. S. Alqahtany, and T. A. Syed, "Integrity and Privacy-Aware, Patient-Centric Health Record Access Control Framework Using a Blockchain," *Applied Sciences (Switzerland)* 13, no. 2 (2023): 1028, https://doi.org/10.3390/app13021028.

61. B. K. Rai, "PcBEHR: Patient-Controlled Blockchain Enabled Electronic Health Records for Healthcare 4.0," *Health Services & Outcomes Research Methodology* 23, no. 1 (2023): 80–102.

62. F. A. Reegu, H. Abas, Y. Gulzar, et al., "Blockchain-Based Framework for Interoperable Electronic Health Records for an Improved Healthcare System," *Sustainability* 15 (2023): 6337.

63. J. Jayabalan and N. Jeyanthi, "Scalable Blockchain Model Using Off-Chain IPFS Storage for Healthcare Data Security and Privacy," *Journal of Parallel and Distributed Computing* 164 (2022): 152–167.

64. M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide* (John Wiley & Sons, 2008).

65. M. A. Uddin, A. Stranieri, I. Gondal, and V. Balasubramanian, "Continuous Patient Monitoring With a Patient Centric Agent: A Block Architecture," *IEEE Access* 6 (2018): 32700–32726.

66. N. K. Dewangan and P. Chandrakar, "Patient-Centric Token-Based Healthcare Blockchain Implementation Using Secure Internet of Medical Things," *IEEE Transactions on Computational Social Systems* 10, no. 6 (2022): 3109–3119, https://doi.org/10.1109/TCSS.2022.3194872.

67. S. Jiang, H. Wu, and L. Wang, "Patients-Controlled Secure and Privacy-Preserving EHRs Sharing Scheme Based on Consortium Blockchain," in *2019 IEEE Global Communications Conference (GLOBECOM)* (2019), 1–6, https://doi.org/10.1109/GLOBECOM38437.2019.9013220.

68. F. Zhao, J. Yu, and B. Yan, "Towards Cross-Chain Access Control Model for Medical Data Sharing," *Procedia Computer Science* 202 (2022): 330–335.

69. M. Kumar and S. Chand, "MedHypChain: A Patient-Centered Interoperability Hyperledger-Based Medical Healthcare System: Regulation in COVID-19 Pandemic," *Journal of Network and Computer Applications* 179 (2021): 102975.

70. D. T. Harrell, M. Usman, L. Hanson, et al., "Technical Design and Development of A Self-Sovereign Identity Management Platform for Patient-Centric Health Care Using Blockchain Technology," *Blockchain Healthcare Today* 22 (2022): 5, https://doi.org/10.30953/bhty.v5.196.

71. S. A. H. Mohsan, A. Razzaq, S. A. K. Ghayyur, H. K. Alkahtani, N. al-Kahtani, and S. M. Mostafa, "Decentralized Patient-Centric Report and Medical Image Management System Based on Blockchain Technology and the Inter-Planetary File System," *International Journal of Environmental Research and Public Health* 19, no. 22 (2022): 14641.

72. A. P. Singh, N. R. Pradhan, A. K. Luhach, et al., "A Novel Patient-Centric Architectural Framework for Blockchain-Enabled Healthcare Applications," *IEEE Transactions on Industrial Informatics* 17, no. 8 (2021): 5779–5789.

73. L. Ismail, H. Materwala, and M. A. B. Khan, "Performance Evaluation of a Patient-Centric Blockchain-Based Healthcare Records Management Framework," in *ACM International Conference Proceeding Series* (Association for Computing Machinery, 2020), 39–50.

74. B. Toshniwal, P. Podili, R. J. Reddy, and K. Kataoka, "PACEX: PAtient-Centric EMR eXchange in Healthcare Systems Using Blockchain," in *Proceedings of the 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2019* (IEEE, 2019), 954–960.

75. H. Ghayvat, M. Sharma, P. Gope, and P. K. Sharma, "SHARIF: Solid Pod-Based Secured Healthcare Information Storage and Exchange Solution in Internet of Things," *IEEE Transactions on Industrial Informatics* 18, no. 8 (2022): 5609–5618.

76. A. Khurshid, C. Holan, C. Cowley, et al., "Designing and Testing a Blockchain Application for Patient Identity Management in Healthcare," *JAMIA Open* 4, no. 3 (2021): ooaa073.

77. H. Kurdi, S. Alsalamah, A. Alatawi, S. Alfaraj, L. Altoaimy, and S. H. Ahmed, "Healthybroker: A Trustworthy Blockchain-Based Multi-Cloud Broker for Patient-Centered Ehealth Services," *Electronics (Switzerland)* 8, no. 6 (2019): 602, https://doi.org/10.3390/electronics8060602.

78. K. P. Satamraju and M. Balakrishnan, "A Secured Healthcare Model for Sensor Data Sharing With Integrated Emotional Intelligence," *IEEE Sensors Journal* 22, no. 16 (2022): 16306–16313, https://doi.org/10.1109/JSEN.2022.3189268.

79. A. Luntovskyy and D. Guetter, "Cryptographic Technology Blockchain and Its Applications," in *Advances in Information and Communication Technologies: Processing and Control in Information and Communication Systems. Lecture Notes in Electrical Engineering*, vol. 560 (Springer, 2019), 14–33.

80. S. Monga and D. Singh, "MRBSChain a Novel Scalable Medical Records Binance Smart Chain Framework Enabling a Paradigm Shift in Medical Records Management," *Scientific Reports* 12, no. 1 (2022): 17660.

81. M. A. H. Wadud, T. M. A. U. H. Bhuiyan, M. A. Uddin, and M. M. Rahman, "A Patient Centric Agent Assisted Private Blockchain on Hyperledger Fabric for Managing Remote Patient Monitoring," in *Proceedings of 2020 11th International Conference on Electrical and Computer Engineering, ICECE 2020* (Institute of Electrical and Electronics Engineers Inc, 2020), 194–197.

82. R. Guo, H. Shi, Q. Zhao, and D. Zheng, "Secure Attribute-Based Signature Scheme With Multiple Authorities for Blockchain in Electronic Health Records Systems," *IEEE Access* 6 (2018): 11676–11686.

83. H. Wang and Y. Song, "Secure Cloud-Based EHR System Using Attribute-Based Cryptosystem and Blockchain," *Journal of Medical Systems* 8, no. 42 (2018): 1–9, https://doi.org/10.1007/s10916-018-0994-6.

84. "How Can We Get Blockchains to Talk to Each Other? - IEEE Spectrum," (2023), https://spectrum-ieee-org.cdn.ampproject.org/c/s/spectrum.ieee.org/amp/blockchain-interoperability-2655912591.

85. F. F. Ozair, N. Jamshed, A. Sharma, and P. Aggarwal, "Ethical Issues in Electronic Health Records: A General Overview," *Perspectives in Clinical Research* 6, no. 2 (2015): 73–76, https://doi.org/10.4103/2229-3485.153997.



15320634, 2025, 12-14, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.70124 by Delhi Technological University, Wiley Online Library on [20/05/2025]. 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

2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT) | 979-8-3315-4335-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/DICCT64131.2025.10986474

# Faces in the Fog: A Deep Dive into Face De-identification Techniques and Their Comparative Analysis

Shaonti Goswami
*Department of CSE*
*Delhi Technological University*
*New Delhi, India*

Sudipty Konica Paul
*Department of CSE*
*Delhi Technological University*
*New Delhi, India*

Sanyam Jawlia
*Department of CSE*
*Delhi Technological University*
*New Delhi, India*

Anurag Goel
*Department of CSE*
*Delhi Technological University*
*New Delhi, India*

***Abstract*—Face de-identification have drawn a lot of attention due to growing privacy concerns and the growing requirement for anonymization in visual data .This review suggests a systematic exploration of current face de-identification methods, emphasizing three main deep learning-based methods: diffusion models, latent Variable based transformations, and GAN-based models. The in-depth analysis of this literature provides a comparison of how well each strategies protect privacy while retaining data utility by focusing on the advantages and disadvantages of them. Our analysis delves deeper into the benefits and limitations of these models, focusing on important issues such as adversarial robustness, systematic biases, and trade-offs between realistic image generation and privacy preservation. This comprehensive overview synthesizes data from several studies, identifies important gaps in existing approaches and provides guidance to researchers and industry practitioners directing to more reliable and ethical de-identification frameworks.**

***Index Terms*—Face de-identification, Generative deep learning ,Privacy bias and fairness, Identity protection.**

## I. Introduction

As artificial intelligence was rapidly advancing and following the extensive deployment of facial recognition systems, the concern regarding protection of personal identity became significantly larger. Governments, industries, and researchers have increasingly reached the awareness of dangers presented by an automated face, which can be used without the permission of others for unauthorized surveillance, theft of identity, or even the generation of deepfakes. Facial images are comprised of sensitive personal information, particularly in the realms of social media, surveillance video, healthcare records, and biometric authentication, which can result in severe violations of privacy if mishandled. The countermeasure against such risks is to develop face de-identification, an area of research that aims to develop algorithms that obscure personal identity effectively but maintain usability and realism of facial images.

From simple pixelation and blurring to more advanced deep learning-based methods which can, at best, produce realistic but anonymized face images, face de-identification techniques have come a long way [1]. The privacy preservation aspect has become more significant than ever due to rigorous data protection laws, such as the GDPR, and growing awareness of digital rights. Images have not been as effective with k-anonymity or differential privacy-like traditional anonymization techniques used with structured data as they would also degrade image utility while preserving anonymity [2], [3]. In an effort to combat this problem, deep learning models, particularly those based on GANs or latent space transformation techniques, were developed, thus showing some level of effectiveness with creating realistic and de-identified faces. However, as de-identification models become more advanced, new challenges arise, such as adversarial reversibility, fairness issues, and the trade-off between privacy and visual quality [2], [4], [5].

### A. Motivation

Face de-identification can generally be defined as the process of modifying or replacing identifiable face features without compromising the natural appearance of the image. Various methods have been proposed, from traditional image processing to cutting-edge deep learning. In this review, we discuss face de-identification methods based on a systematic categorization of strategies and their feasibility in maintaining privacy. To extensively research development within this area, we consider the following key research questions: Face de-identification can generally be defined as the process of modifying or replacing identifiable face features without compromising the natural appearance of the image. Various methods have been proposed, from traditional image processing to cutting-edge deep learning. In this review, we discuss face de-identification methods based on a systematic categorization of strategies and their feasibility in maintaining privacy. To extensively research development within this area, we consider few key research questions shown in Table I.

### B. Contributions

This review gives a systematic survey of face de-identification methods through a categorization into three major advanced methods: GAN-based methods, latent space manipulation methods, and diffusion methods for de-identification. These methods are described with a comparative categorization based on their underlying mechanism, algorithm type utilized, and their implications on privacy preservation. The review also emphasizes major contributions of various authors, such as the models utilized, their privacy preservation

979-8-3315-4335-8/25/$31.00 ©2025 IEEE



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:36 UTC from IEEE Xplore. Restrictions apply.

TABLE I
RESEARCH QUESTIONS & MOTIVATIONS

| RQ# | Research Question | Motivation |
|---|---|---|
| RQ1 | What is the current state-of-the-art in face de-identification? | To enumerate the latest de-identification methods in various contexts. |
| RQ2 | How can GAN-based de-identification models be optimized to prevent full identity exposure while maintaining facial expressions? | To investigate optimizations in GAN-based approaches for best privacy as well as visual realism. |
| RQ3 | How can latent space manipulation techniques be optimized to prevent reconstruction of sensitive features via latent space interpolation? | To investigate methodologies that are resistant to adversary attacks based on latent space loopholes. . |
| RQ4 | What new mechanisms of privacy protection can be integrated into diffusion-based face de-identification methods for enhanced reversibility attack robustness? | To investigate the diffusion model application towards highly secure and natural-looking de-identified faces. |

solutions, the utilized evaluation metrics, and the problem of privacy preservation.

The paper is structured as follows: The next section gives a comprehensive literature survey of the GAN-based, latent space-based, and diffusion-based de-identification methods. The pros, drawbacks, and privacy issues of each approach are discussed in comparative analysis that follow. Finally, the conclusion highlights the main obstacles to confront de-identification and suggests methods for further study to improve privacy protection, equity, and real-time application.

## II. DETAILED LITERATURE REVIEW

State-of-the-art efforts in facial image anonymization and de-identification have explored an array of generative models, including diffusion models, latent space approaches, and GANs, to reach a tradeoff between data utility and privacy assurance. Diffusion models have been favored to generate high-quality anonymized outputs, with dual-conditional models delivering greater anonymization control [6]. In order to ensure privacy, latent space techniques have been investigated to encode and modify face features while maintaining a concise description for safe data transfer [7]–[10]. With modifications centered on maximizing stability and minimizing artifacts, GANs have been in the vanguard of the synthesis of realistic yet anonymous face images . These strategies are complemented by research efforts in computational cost and scalability for practical application to real-world practice [11]. In aggregate, these studies testify to the universality of multiple generative models in addressing the problem of privacy, leaving stable solutions to the protection of facial data across many applications [12]–[14] as shown in Table II.

### A. *GAN Based Model for Face De-identification*

GAN-based models stand as essential tools for face de-identification because their generation of realistic facial images maintains both privacy standards. The several GAN based face de-identification models are summarized in Table III.

***GAN Variants and Techniques*** Conditional GANs (cGANs), StyleGAN and Landmark-Based GANs (L-GAN) that serve different purposes in de-identification tasks. The generation of de-identified faces with cGANs operates based on specific attributes like age or gender while StyleGAN uses latent space manipulation to achieve high-quality output generation . The disruption of identity recognition happens through methods which include latent space manipulation together with face swapping and adversarial perturbation techniques while maintaining non-identity features such as facial expressions and poses [2]. The integration of GANs with autoencoders or diffusion models has become popular for enhancing both de-identification quality and utility preservation. De-identification approaches found between privacy safeguards and perceptually natural face appearances make them applicable for security, emotional analysis and medical information exchange applications.

***Performance Evaluation and Output Effectiveness*** Assessment of the performance of GAN-based face de-identification techniques involves a trade-off between privacy protection, visual quality, and utility to downstream applications [15]. A number of key performance metrics are used.

**Privacy Metrics:** Privacy effectiveness is generally measured by comparison of face recognition similarity scores both prior to and following de-identification. As the similarity score gets lowers it indicates more effective privacy protection [1]. Furthermore , when advanced face recognition models (e.g., FaceNet, ArcFace) are used , adversarial evaluation are seen in order to estimate re-identification hazards .

**Image Quality Metrics:** To maintain visually realistic de-identified images, GAN-based models are evaluated using metrics such as Fréchet Inception Distance (FID), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR). High SSIM and low FID scores show that the generated face images are highly human-like while successfully de-identifying the person.

**Utility Preservation:** There are some methods which emphasize preserving usability for detecting emotion, gaze estimation, and video analysis. For instance, Facemotionpreserve performs facial de-identification in a way that medical diagnostic information is not biased, and so it beomes suitable for healthcare and biometric use cases. Additionally, feature-space adversarial perturbation-based de-identification is examined in terms of impact on machine learning models employed for human behavior analysis [12].

**User Perception and Human Factors:** A critical consideration of real-world deployments is whether perceptually natural de-identified images are perceived by human beings. Human perception is evaluated in experiments by Mean Opinion Score (MOS) surveys, where the subjects rate the realism of de-identified images . Additionally, in driver monitoring systems, GAN-based face de-identification is evaluated in nat-



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:36 UTC from IEEE Xplore. Restrictions apply.

TABLE II
FACE EDITING TECHNIQUES: FROM TRADITIONAL TO ADVANCED DL-BASED METHODS

| Approaches | Subcategory | Techniques |
|---|---|---|
| Traditional Methods | Pixelation | Minimizing image resolution to anonymize facial details. |
| | Blurring | Gaussian or motion blurring to conceal facial features |
| | Masking | Overlaying patterns or shapes to conceal facial features.. |
| | K-Same Algorithm | Substituting faces with average faces from a dataset to anonymize. |
| | Eigenface-Based Anonymization | Utilizing eigenfaces to anonymize faces in images. |
| Advanced DL-Based Methods | GAN-Based Models | StyleGAN: Style-based face generation for realistic face synthesis. |
| | | StarGAN: Multi-domain face modification using a single generator. |
| | | AttGAN: Attribute-based face editing (e.g., age, gender, hair color). |
| | Latent Variable Models | Latent Space Interpolation: Smooth face transitions. |
| | | Latent Space Projection: Images projected to latent space for editing. |
| | | Latent Space Manipulation: Individual attributes edited in latent space. |
| | Diffusion-Based Models | DDPM (Denoising Diffusion Probabilistic Models): Iterative denoising. |
| | | Latent Diffusion Models (LDM): Integrates diffusion with latent space. |
| | | Stable Diffusion: High-quality face development and editing. |

uralistic driving scenarios, hence not interfering with safety-critical use cases [1].

### *B. Latent Variable Model for Face de-identification*

Latent space-based face de-identification employs learned feature representations in generative models to anonymize facial identities without loss of significant attributes such as pose, expression, and facial structure. The overview is shown in Table IV.

**Latent Variants and Techniques:** Latent space mapping is one of the widely employed techniques, which manipulates identity-sensitive features without deforming the natural look of the face. This technique ensures the withholding of significant facial attributes at the same time preventing identity leakage [8]. Another strong technique is layer swapping, which determines that a limited number of layers in the latent representation are swapped to generate anonymization while retaining key facial details. By specifically modifying deep layers containing identity, this technique generates realistic but anonymized faces [3].

Apart from these transformation-based techniques, latent code optimization enhances the latent representation to perform optimal de-identification by minimizing identity similarity. This technique adjust the latent vector to reduce re-identification risks while maintaining facial attributes such as age and gender [2]. Similarly, identity-aware variational autoencoders (IA-VAE) implements a organized disentanglement of identity-related and identity-irrelevant features. By learning a disentangled enactment, this method supports efficient face-swapping while retaining natural appearance [16]. Another important technique is secure dual networks which employs reversible transformations in the latent space, while supports supervised retrieval of identity under authorized conditions [17]. These techniques altogether provide a trade-off between privacy maintainance and the retention of significant facial attributes.

**Performance Evaluation and Output Effectiveness** The evaluation metrics for Latent Variable based Face de-identification is estimated by :

**Privacy Metrics:** Face Verification Rate (FVR) and Cosine Similarity Score (CSS) evaluates privacy protection. When there is lower CSS value it reveals better anonymization, with HFORD offering exceptional privacy when CSS is below 0.3 [16], [18].

**Image Realism Measures:** PSNR, SSIM, and FID estimates image realism. SSIM above 0.85 yields natural faces, while below 0.6 causes distortion. FID less than 25 offers high realism, while over 40 implies excessive feature loss [17], [19].

**Identity Preservation and Usefulness:** L-GAN eliminates identity details but preserves expressions for applications like emotion detection. APLO preserves age and gender attributes with a success rate of over 90%, striking a balance between usability and privacy [2], [9].

**Bias and Fairness Considerations:** Equalized Odds Difference (EOD) and Demographic Parity (DP) facilitate minimization of bias. Models like People Taking Photos That Faces Never Share learn transformations in latent space for fair de-identification over groups [11].

### *C. Diffusion Based De-identification:*

Diffusion-based face de-identification methods apply generative diffusion models to iteratively optimize images for privacy maintenance while preserving crucial attributes.

**Diffusion Variants and Techniques:**The front-running variants are score-based generative models, denoising diffusion probabilistic models (DDPMs), and conditional latent diffusion models. IDDiffuse proposes a dual-conditional diffusion model with identity priors to control anonymization and attribute preservation [6]. Diff-Privacy leverages a noise-guided diffusion process that continuously deforms and reconstructs images to achieve irreversible masking of identities [21].

Latent diffusion models (LDMs) were found to be a prospective attribute-preserving de-identification option. LDMs act upon data within a latent compressed domain to optimize computation cost at realism preservation. Piano et al. demonstrate an LDM-based anonymization criterion with assured identity-controlled removal by protecting crucial facial attributes [10]. Campos et al. apply LDMs to obscure medical images with domain applicability demonstrations [7]. Other techniques, like Realistic Anonymization using Stable Diffusion (RAD), operates pre-trained stable diffusion models to introduce visually realistic anonymized images [22].



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:36 UTC from IEEE Xplore. Restrictions apply.

TABLE III
COMPREHENSIVE OVERVIEW OF MODELS AND VARIANTS

| Model Name | Variant Name | Authors | Technique Used | Mathematical Formulation | Performance Evaluation Metric | Strengths | Limitations |
|---|---|---|---|---|---|---|---|
| GAN-Based Models | CycleGAN | [13] | Optical face anonymization | $L_{GAN}(G,D) = \mathbb{E}[\log D(x)] + \mathbb{E}[\log(1 - D(G(z)))]$ | High-quality image translation with minimal artifacts in case of high PSNR and SSIM value | Vigorously handle adversarial attacks | Facial datasets are limited and so requires advanced hardware structures. |
| | Conditional GAN | [1] | Two-step face swapping framework | $L_{cGAN}(G,D) = \mathbb{E}[\log D(x,y)] + \mathbb{E}[\log(1 - D(G(z,y),y)]$ | High SSIM and FID ensure realistic face swaps with preserved details. | Robust to pose and lighting variations | Computationally expensive, limited to face datasets |
| | Wasserstein GAN | [5] | Anomaly detection in IoT networks | $L_{WGAN}(G,D) = \mathbb{E}[D(x)] - \mathbb{E}[D(G(z))]$ | High F1-Score ensures effective detection of rare anomalies. | Scalable and effective in detecting anomalies | High false-positive rate in noisy environments |
| | AttnGAN | [1], [14] | Text-to-image synthesis | $L_{AttnGAN}(G,D) = L_{GAN} + \lambda L_{KL}$ | High R-precision and FID ensure accurate alignment of text and images. | Aligns well with text descriptions | Struggles with complex textual inputs |
| | Temporal GAN | [5], [15] | Video frame interpolation | $L_{Temporal}(G,D) = \mathbb{E}[\log D(x_t, x_{t+1})] + \mathbb{E}[\log(1 - D(G(z)))]$ | High VMAF and PSNR ensure smooth frame interpolation with minimal artifacts. | Handles motion blur effectively | Struggles with fast-moving objects |
| | 3D-GAN | [14] | 3D object reconstruction from 2D images | $L_{3D-GAN}(G,D) = \mathbb{E}[\log D(x)] + \mathbb{E}[\log(1 - D(G(z)))]$ | Low Chamfer Distance and FID ensure accurate 3D reconstructions. | Robust to noisy inputs | Limited to small-scale objects |
| | Progressive GAN | [15], [12] | High-resolution medical image synthesis | $L_{Progressive}(G,D) = \sum_{i=1}^{n} L_{GAN}(G_i, D_i)$ | High Dice Coefficient and FID ensure high-quality medical image synthesis. | Enhances dataset diversity for medical AI models | Limited to specific medical imaging modalities |
| | SRGAN | [20] | Super-resolution for satellite images | $L_{SRGAN}(G,D) = L_{GAN} + \lambda L_{perceptual}$ | High PSNR and SSIM ensure enhanced image resolution for remote sensing. | Enhances image resolution for remote sensing | Requires high computational resources |
| | StyleGAN | [12], [8] | High-resolution image synthesis | $L_{StyleGAN}(G,D) = L_{GAN} + \lambda L_{style}$ | High IS and FID ensure high-quality image synthesis with fine details. | Captures fine details in images | Requires large computational resources |
| | AnoGAN | [20], [9] | Anomaly detection in industrial systems | $L_{AnoGAN}(G,D) = L_{GAN} + \lambda L_{anomaly}$ | High Precision and Recall ensure effective detection of rare anomalies. | Effective in detecting rare anomalies | High false-positive rate in noisy environments |

Denoising-based de-identification models play a role too in the process of eliminating identity. Zhu et al. propose diffusion autoencoders that advance video anonymization by promoting consistency in the frame while preserving details of motion and expression [23]. Threats to security are present nonetheless since Jiang et al. present the Denoising Diffusion Cryptanalytic Model (DDCM), an alternative that could be used for extracting anonymized images, leading to compromised privacy.

**Performance Evaluation and Output Effectiveness:** The metrics for diffusion based de-identification determines how well a model can anonymize faces while preserving essential visual attributes.

**Privacy Metrics:** Cosine Similarity Score (CSS) and Face Verification Rate (FVR) determines how well de-identification works. By reducing CSS below 0.3, IDDiffuse and Diff-Privacy perform strong anonymization and minimize the risk of re-identification [6], [21]. However, by re-generating de-identified pictures, adversarial recovery models such as DDCM reveal possible vulnerabilities and challenge robustness.

**Image Quality Metrics:** Visual fidelity is performed using the Fréchet Inception Distance (FID) and the Structural Similarity Index (SSIM). Natural face composition are preserved by Stable Diffusion-based models such as RAD, which maintain high SSIM (>0.85) [22]. Significant face distortions take place when SSIM falls below 0.6. When FID values is greater than 40 , it generates observable changes whereas values below 25 implies realistic results [10].

**Identity Retention and Attribute Preservation:** Diffusion autoencoders afflict a trade-off between anonymization and usefulness by retaining some characteristics, such as age and expression. While suppressing identity-relevant characteristics, LDM-based methods assure an attribute preservation rate of over 90%. [7], [10]. By avoiding frame-wise inconsistencies, video anonymization algorithms also preserve temporal coherence [23].

## III. CHALLENGES AND COMPARISON

Different approaches are used by face de-identification models to influence a trade-off between computing productivity, realism, and privacy. Each of the three methods—GAN-based, latent variable-based, and diffusion-based—has strengths and weakness with regard to robustness, picture quality, and identity removal. GAN based models are prompt and realistic yet prone to adversarial attacks. Latent variables prefer precise control over privacy protection but they often face difficulties in complete de-identification. Advanced privacy protection is provided by diffusion-based models, but algorithmic efficiency is sacrificed in the process. The particular application will determine which method is best, whether privacy, visual quality, or real-time processing are the top priorities. Table V depicts the advantages and limitations of various face de-identification approaches.

### A. *Advantages:*

*1) Advantages of GAN Based Model::* **GAN-based** approaches are preferable at producing incredibly precise de-identified faces at the same time maintaining crucial non-identifiable characteristics. By learning disentangled facial features, they allow attribute-preserving de-identification, which effectively protects privacy without considerably sacrificing visual quality. Though StyleGAN-based methods guarantee utility preservation, approaches such as RBGAN and L-GAN balance anonymity and authenticity [24]. Addition to this, real-time processing is supported by GAN-based solutions, which makes them perfect for streaming settings and video analytics [9], [15].

*2) Advantages of Latent Variable Based Model::* Variational autoencoders (VAEs) and latent diffusion models under **Latent space manipulation** methods provide precise benchmark over identity abolition while maintaining significant facial attributes . In spite of totally altering the original picture the methods, such as embedding-based de-identification



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:36 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
COMPREHENSIVE OVERVIEW OF MODELS AND VARIANTS

| Model Name | Variant Name | Authors | Technique Used | Mathematical Formulation | Performance Evaluation Metric | Strengths | Limitations |
|---|---|---|---|---|---|---|---|
| Latent Variable-Based Models | VAE-GAN | [8], [9] | Image generation with latent variable modeling | $L_{VAE-GAN} = L_{GAN} + \lambda L_{KL}$ | High FID and IS ensure high-quality image generation with latent space control. | Combines generative and latent space modeling | Struggles with complex datasets |
| | $\beta$-VAE | [9], [3] | Disentangled representation learning | $L_{-VAE} = \mathbb{E}[D(x)] - \beta\mathbb{E}[D(G(z))]$ | High disentanglement score ensures effective separation of latent factors. | Effective for unsupervised learning | Limited to small-scale datasets |
| | InfoGAN | [3] | Unsupervised disentangled representation learning | $L_{InfoGAN} = L_{GAN} + \lambda L_{MI}$ | High mutual information ensures effective disentanglement of latent variables. | Captures interpretable latent factors | Computationally expensive |
| Diffusion-Based Models | DDPM | [17] | Image generation via iterative denoising | $L_{DDPM} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\|\epsilon - \epsilon_\theta(x_t, t)\|\|^2\right]$ | High PSNR and SSIM ensure high-quality image generation. | Stable training and high-quality outputs | Computationally expensive |
| | Score-Based Generative Model | [17], [4] | Image generation via score matching | $L_{Score} = \mathbb{E}_{x\sim p(x)}\left[\|\|\nabla_x \log p(x) - s_\theta(x)\|\|^2\right]$ | High FID ensures realistic image generation. | Effective for high-resolution image synthesis | Requires large computational resources |

TABLE V
MODEL TYPES, ADVANTAGES, AND LIMITATIONS

| Model Type | Advantages | Limitations |
|---|---|---|
| **GAN-Based Models** | a. High-quality anonymization takes place. (e.g., L-GAN, GAN-based de-identification) [9]<br>b. The facial features are maintained at the time of anonymization (e.g., StyleGAN, L-GAN) [9]<br>c. Ethical preservation of identity-consistency (e.g., RBGAN, L-GAN) [1]<br>d. Capable for reversible anonymization (e.g., RBGAN, Diffuse) | a. Training and inference need high computational cost.(e.g., StyleGAN) [9]<br>b. Issues related to model interpretability. (e.g., L-GAN) [9]<br>c. Problems related to non-human faces and so limited to certain facial features.(e.g., RBGAN) [1]<br>d. Issues prevails related to relaistic image generation when facing obstruction. (e.g., GAN-based models) |
| **Latent Variable-Based Models** | a. High-quality image generation even with occlusion and performs robust face de-identification (e.g., HFORD, Identity-aware VAEs) [18]<br>b. Better maintenance of privacy preservation. (e.g., RBGAN, Face anonymization via latent space mapping)<br>c. Reversible facial features are allowed in some methods (e.g., HFORD) [16]<br>d. Facial attributes are protected during anonymization (e.g., Latent Space Manipulation) [8] | a. Generalization might not be well across all datasets (e.g., latent space manipulation) [8]<br>b. Rigorous computation is seen. (e.g., VAE-based models) [16]<br>c. In order to train effectively large datasets are required. [8] |
| **Diffusion-Based Models** | a. Well organized for privacy-preserving face de-identification. (e.g., DDCM, IDDiffuse)<br>b. Denoising and perturbation maintains improved identity preservation. (e.g., Diff-privacy) [21]<br>c. Re-identification might not take place . (e.g., Diffusion-based Face Privacy Protection) [6] | a. Resource consumption is potentially high.(e.g., large diffusion models)<br>b. Natural images geberation might be less compared to GAN-based models (e.g., Diffusion-based methods) [21]<br>c. Training phase required large labeled datasets. [10] |

and latent space optimization enable structural changes that protect privacy . In sensitive applications such as medical imaging, secure dual-network and attribute-preserving latent space models enhance usability and reversibility [2], [17]. Furthermore, in comparison to direct pixel-space alterations, latent space approaches provide more advanced generalization across a variety of datasets and applications [8].

*3) Advantages of Diffusion Based Model::* **Diffusion models** provide anonymized faces with exceptional persistent quality and privacy guarantees by utilizing probabilistic changes. In terms of preserving identity reconstruction attempts while retaining natural look, they perform far better than conventional GANs. Identity-consistent anonymization and enriched security in privacy-sensitive areas are offered by techniques like IDDiffuse and Diff-Privacy [6], [21]. Latent diffusion models are very favorable at anonymizing medical and forensic datasets because they further improve attribute preservation [7], [10]. Diffusion models also provide controlled de-identification at diverse noise levels, assuring flexibility in a range of real-world situations.

### *B. Limitations:*

Although **GAN-based face de-identification** techniques have their benefits, they are prone to confronting attacks, identity leakage, and mode collapse [15], [24]. They also struggle to create a balance between privacy and functionality, which subsequently results in severe modifications that produce unrealistic images. [9] **Latent variable-based** approaches might result in undesirable biases since they conflict ensuring strong identity removal in spite of being functional for preserving facial features [2], [17]. Because of advanced training and learning processes, this model need a lot of computing resources [8]. Even though **diffusion-based methods** are more secure and of higher quality, they are computationally costly and have extensive inference times [6], [21]. Also, their clarity and control over de-identification levels are bounded which are making them incompatible for real-time applications [7], [10]. Limitations of these various techniques are shown in Table V.

## IV. FUTURE RESEARCH DIRECTIONS:

Context-aware anonymization is a likely-looking approach in which the model dynamically adjusts the intensity of de-identification in response to risk estimation. Reinforcement learning or uncertainty-aware models can be used for this that measure the necessary degree of anonymization while maintaining image functionality . Since many current methods focus on combining sensitive face data for training, robust models may be trained without collecting user data by incor-



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:36 UTC from IEEE Xplore. Restrictions apply.

porating federated learning (FL) with self-supervised learning (SSL). FL ensures that training takes place locally on user devices, and SSL can improve feature extraction without labeled data, reducing the possibility of privacy risks. Face de-identification often rely only on visual data, but real-world privacy exposition also come from textual and biometric derivations. Another future direction can be developing multi-modal anonymization frameworks that anonymize facial images, textual descriptions, and biometric markers (e.g., gait patterns, voice data) together. As a result, privacy protection will be ensure in applications like social media, surveillance, and medical imaging.

## V. Conclusion

This study examines the evolution in face de-identification methods, evaluating their efficacy, drawbacks, and implementations for privacy protection. This review's aim is to provide readers a comprehensive overview of the development of de-identification techniques, from conventional obfuscation techniques like blurring and pixelation to more complex deep learning approaches like GANs and diffusion models . By looking at different methods, this survey shows how these techniques may handle issues like identity exposure, fairness, and adverse weaknesses while maintaining a balance between privacy and data storage [15]. The paper also critically estimates how self-supervised learning and federated learning might improve privacy without disclosing sensitive data. With the use of this study, the review hopes to clarify on the advantages and disadvantages of current methods, offering a structured comparison that helps researchers and and practitioners in deciding on appropriate de-identification methods for real-world applications.

## References


[1] S. Thapa and A. Sarkar, "Gan-based deidentification of drivers' face videos: An assessment of human factors implications in nds data," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–6, 2023.

[2] S. Barattin, C. Tzelepis, I. Patras, and N. Sebe, "Attribute-preserving face dataset anonymization via latent code optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8001–8010, 2023.

[3] E. BenSaid, M. Neji, and A. M. Alimi, "Natural face anonymization via latent space layers swapping," in *2023 IEEE Symposium on Computers and Communications (ISCC)*, pp. 562–567, IEEE, 2023.

[4] G. Hanawa, K. Ito, and T. Aoki, "Face image de-identification based on feature embedding," *EURASIP Journal on Image and Video Processing*, vol. 2024, no. 1, p. 25, 2024.

[5] R. More, A. Maity, G. Kambli, and S. Ambadekar, "Privacy-preserving video analytics through gan-based face de-identification," in *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*, pp. 1–6, 2024.

[6] M. Shaheryar, J. T. Lee, and S. K. Jung, "Iddiffuse: Dual-conditional diffusion model for enhanced facial image anonymization," in *Proceedings of the Asian Conference on Computer Vision*, pp. 4017–4033, 2024.

[7] F. Campos, L. Petrychenko, L. F. Teixeira, and W. Silva, "Latent diffusion models for privacy-preserving medical case-based explanations," *submitted to EXPLIMED, ECAI*, p. 3, 2024.

[8] E. BenSaid, M. Neji, and A. M. Alimi, "Faceanonym: Face anonymization model via latent space mapping," in *2023 International Conference on Cyberworlds (CW)*, pp. 480–483, IEEE, 2023.

[9] S.-s. Jang, C.-j. Kim, S.-y. Hwang, M.-j. Lee, and Y.-g. Ha, "L-gan: landmark-based generative adversarial network for efficient face de-identification," *The Journal of Supercomputing*, vol. 79, no. 7, pp. 7132–7159, 2023.

[10] L. Piano, P. Basci, F. Lamberti, and L. Morra, "Latent diffusion models for attribute-preserving image anonymization," *arXiv preprint arXiv:2403.14790*, 2024.

[11] J. Zhu, L. Gu, X. Wu, Z. Li, T. Harada, and Y. Zhu, "People taking photos that faces never share: Privacy protection and fairness enhancement from camera to user," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 14646–14654, 2023.

[12] H. Xue, B. Liu, X. Yuan, M. Ding, and T. Zhu, "Face image de-identification by feature space adversarial perturbation," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 5, p. e7554, 2023.

[13] J. Lopez, C. Hinojosa, H. Arguello, and B. Ghanem, "Privacy-preserving optics for enhancing protection in face de-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12120–12129, June 2024.

[14] J. Song, J. Kim, and J. Nang, "Face de-identification using convolutional neural network (cnn) models for visual-copy detection," *Applied Sciences*, vol. 14, no. 5, p. 1771, 2024.

[15] Y. Zhang, Y. Fang, Y. Cao, and J. Wu, "Rbgan: Realistic-generation and balanced-utility gan for face de-identification," *Image and Vision Computing*, vol. 141, p. 104868, 2024.

[16] Z. Li, Z. Zhang, S. He, Q. Meng, S. Zhang, B. Zhong, and R. Ji, "Identity-aware variational autoencoder for face swapping," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5466–5479, 2024.

[17] Y.-L. Pan, J.-C. Chen, and J.-L. Wu, "Secure dual network for reversible facial image anonymization through the latent space manipulation.," *Electronics (2079-9292)*, vol. 13, no. 22, 2024.

[18] D. Chen, M. Zhu, N. Wang, and X. Gao, "Hford: High-fidelity and occlusion-robust de-identification for face privacy protection," *arXiv preprint arXiv:2311.08786*, 2023.

[19] Y. Cao, Y. Zhang, J. Wu, and Y. Fang, "Multi-channel attribute preservation for face de-identification," *Multimedia Tools and Applications*, pp. 1–25, 2024.

[20] Y. Wen, B. Liu, L. Song, J. Cao, and R. Xie, "High quality face de-identification with model explainability," in *Face De-identification: Safeguarding Identities in the Digital Era*, pp. 127–146, Springer, 2024.

[21] X. He, M. Zhu, D. Chen, N. Wang, and X. Gao, "Diff-privacy: Diffusion-based face privacy protection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[22] S. Malm, V. Rönnbäck, A. Håkansson, M.-h. Le, K. Wojtulewicz, and N. Carlsson, "Rad: Realistic anonymization of images using stable diffusion," in *Proceedings of the 23rd Workshop on Privacy in the Electronic Society*, pp. 193–211, 2023.

[23] Y. Zhu, J. Cao, B. Liu, T. Chen, R. Xie, and L. Song, "Identity-consistent video de-identification via diffusion autoencoders," in *2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, IEEE, 2024.

[24] W. Khan, L. Topham, U. Khayam, S. Ortega-Martorell, P. Heather, D. Ansell, D. Al-Jumeily, and A. Hussain, "Person de-identification: A comprehensive review of methods, datasets, applications, and ethical aspects along-with new dimensions," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:36 UTC from IEEE Xplore. Restrictions apply.

Check for updates

ChemistrySelect

Research Article
doi.org/10.1002/slct.202500293

Chemistry Europe
European Chemical Societies Publishing

www.chemistryselect.org

# FeO NPs-Loaded Guar Gum/Xanthan Gum-Based Hydrogel: A Sustainable Approach for Cationic Dye Removal

Priyanka Yadav,[a] Kanishka Kashyap,[a] Sudhir G. Warkar,[a] and Anil Kumar*[a]

In this research article, facile synthesis of iron oxide nanoparticles (FeO NPs) was done, and novel FeO NPs loaded guar gum (GG): xanthan gum (XG)-based hydrogel has been fabricated to treat wastewater via the adsorption process. The structural and morphological properties of the optimized sample were characterized and analyzed by FTIR, PXRD, TGA, DLS, HRTEM, and SEM-EDAX. The swelling ratio was found to be maximum (6201%) in the case of Fe3:GG:XG hydrogel nanocomposite. The gel content (%) of the hydrogel nanocomposite was found to be 86. The maximum adsorption capacity ($q_{max}$) was calculated using the Langmuir isotherm and found to be 130 mg/g. In addition, the result revealed that the iron-loaded hydrogel has better adsorption capacity, such as ~97% removal of dye with 0.35 g in 90 min of contact time. Moreover, the isotherms and kinetics were fitted, and it was found that the synthesized hydrogel is best fitted in the Langmuir isotherm with pseudo-second-order kinetics, with the regression coefficients 0.988 and 0.998, respectively.

## 1. Introduction

Water is crucial to our day-to-day existence on Earth, and a constant need to preserve and improve its quality is necessary for mankind and other living beings. According to a survey, it was found that in the forthcoming decades, the only alternatives for procuring fresh water will be water treatment and recycling techniques.[1] Various hazardous materials are discharged into water bodies, including textiles, industrial waste, and so on. Still, synthetic dyes, specifically cationic dyes, affect water quality the most, leading to a growing and significant global hazard to the ecosystem. Synthetic organic dyes have harmful consequences for all life in the ecosystem due to nonbiodegradability, toxicity, allergic problems, skin irritations, and carcinogenic effects. Besides these significant concerns, dyes are utilized worldwide in many sectors, including plastics, textiles, leather, paper, printing, and pharmaceuticals, and are substantially sequestrated into water bodies as waste products.[2,3]

To overcome the concern of water quality, dye treatment techniques that are simpler, more precise, and cost-effective need to be developed. Several methods are available for dye sequestration, such as coagulation, flocculation, membrane filtration, photocatalysis, and adsorption.[4] Adsorption is one of the best techniques as it is facile, easy to use, and cost-effective.[5]

Nowadays, "Hydrogels" are an emerging class of materials that can be used as efficient adsorbents. It is defined as a class of polymeric materials consisting of biopolymer as well as a synthetic polymer that can hold plenty of water in their three-dimensional networks due to the availability of hydrophilic moiety in its structure such as carboxylic, amide, amino groups, sulfonic acid, and hydroxyl.[6] Bio-polymeric-based super-adsorbent hydrogels are also utilized explicitly in wastewater remediation through adsorption mechanisms such as ion exchange, physisorption, chemisorption, and interactions of targeted dye molecules and adsorbents due to their high adsorption capabilities, low toxicity, and regeneration capacities.[7] Numerous studies have been going on hydrogel-based adsorbents to remove pollutants from wastewater.[8] Huang and coworkers developed a dye adsorbent made of graphene oxide (GO) modified zeolite to sequestrate methylene blue from wastewater.[9] Gau and group fabricated a hydrogel and utilized it in the removal of dyes.[10]

Biopolymer-Xanthan gum (XG)-based hydrogels have sharply risen in various fields due to their immense applications in various fields.[11] XG is an exopolysaccharide that was identified in the 1950s. It is primarily derived from strain X of the family Xanthomonas, a plant-pathogenic microorganism.[12] Kumar and the group have utilized Xanthan gum-based hydrogel in the sequestration of dye.[13]

To lower the production expenses and enhancement properties like bioavailability and water retention capacity, Guar gum (GG) and Locust bean (LB) gum are often combined with XG.[14] Guar gum (GG), also known as cymopsis tetragonolobus, is a significant agrochemical active produced from the guar plant's seed endosperm. It is a novel thickener and stabilizer because of its strong propensity to create hydrogen bonds in water.[15,16] Further, synthetic monomers such as acrylic moieties can be used to improve the properties of existing moieties due to hydrophilic groups to enhance the strength of biopolymers.[17] Table 1 represents some reported literature based on guar gum (GG) and xanthan gum (XG)-based hydrogel. However, no reports are available on the exact combination that is utilized in this article.

In order to further enhance the properties of the hydrogel, the focus has been shifted toward the reinforcement of nanoparticles into the hydrogel to fabricate the hydrogel

[a] P. Yadav, K. Kashyap, S. G. Warkar, A. Kumar
Department of Applied Chemistry, Delhi Technological University (DTU), Delhi 110042, India
E-mail: anil_kumar@dce.ac.in

Supporting information for this article is available on the WWW under https://doi.org/10.1002/slct.202500293



© 2025 Wiley-VCH GmbH

**Table 1.** Comparative analysis of different hydrogels based on guar gum and xanthan gum.

| S. No. | Hydrogel Based on Guar Gum and Xanthan Gum (XG) | Swelling Ratio (%) | Adsorption Capacity/% Removal | Ref. |
|---|---|---|---|---|
| 1. | GG/XG | 2213.20 | – | [18] |
| 2. | XG//PAM/Alginic acid | 96 | 114.9/84 | [19] |
| 3. | GG/*Ulva lactuca L.*/XG | – | 62.83 / 92 | [20] |
| 4. | GG/polyacrylic acid-aniline | – | 80/66 | [21] |
| 5. | GG/PAM | – | 31/90 | [22] |

**Table 2.** Different adsorption materials utilized in adsorption.

| S.No. | Adsorption Materials | Concentration of Dye | Adsorption Capacity | Ref. |
|---|---|---|---|---|
| 1. | PVC/CNC/$TiO_2$ | 30/0.1 | 14 | [27] |
| 2. | MgO-bentonite clay | 120or 80/1g | 17.2 | [28] |
| 3. | Orange peel-N-vinyl-2pyrrolidone | 20/50 | 26 | [29] |
| 4. | CMTKG/PAM/GO | 40 | 100 | [30] |
| 5. | Zeolite-CMTKG-PMA | 20/0.1 | 126 | [31] |

nanocomposite, which results in enhanced properties. The nanocomposite-based hydrogel has great potential to increase the effectiveness of the hydrogels, such as improved medication administration in drug delivery applications and improved degradation accuracy with lower undesirable or toxic side effects. In the case of adsorption, improved dispersion, better stability, and better adsorption efficacy were observed. A variety of inorganic/organic nanomaterial-based systems have been reported, such as graphene and carbon nanotubes, metal chalcogenides, iron oxide nanoparticles (FeO NPs), activated charcoal, biochar, clay, and silica are just a few examples that can be used to treat wastewater.[23] Synthesis of FeO NPs has been extensively researched and developed for various biological and technological applications, including the extraction of genomic DNA and medicine, such as targeted drug delivery and the energy field.[24,25] Among other transition metals, FeO NPs have the most vital magnetism and are also stable at ambient temperatures. Due to the effectiveness of FeO NPs, it is incorporated into the hydrogel matrix and used in wastewater treatment. Afkhami and group have synthesized magnetic nickel zinc ferrite nanocomposite as an efficient, effective electrostatic interaction that primarily promotes dye adsorption. FeO NPs are widely used in the synthesis of hydrogel nanocomposites due to their minimal toxicity and bioenvironmental activities. Lately, FeO NPs have garnered a lot of interest from researchers due to their various advantages, such as surface-to-volume ratio and larger surface area. FeO NPs have opposite charges to the hydrogel matrix, susceptible to pH and ionic strength changes.[26] The incorporation of nanoparticles into the hydrogel matrix creates a hydrogel nanocomposite, which has improved water treatment performance and made it possible to increase their mechanical strength, swelling characteristics, and ability to reuse.[6] Table 2 represents some reported data in which different adsorbent materials have been utilized in adsorption.

To the best of our knowledge, a combination of Fe:GG:XG:PAM is not known yet and has not been explored to check its potential in the application of drug delivery. Therefore, the components utilized in the fabrication of Fe:GG:XG:PAM may serve as a potent carrier to adsorb the pollutant. This study includes the synthesis of FeO NPs using a coprecipitation approach and a novel low-cost sorbent material developed by incorporating FeO NPs into a hydrogel matrix using affordable, environmentally friendly, naturally occurring XG and GG. Polyacrylic acid (PAA) was also utilized to strengthen the hydrogel, and the synthesized hydrogel nanocomposite was utilized in a rapid and highly effective approach for sequestration of crystal violet (CV).

## 2. Materials and Methods

### 2.1. Materials

Xanthan gum, acrylic acid (AA), *N,N′*-methylene bis(acrylamide), and sodium hydroxide were obtained from CDH, New Delhi. Guar gum was kindly provided by Hindustan Gum and Private Limited, India. Crystal violet (CV), potassium persulfate (KPS), iron (II) chloride tetrahydrate ($FeCl_2.4H_2O$), and anhydrous iron (III) chloride ($FeCl_3$) were purchased from Merck, Germany. All the studies were performed in double-distilled water. The error bars were inserted on data points using standard deviation.

### 2.2. Synthesis

#### *2.2.1. Synthesis of Iron Oxide Nanoparticles (FeO NPs)*

The FeO NPs synthesis was done using the coprecipitation method.[32] For the synthesis, 0.301 M $FeCl_2$ and 0.665 M $FeCl_3$ were suspended in distilled water and stirred for 30 min at 95 °C. Then 1 M NaOH was added dropwise to maintain solution pH 10, followed by the addition of ammonia (10 mL), and an instant color change from yellow to black was observed. Then, the solution was stirred for 2 h at 95 °C, resulting in the formation of a black precipitate. The above mixture was then centrifuged at 6000 rpm for 10 min, and the excess solution was removed to separate the precipitate and washed with ethanol. Then, the obtained product was placed in the oven. Further, the product was calcined at 300 °C in oxygen atmosphere for 3 h to get the final product.

#### *2.2.2. Synthesis of FeO-XG/GG/PSA Hydrogel Nanocomposite*

The optimized amount of Xanthan gum (0.16 g) and Guar gum (0.16 g) was dispersed in 40 mL of distilled water as the preliminary study presented in Table S1. The mixture was agitated until a homogenous dispersion was obtained, followed by the addition of acrylic acid (5 mL) and NaOH (1 g) into the mixture. Then, initiator KPS (0.05 g) and crosslinker MBA (0.05 g) were


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**Table 3.** Composition of Constituents in the formulation of hydrogels.

| Sample Code | XG (g) | GG (g) | MBA (g) | KPS (g) | FeO NPs (g) |
|---|---|---|---|---|---|
| GG:XG | 0.16 | 0.16 | 0.05 | 0.05 | 0 |
| Fe2:GG:XG | 0.16 | 0.16 | 0.05 | 0.05 | 0.02 |
| Fe3:GG:XG | 0.16 | 0.16 | 0.05 | 0.05 | 0.03 |
| Fe4:GG:XG | 0.16 | 0.16 | 0.05 | 0.05 | 0.04 |
| Fe5:GG:XG | 0.16 | 0.16 | 0.05 | 0.05 | 0.05 |

added and stirred for 30 min. Further, FeO NPs were added dropwise by varying amounts (0, 20, 30, 40, and 50 mg), and stirring was maintained for 40 min. Finally, the homogeneous mixture was poured into the test tubes, and labeled GG:XG, Fe2:GG:XG, Fe3:GG:XG, Fe3:GG:XG, and Fe3:GG:XG were kept in the water bath at 75 °C for 2 h. Then, the obtained products were cut into pieces, submerged in distilled water overnight followed by oven drying at ambient temperature, compositions are as represented in Table 3.

## 3. Characterization

The investigation of synthesized materials was done using various techniques. A Perkin Elmer Fourier-transform infrared spectrometer (FTIR) was used to analyze the synthesized materials in a range of 400–4000 $cm^{-1}$; a high-resolution X-ray diffractometer (PXRD), model Bruker D8 advanced, having angle $2\theta = 5 - 70°$ and $\lambda = 1.5418$ Å, was used for the analysis of samples. Scanning electron microscopy (SEM) was used to identify the surface morphology of the synthesized material using JEOL, JSM 6610LV. Further, to assess the thermal stability of synthesized materials, TGA analysis was done in a nitrogen ($N_2$) atmosphere with a temperature ranging from 30 to 900 °C at a uniform heating rate. HRTEM imaging of FeO NPs was performed using TECNAI G 20 (200 KV). Size distribution measurement of FeO NPs was carried out in distilled water using a dynamic light scattering (DLS), (Model ZSP, Malvern instrument, UK). The UV–vis spectrophotometer model UV-1800 was used to record the absorption spectra.

### 3.1. Swelling Study

The well-known gravimetric method was used to evaluate GG:XG, Fe2:GG:XG, Fe3:GG:XG, Fe3:GG:XG, and Fe3:GG:XG. The hydrogels were weighed and placed in 100 mL of deionized water for 24 h. Then, hydrogels were removed after a regular interval of time, excess water was wiped off from the surface using filter paper, and hydrogels were re-weighed.[33] The same experiment was done in triplicate to avoid errors. The percentage of swelling ratio (%SR) was calculated using Equation (1).

$$\text{Swelling ratio}\,(\%) = \frac{w_t - w_i}{w_i} \times 100 \quad (1)$$

$W_t$ is the weight of swelling hydrogels at time $t$, and $W_i$ is the initial weight of dried hydrogels.

### 3.2. Sol–Gel Content and Porosity Analysis

The analysis was performed to calculate the uncrosslinked part of the fabricated FeO NP-loaded hydrogels. Sol implies the soluble uncrosslinked portion within the hydrogel, while a gel is the insoluble cross-linked part of the synthesized formulation of the hydrogels. To assess this Soxhlet extraction method was used. Initially, the dried hydrogel discs were weighed and subjected to Soxhlet and distilled water as solvent at 100 °C for 2 h. After 2 h, the extracted hydrogel discs were oven-dried at 45 °C, and their weight was measured. The sol and gel fraction calculations were performed using the following equations.[34]

$$\text{Sol Fraction} = \frac{w_i - w_f}{w_i} \times 100 \quad (2)$$

$$\text{Gel fraction} = 100 - \text{Sol fraction} \quad (3)$$

where $W_i$ implies the weight of the initially dried hydrogel, and $W_f$ implies the final weight of the extracted dried hydrogel after being subjected to the Soxhlet extraction method.

The liquid displacement technique was employed to calculate the hydrogel's porosity. To perform this, a fixed amount of hydrogel ($W_1$) was submerged in ethanol for 24 h; after 24 h, the samples were taken out, and the surface liquid was removed using tissue paper, and then weighed ($W_2$). The hydrogels' porosity was calculated using Equation (4).[35]

$$\text{Porosity}\,(\%) = \frac{w_1 - w_2}{\rho v} \times 100 \quad (4)$$

where $\rho$ denotes the density of ethanol, and $V$ represents volume ($V = \pi r^2 h$ of hydrogel disc; here, $r$ is the radius of the disc, and $h$ is the height of the disc. The vernier caliper was used for the measurements).

### 3.3. Adsorption Study

The crystal violet (CV) removal was accomplished by attempting a variety of parameters, including contact time, concentration study, adsorbent dosage, and pH variation. Isotherm and kinetic studies were also performed to validate the experimental data. A 100 mg/L CV solution was used to study the above-mentioned parameters. For contact time, 0.035 g adsorbent was taken in 100 mL of dye solution (40 mg/L) and kept in the dark condition for 90 min; after a regular interval of 10 min, 2 mL of aliquots were withdrawn, and UV–vis spectra were recorded at $\lambda_{max}$ 590 nm. In the concentration study, concentration was varied from 10 to 60 mg/L, kept in stirring for 90 min, and after a regular time interval, the sample was aliquoted, and $\lambda_{max}$ was recorded at 590 nm using a UV–vis spectrophotometer. Adsorbent dosage was studied by varying amounts from 0.015–0.06 g; the same experiments were followed as above, and absorbance was recorded. To examine the effect of pH, experiments were conducted using fixed concentrations of dye solutions (40 mg/L) and adsorbent dose (0.035 g) for 90 min at various pH ranges from solution pH 3–9. The pH study was done by varying from 3


© 2025 Wiley-VCH GmbH

to 9, and the pH of the solution was maintained by adding 0.1 N HCl and 0.1 N NaOH. A similar experiment was performed using a UV–vis spectrophotometer. A UV–vis spectrophotometer (UV-1800) was used to calculate the sorbent's removal effectiveness (% R) and maximum adsorption capacity ($q_e$) for CV by following Equations (5) and (6):

$$\% R = \frac{C_i - C_f}{C_i} \times 100 \quad (5)$$

$$q_e = (C_i - C_f) \times \left(\frac{V}{W}\right) \quad (6)$$

where $C_i$ and $C_f$ (mg/L) are the initial and final concentration of CV dye, $V$ is the volume of dye solution (mL), and $W$ (g) is the mass of sorbent.

Further, isotherms were plotted, and kinetics studies were accomplished by contact time using 0.035 g adsorbent having 90 min of reaction time.[36]

### 3.4. pH of Electric Point Zero ($pH_{ZC}$)

The point where the surface of a substance turns neutral is known as the point zero charge ($pH_{ZC}$). A material's surface displays a positive charge when its pH is below $pH_{ZPC}$ and a negative charge when it is above $pH_{ZPC}$. It was determined using an electrolytic aqueous solution ranging in pH from 2 to 11. After adjusting the pH of the solution, 0.035 g adsorbent was added to the pH solution and then kept stable for 48 h. Then, the final pH was measured, a plot was made between ΔpH versus pH, and the point where the plotted line intersects the x-axis refers to $pH_{zc}$.[37]

### 3.5. Desorption and Reusability

The desorption and reusability of hydrogels loaded with FeO NPs were investigated over five consecutive cycles. To perform this study, 0.035 g Fe3:GG:XG was taken into 100 mL dye solution (40 mg/L) in a beaker and placed in dark condition at room temperature. After a regular interval, a 2 mL sample solution was taken, and absorbance at $\lambda_{max}$ 590 nm was recorded. Then, the desorption was carried out using 0.1 N HCl and 0.1 N NaOH solutions. The adsorbent was stirred with a 100 mL mixture of 0.1 N HCl and 0.1 N NaOH for 40 min, followed by washing with distilled water. Then, the obtained sorbent was dried in an oven to get a persistent weight, and the adsorbent was weighed for the next cycle using the similar adsorption process as above.

## 4. Results and Discussion

### 4.1. Mechanism-Synthesized Hydrogel

The loading of FeO NPs was done to formulate different hydrogels, as shown in Table 3. The synthesized hydrogel was effectively created by free radical polymerization. The mechanism involved biopolymer-XG and GG, polymer-poly (sodium acrylate) (PSA), initiator-KPS, and crosslinker-MBA. Initially, the redox initiator KPS breaks down to produce the active sulfate radicals at 60 °C that promote polymerization by removing hydrogen from the unsaturated double bond of SA. Then, the vinyl bond of the sodium acrylate was broken down by an anionic radical created by KPS, producing sodium acrylate radicals. Then, the generated radical attacks another sodium acrylate monomer by the sodium acrylate radical, resulting in chain propagation. Further, the vinylic group of PSA, the -OH group of XG, and GG are also cross-linked by the MBA radical, which serves as a bifunctional crosslinking agent. The loaded FeO NPs show electrostatic interaction with the hydrogel matrix, which results in fabricating the FeO NPs-loaded XG/GG/PSA hydrogel, as shown in Scheme 1.

### 4.2. Characterization

#### *4.2.1. Swelling Behavior*

The swelling behavior of synthesized hydrogel was examined in double-distilled water, as presented in Figure 1a. It was observed that the swelling ratio (%SR) was increased from GG:XG to Fe3:GG:XG hydrogels and decreased in Fe4:GG:XG and Fe5:GG:XG hydrogels, respectively. The %SR was found to be maximum in Fe3:GG:XG hydrogel, suggesting the most robust possible synergistic interactions between the FeO NPs and the polymers in the hydrogel matrix. Meanwhile, the reduction in swelling ratio for Fe4:GG:XG and Fe5:GG:XG hydrogels was noted because there were more FeO NP particles present in the lattice, which significantly decreased the synergistic interactions between the FeO NPs and the hydrogel matrix. Furthermore, the quantity of FeO NPs rose along with the stiffness of the hydrogels, which is also responsible for the hydrogel's decreased swelling. Hence, it can be concluded that the optimized amount for adding filler in the matrix is 0.03 g because Fe3:GG:XG hydrogels showed the highest %SR, and Fe3:GG:XG hydrogels were employed for further studies.

#### *4.2.2. Sol–Gel Content*

Figure 1b represents the sol–gel analysis of synthesized FeO NPs loaded XG/GG/PSA hydrogels. A regular increase in the trend of gel fraction was observed with increases in the amount of FeO NPs (0, 20, and 30 mg) in (GG:XG, Fe2:GG:XG, and Fe3:GG:XG). In addition, when FeO NPs (40 and 50 mg) are further increased (Fe4:GG:XG and Fe5:GG:XG), resulting in a higher sol fraction and a lesser gel content. The possible cause of this is the significant interaction of functional groups present in FeO NPs and the three-dimensional network of the hydrogel matrix, which leads to a decrease in gel content.[38] In a study, Saleh and group reported that the amount of $CeO_2$-NPs gel content decreased with an increase in the amount of $CeO_2$-NPs.

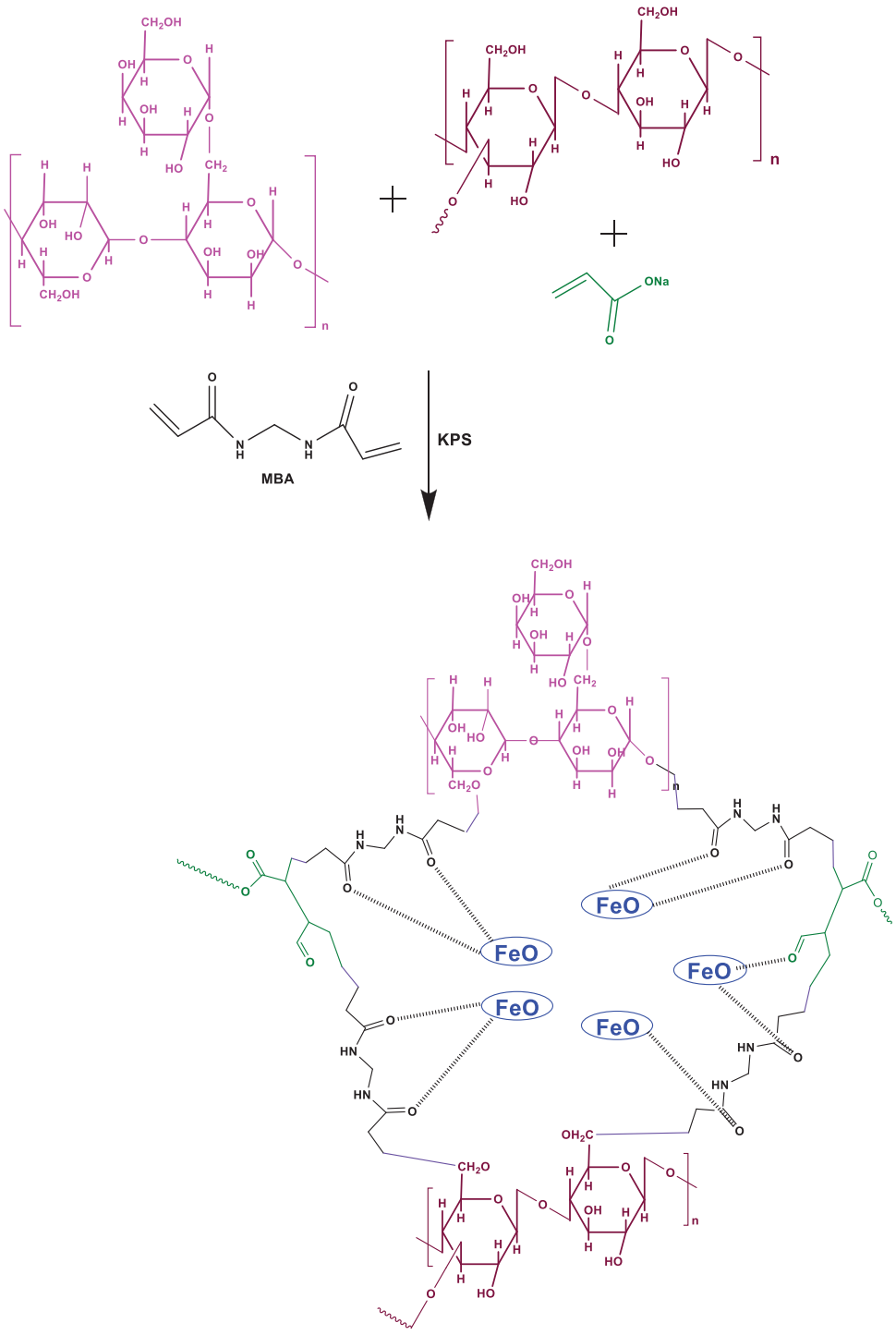
It was found that as the amount of FeO NPs in the synthesized hydrogel is introduced, the porosity of the system also


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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 Article
doi.org/10.1002/slct.202500293

Chemistry Europe
European Chemical Societies Publishing




**Scheme 1.** Mechanism of the synthesized hydrogel nanocomposite.

increases. As seen in Figure 1c, the observed porosity ranges from 68% to 80%. The incorporation of FeO NPs results in the enhancement of the porosity due to the hydrogel's ability to expand the network in the hydrogel matrix. FeO NPs are introduced as a filler into the hydrogel matrix. As per available literature, it was found that less amount of nanoparticles in the hydrogel matrix results in the expansion of the hydrogel network structure, which leads to high porosity. While the amount of filler increases, the structure of the matrix becomes dense and comparatively tightly packed, leading to the tightening of the polymer network.[39] Hence, it can be concluded that as the concentration of FeO NPs increases from 0 to 30 mg, the porosity increases, which may be attributed to the expansion of the hydrogel network. As the concentration increased from 30 mg to 50 mg, a reduction in porosity was observed, which may be due to the dense structure of the hydrogel. Hence, the porosity of Fe3:XG:GG was found to be maximum.



© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

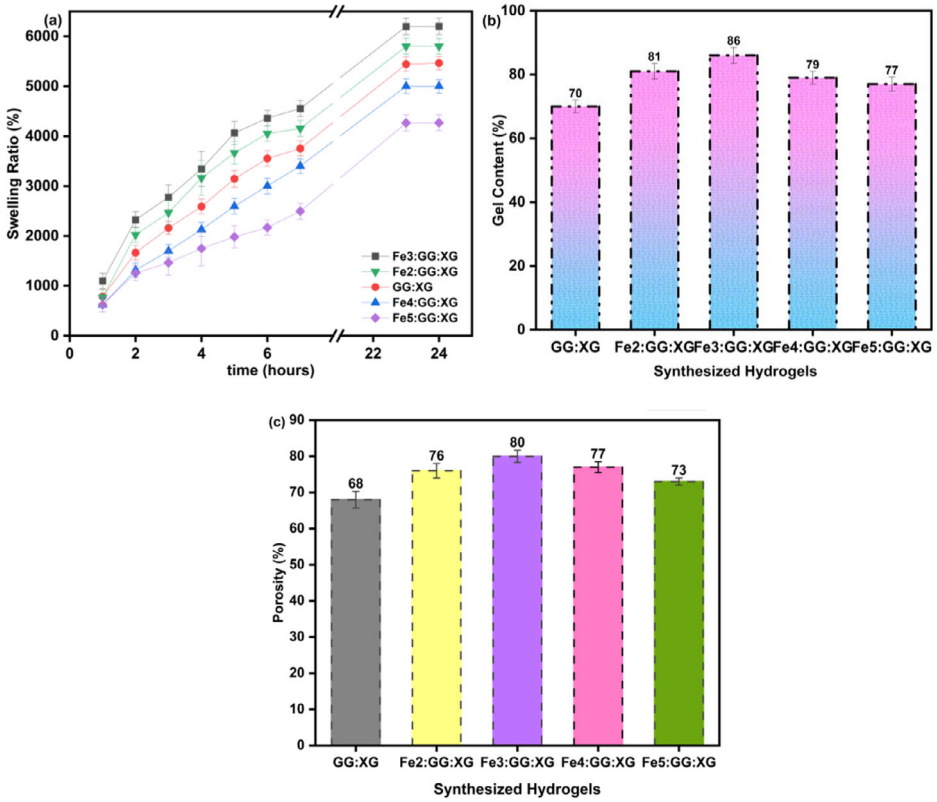



**Figure 1.** Plots for (a) swelling studies, (b) gel content of the hydrogel, and (c) porosity analysis.

#### *4.2.3. FTIR*

The FTIR spectra of guar gum (GG), xanthan gum (XG), MBA, control hydrogel (GG:XG), and Fe3:GG:XG hydrogel are presented in Figure 2. The spectrum of GG and XG revealed that the band appears at 3264 and 3365 $cm^{-1}$, respectively, while in GG:XG hydrogel and Fe3:GG:XG, the broadband appears at 3357 and 3344 $cm^{-1}$ are attributed to the overlapping of -NH and -OH groups stretching. The peaks at 1557 and 1560 $cm^{-1}$ suggested the asymmetric stretch of the $COO^{-}$ group, while the symmetric stretch at 1412 and 1408 $cm^{-1}$, respectively.

The spectrum of MBA imparts a peak at 1226 $cm^{-1}$ associated with C—N stretching, which shifted to 1032 and 1041 $cm^{-1}$ in control hydrogel (GG:XG) and Fe3:GG:XG hydrogel.[40,41] The peak shifting suggests that there is some interaction between the hydrogels and the MBA, resulting in lesser conjugation due to crosslinking in the matrices. In addition, a peak was observed at 547 attributed to the Fe—O bond in Fe3:GG:XG hydrogel, which is in good correlation with the reported literature.[42,43] Hence, it confirms the successful incorporation of FeO NPs into the hydrogel matrix.

#### *4.2.4. PXRD*

The PXRD patterns of FeO NPs, GG:XG, and Fe3:GG:XG hydrogels are shown in Figure 3. In the case of FeO NPs, the peaks of higher intensity were obtained at 31.89°, 33.32°, 35.80°, 45.53°, and 54.14°, while peaks with lower intensity were obtained at 24.25°, 40.91°, 49.46°, 56.54°, 57.71°, 62.54°, and 64.19° are in good agreement with the available reported literature.[44,45] The average size of FeO NPs was calculated using the Debye–Scherrer equation.[46]

$$D = \frac{0.89\lambda}{\beta Cos\ \theta} \tag{7}$$

here $D$ represents particle size in nm of crystals, $\lambda$ (0.15418 nm) is wavelength of X-ray, $\beta$ (radians) is full width at high maximum (FWHM), and $\theta$ is diffraction angle. Hence, the average size of FeO NPs was calculated using Instanano software, and it was found to be 18 nm.

In the case of GG:XG and Fe3:GG:XG hydrogel, a similar PXRD pattern was observed, as shown in Figure 3c,d. It was observed that the broad peaks were obtained at ~20° and ~36°, suggesting the amorphous nature of hydrogels. The crystallinity of FeO NPs was diminished, but a slight enhancement in the intensity of Fe3:GG:XG hydrogel was observed, suggesting the uniform dispersion of FeO NPs to the hydrogel matrix. Hence, it can be concluded that FeO NPs have not affected the hydrogel matrix.[47]

#### *4.2.5. TGA*

The thermal stability of synthesized nanoparticles, GG:XG, and Fe3:GG:XG was performed to assess the interaction between the FeO NPS and hydrogels, as presented in Figure 4. The thermogram of FeO NPs does not show significant weight loss. It was found that ~4% of weight loss was observed in FeO NPs, while four steps of weight loss were observed in GG:XG and


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

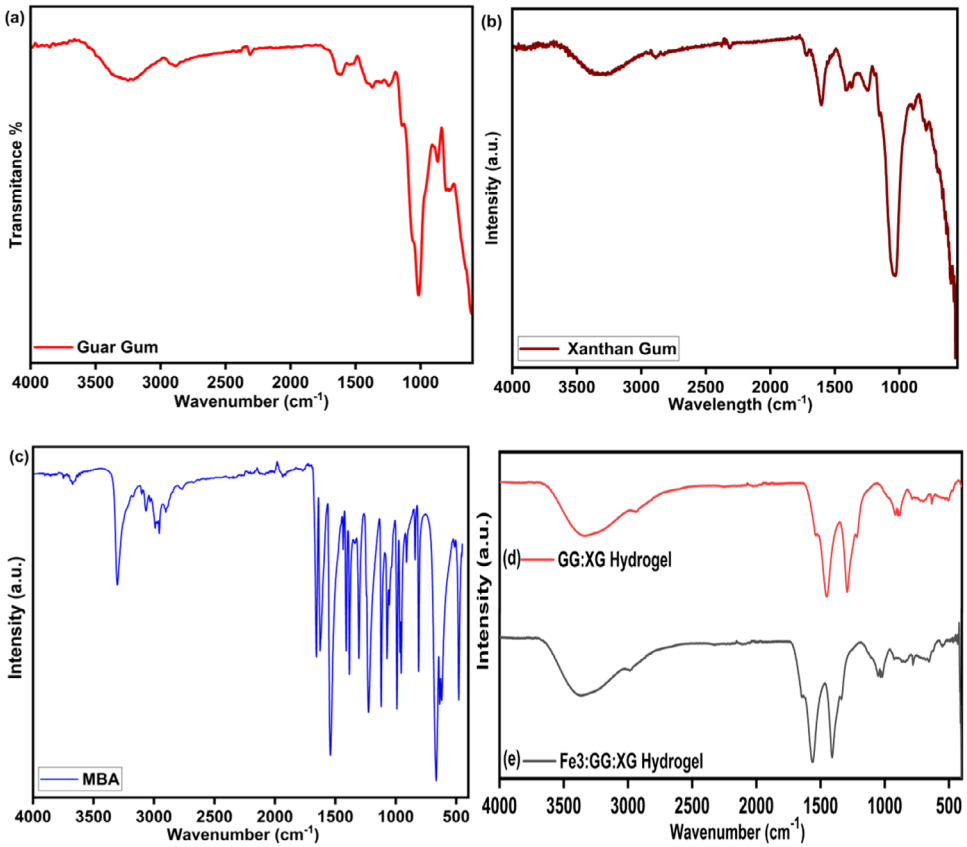



**Figure 2.** Plots for FTIR of (a) GG, (b) XG, (c) MBA, (d) GG:XG hydrogel, (e) Fe3:GG:XG hydrogel nanocomposite.

Fe3:GG:XG hydrogel. The initial stage is associated with the loss of water molecules with a weight loss of 18.83% and 14.53% at 30–273 °C, respectively. The other stage is associated with the degradation of the functional group's polymer backbone, that is, -OH group of hydrogel matrices of GG:XG and Fe3:GG:XG hydrogel, with weight loss of 31.99% at 273–348 °C, and 27.92% at 273–370 °C, respectively. The third stage is attributed to the breakdown of crosslinking within the polymeric network, with weight loss of 63.15% at 348–407 °C, and 60.07% at 348–525 °C. Finally, the fourth stage suggested the complete degradation of the hydrogel, with 4.5% of residue left behind at 407–768.05 °C and 12.54% at 525–666.43 °C, respectively.[48] It was found that both the hydrogels have almost similar patterns of thermogram, but the thermogram of FeO NPs-loaded XG/GG/PSA hydrogel nanocomposite has slightly higher thermal stability as compared to XG/GG/PSA hydrogel, suggesting that incorporation of filler enhanced the thermal stability.[42]

#### 4.2.6. SEM

SEM micro images of FeO NPs, GG:XG, and Fe3:GG:XG hydrogels are shown in Figure 5. The SEM microimage revealed the spherical and rod-shaped morphology of FeO NPs. The average particle size of FeO NPs was calculated using Image J software, and it was found to be 0.45 µm. The SEM micrograph of GG:XG shows an irregular and porous surface, presented in Figure 5b, while in Figure 5c, the SEM micrograph of Fe3:GG:XG hydrogel has irregular surfaces and fewer pores are visible, suggesting that pores were covered by FeO NPs[13]. However, FeO NPs may interact with the polymeric matrix, which can increase hydrophilicity in the matrix or may change the crosslinking density of the hydrogel matrix. The incorporation of FeO NPs into hydrogel may enhance free volume or may help to loosen the polymeric network, which allows the water molecules to penetrate and retain them in the vicinity of the hydrogel network. Further, to confirm the presence of FeO NPs in the Fe3:GG:XG hydrogel, EDAX analysis was done. Figure 5d presents an EDAX plot of the optimized Fe3:GG:XG hydrogel.

#### 4.2.7. DLS

The dynamic light scattering method was used to measure the hydrodynamic average size of the FeO NPs. It was concluded that the average size of the particle was 214.7 (d. nm), presented in Figure 6a.

#### 4.2.8. HRTEM

The HRTEM image of synthesized FeO NPs is presented in Figure 6b, which shows that particles have spherical and rod-type shapes. The average size of FeO NPs was calculated using ImageJ software, and it was found to be 21 nm. Moreover, the average particle size of FeO NPs was also calculated by XRD, and it was found to be 18 nm, which shows a good correlation with


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

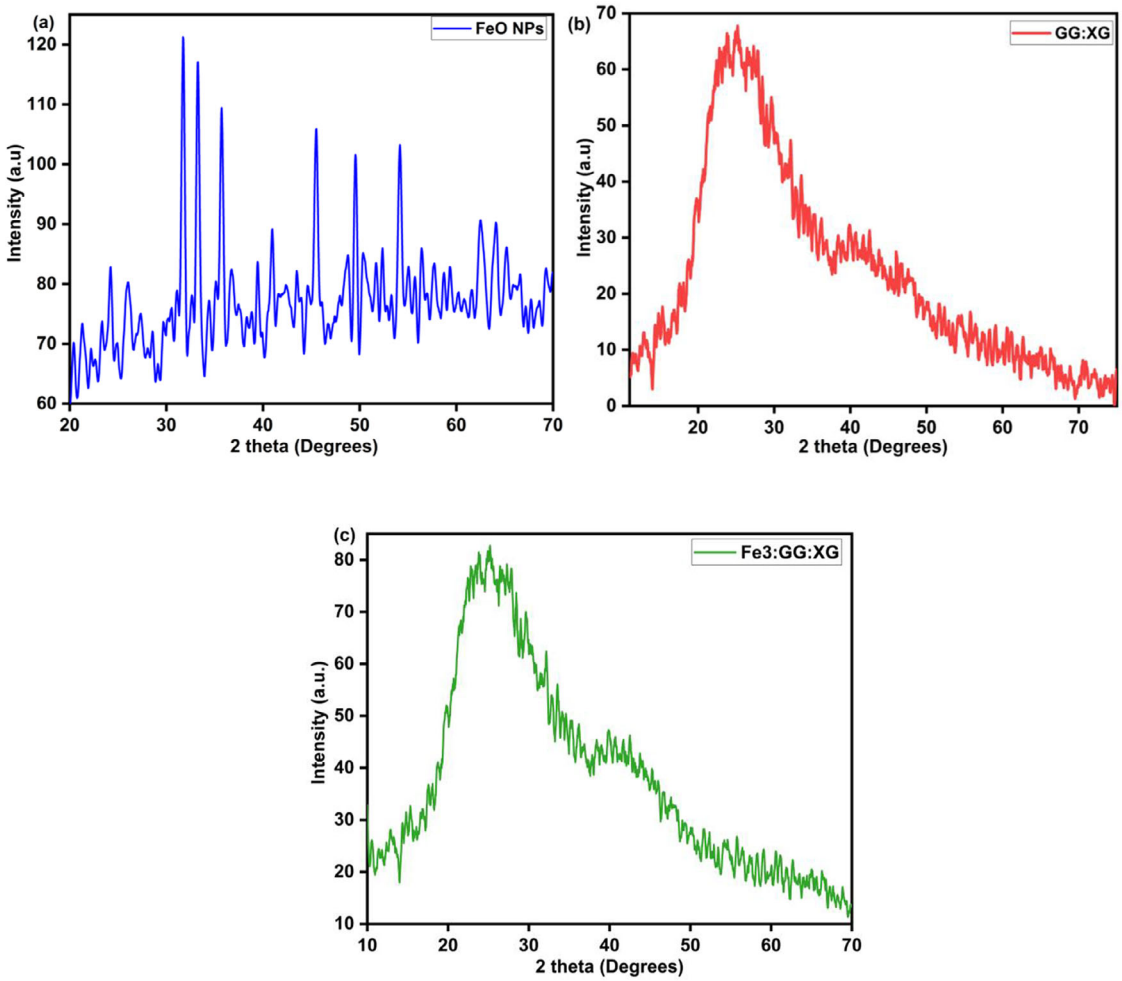



**Figure 3.** PXRD of (a) FeO NPs, (b) GG:XG hydrogel, and (c) Fe3:GG:XG hydrogel nanocomposite.

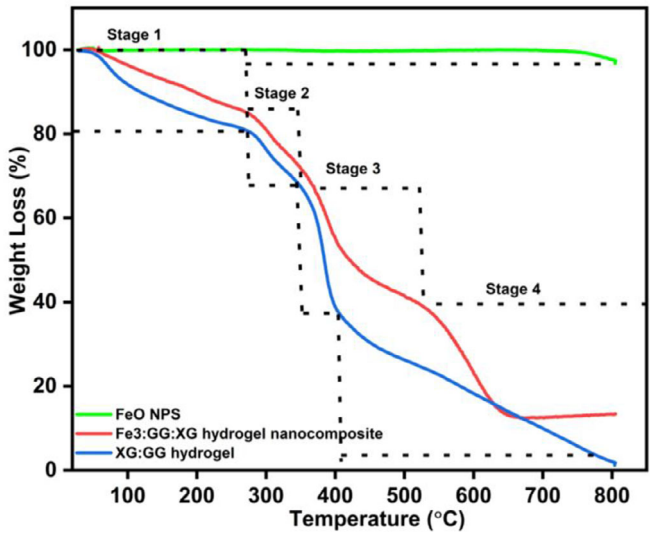



**Figure 4.** Plot of TGA of (a) FeO NPs, (b) GG:XG hydrogel, and (c) Fe3:GG:XG hydrogel nanocomposite.

the TEM analysis data. Hence, it confirms that the synthesized particles are in nanoparticles.

#### *4.2.9. X-ray Photoelectron Spectroscopy (XPS)*

It is a flexible surface analysis method that may be applied to identify the chemical states and composition of a material. The ionic states of iron have been demonstrated to influence the peak positions of Fe $2p_{1/2}$ and Fe $2p_{3/2}$. The Fe $2p_{1/2}$ and Fe $2p_{3/2}$ satellite peak positions are likewise very sensitive to the oxidation states, and these peaks have been utilized to qualitatively identify the ionic states of iron.[49] In the Fe 2p spectrum, two major peaks, such as $Fe2p_{3/2}$ and Fe $2p_{1/2}$ at 711 eV and 725 eV, respectively, were observed, as presented in Figure 6c. It was reported that the satellite peak near to 724 eV suggests Fe is in +3 oxidation state, which is commonly present in nanoparticles.[50]

### 4.3. Adsorption Study

#### *4.3.1. Contact Time*

The effect of contact time (0–90 min) was investigated for CV dye removal, as shown in Figure 7a. The results demonstrated that as time increased, the percentage of dye removal increased. The


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

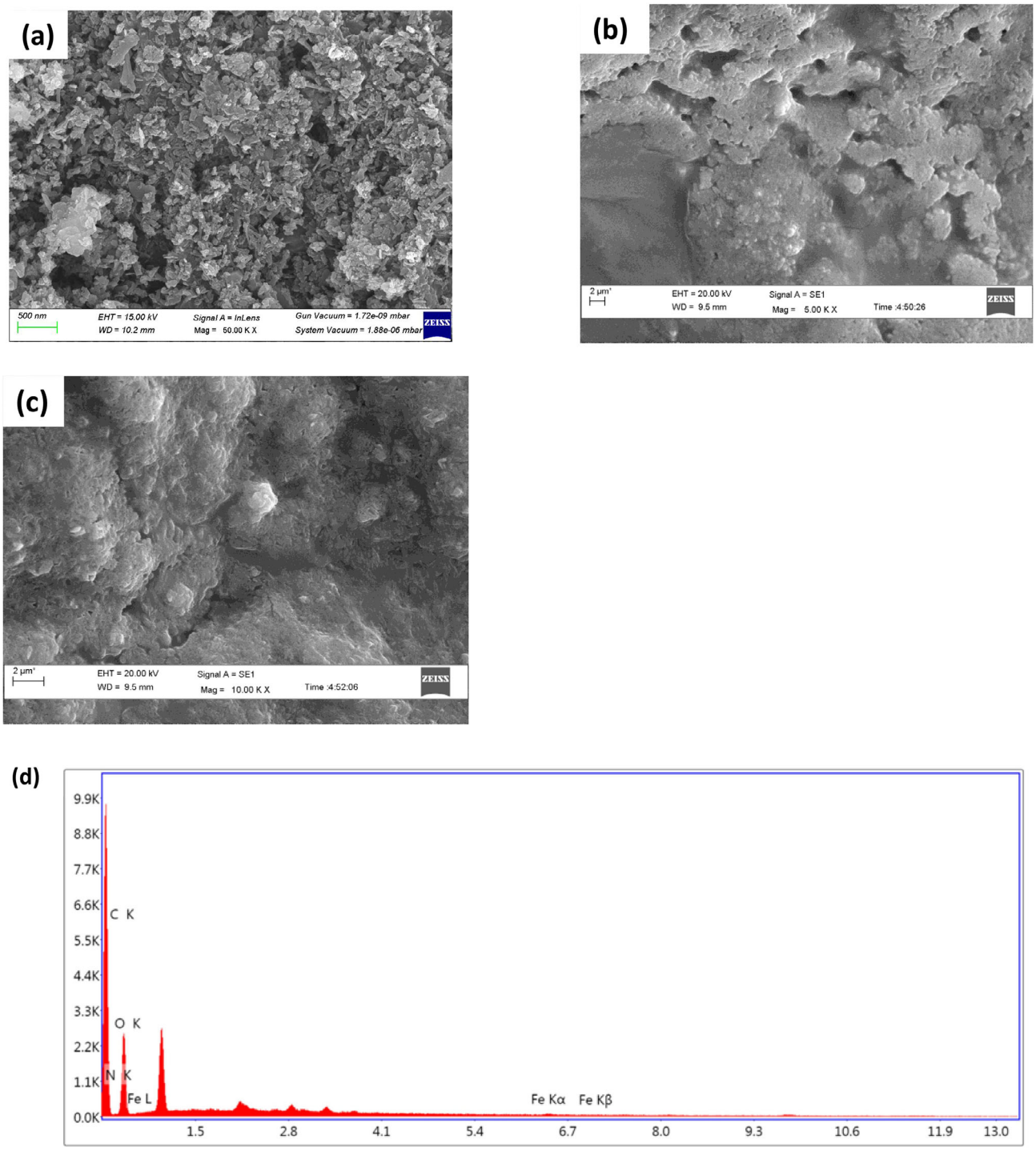



**Figure 5.** SEM of (a) FeO NPs, (b) GG:XG hydrogel, (c) Fe3:GG:XG hydrogel, and (d) EDAX of Fe3:GG:XG hydrogel nanocomposite.

maximum percentage of dye removal was found to be 96.8% in 90 min. After 90 min, no valuable change was observed, suggesting that equilibrium has been reached, as presented in Figure 7b. The reason for this result may be attributed to the fact that, initially, there might have been a substantial amount of vacant space for the adsorption of CV dye molecules, leading to fast adsorption. Further, after a particular time, the voids are occupied by dye molecules, leading to a decrease in adsorption, resulting in the saturation of the synthesized material.[51]

#### *4.3.2. Concentration of Dye and Adsorbent Dosage*

For the concentration study, the dye solution was varied from 10 and 60 mg/L with 0.035 g of adsorbent for 90 min, and the observation trend is shown in Figure 8a. The adsorption capacity at equilibrium ($q_e$) was increased from 6.15 to 37.75 mg/g for CV dye, showing that as the dye concentration increases, the adsorption capacity at equilibrium also increases. Yadav and the group have reported similar results. An increase in mass transfer from


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

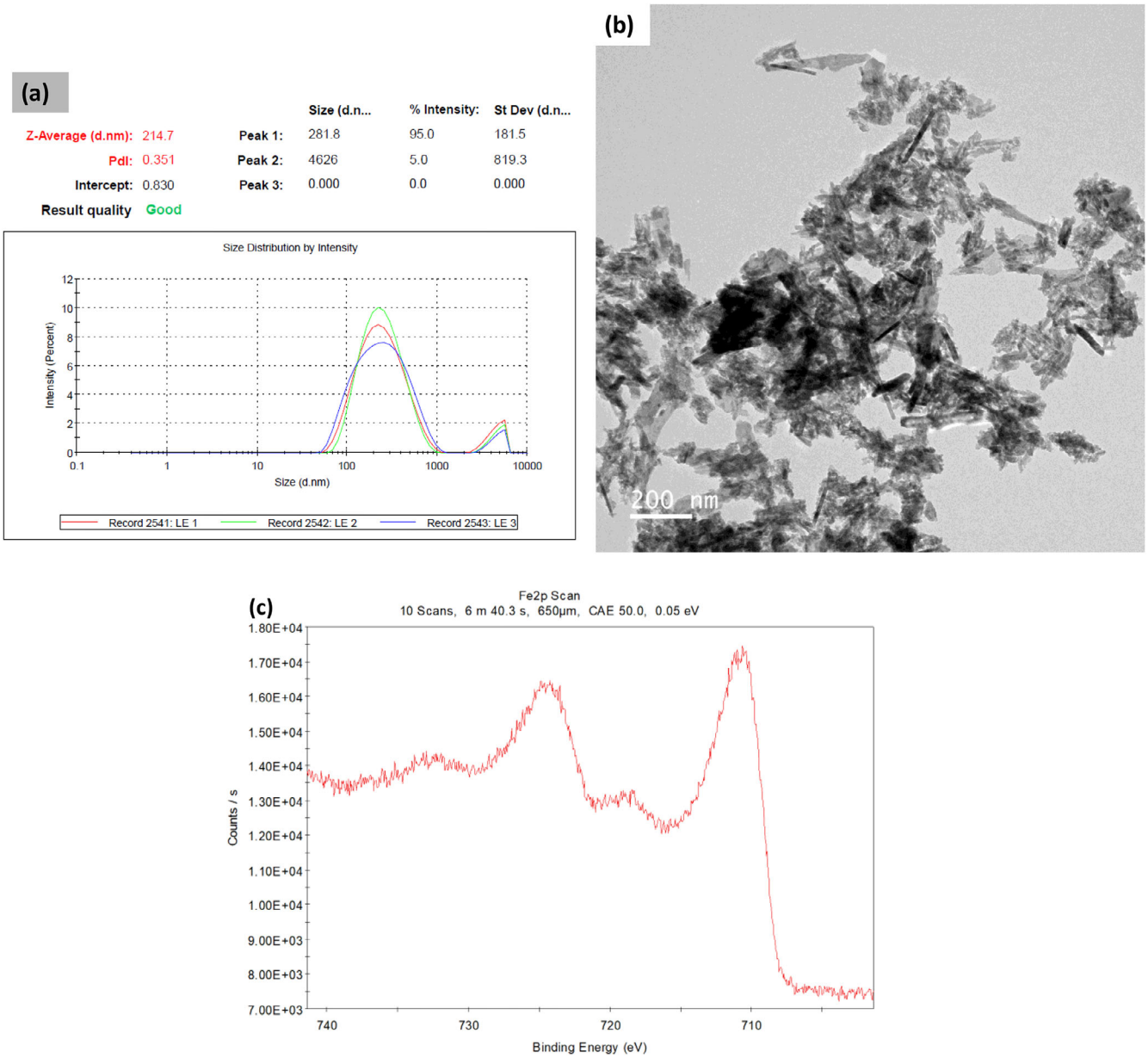



**Figure 6.** (a) Size distribution by DLS, (b) HRTEM image of FeO NPs, and (c) XPS of FeO NPs.

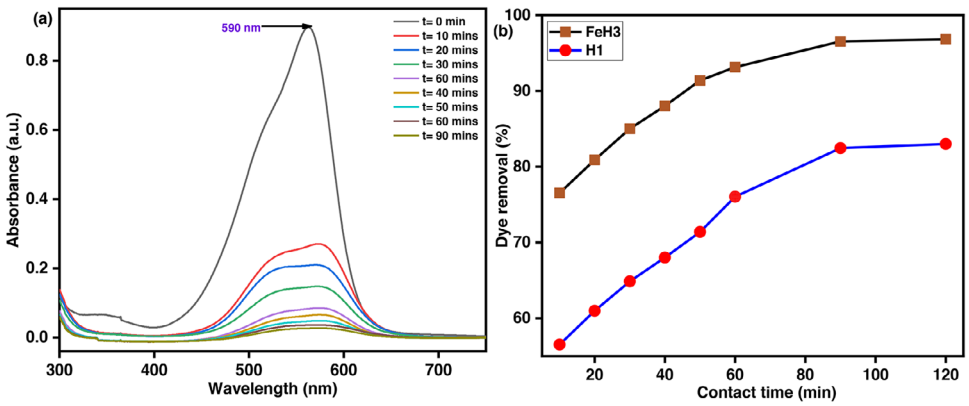



**Figure 7.** Plots for (a) UV–vis adsorption spectra, (b) contact time on dye removal capacity.


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

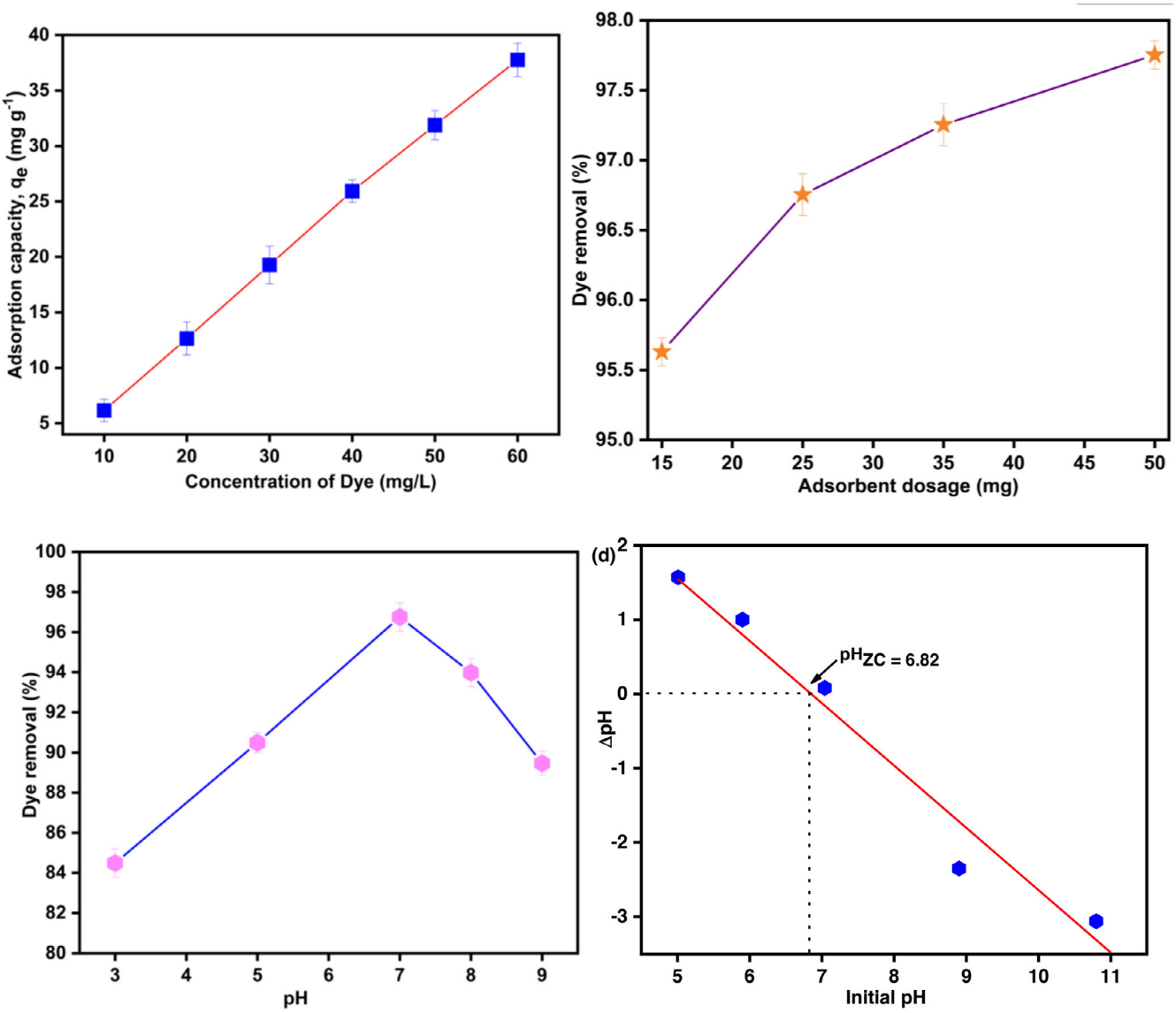



**Figure 8.** Plot for the effect of (a) concentration, (b) adsorbent dosage, (c) pH on dye removal capacity, and (d) $pH_{zc}$ for as adsorbent.

solution to solid is attributed to a higher concentration because it raises the probability of molecular collisions with adsorbent surfaces. The molecules in the solution phase are removed at a higher rate due to more collisions. In conclusion, the initial higher concentration of dye shows a more significant concentration slope between the dye solution and the adsorbent phase, which raises the adsorption rate.[52]

The experiments were conducted using various adsorbent doses (0.015–0.06 g), while maintaining constant values for other parameters such as reaction time of 90 min and dye concentration of 40 mg/L. This was done to examine the effects of variation of adsorbent dosages on the dye removal efficiency. The adsorption efficiency increases with an increase in the amount of adsorbent, as illustrated in Figure 8b. The percentage of CV removal increased significantly from 85.63% to 98.19%, indicating that the more active sites available for sorption lead to a higher removal percentage of dye.[53]

#### *4.3.3. Influence of pH and Point Zero Charges ($pH_{ZC}$)*

The pH of the dye solution plays a crucial role in controlling the effectiveness of adsorbents in water treatment applications. The dye removal percentage is shown in Figure 8c. The point zero charge was found to be 6.82, as shown in Figure 8d, which suggests that pH > $pH_{ZC}$ shows the surface of the adsorbent is negatively charged, and pH < $pH_{ZC}$ shows the positive surface of the adsorbent.

It was found that when the pH increases from 3 to 7, the percentage removal of dye increases from 84.49% to 96.75%. It may be attributed that at lower pH, the adsorbent surface has a positive charge, which results in less adsorption of cationic dye molecules on the adsorbent surface. The maximum dye removal was observed in the case of pH 7, attributed to the fact that there is no interfering ion present in the solution, thereby leading to maximum adsorption of dye. Further, on moving


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

from pH 7 to 9, removal efficacy decreased from 96.75% to 89.46%, and a decrease in adsorption efficiency was observed due to the excess of $Na^+$ in the basic medium, which indicates that there is competition between the dye molecule and $Na^+$ ion, leading to slight decrease in adsorption. Hence, it is concluded that the synthesized hydrogel shows an anionic surface ($pH > pH_{ZC}$), which leads to the adsorption of cationic dye molecules.[54]

### 4.4. Adsorption Isotherm

The present study used three isotherm models to analyze equilibrium adsorption. The primary model, Langmuir isotherm, is commonly employed to describe monolayer adsorptions and suggests that adsorption exhibits on uniform, homogenous surfaces with identical sites on an adsorbent. The model's linearized equation is as follows:[55]

$$\frac{1}{q_e} = \frac{1}{q_m C_e k_L} + \frac{C_e}{q_m} \tag{8}$$

$q_e$ is the absorption capacity (mg/g), $C_e$ is the dye concentration at equilibrium (mg/L), $k_L$ refers to the Langmuir constant (L/mg), and $q_m$ defines "monolayer" or "maximum adsorption capacity" (mg/g). The $R_L$ is a dimensionless factor represented in the equation that allows us to assess the Langmuir isotherm and is described as follows:

$$R_L = \frac{1}{1 + k_L C_0} \tag{9}$$

where $C_0$ denotes the initial concentration (mg/L) and the $R_L$ value shows the adsorption's features, such as favorable ($0 < R_L < 1$), unfavorable ($R_L > 1$), linear ($RL = 1$), and irreversible ($RL = 0$).

The second model that might be very helpful in comprehending the sorption mechanism for heterogeneous surfaces is the Freundlich isotherm. The model's linearized form is as follows:[56]

$$\log q_e = \log K_f + \frac{1}{n_f} \log C_e \tag{10}$$

where constants $K_F$ and $1/n_f$ represent the adsorption capacity and intensity, respectively. The values of $1/n$ and $K_F$ can be examined by the slope and intercept of the graph of ln $q_e$ versus ln $C_e$. If the adsorption process is favorable ($0 < 1/n_f < 1$), or unfavorable ($1/n_f = 1$), and the adsorption process is irreversible, $1/n_f$ should be zero.

The third Temkin isotherm model accounts for the indirect interactions between adsorbents and adsorbates. The heat of adsorption is expected to drop linearly rather than logarithmically as surface coverage increases. The linearized forms of the Temkin adsorption model are as follows:[57]

$$q_e = B \ln A_T + B \ln C_e \tag{11}$$

where $B$ and $A_T$ represent the heat of sorption and the binding constant at equilibrium, respectively.

Figure 9 shows the results for the various adsorption isotherms used in the current investigation, and Table 4 provides the required parameters for these isotherms.

The calculated adsorption data correlate with the Langmuir isotherm, with $R^2 = 0.988$. This demonstrates the monolayer adsorption of CV dye on the synthesized hydrogel as an adsorbent, suggesting that CV dye was adsorbed favorably. The $R_L$ was found to be 0.235. Furthermore, the maximum adsorption capacity ($q_m$) was calculated and found to be 130 mg/g.

### 4.5. Kinetics

The adsorption rates, design, and model adsorption processes can be predicted by kinetic parameters. The three distinct kinetic models were coupled with the acquired data to examine the adsorption mechanism. To illustrate adsorption rates, one can use linear Lagergren's first-order rate equations, which are represented as follows:[58]

$$\ln (q_e - q_t) = \ln q_e - K_1 t \tag{12}$$

The uptake capacity at equilibrium is represented by $q_e$, and time $t$ is represented by $q_t$ (mg/g), respectively. $K_1$ represents the pseudo-first-order rate constant ($min^{-1}$), calculated from the linear plot of ln ($q_e - q_t$) versus t.

Ho and McKay suggested that a pseudo-second-order model might also explain the kinetics of adsorption. The linearized equation for the kinetic model is as follows:[59]

$$\frac{t}{q_t} = \frac{1}{k_2 q_e^2} + \frac{t}{q_e} \tag{13}$$

where $k_2$ represents the rate constant (g/mg/min).

The Elovich model clarifies the adsorption–desorption kinetics and has linear equations:[60]

$$q_t = \frac{1}{\beta} \ln (\alpha\beta) + \frac{1}{\beta} \ln t \tag{14}$$

where $q_t$, $\alpha$, and $\beta$ represent the adsorbent's absorption capacity at time $t$ (mg/g), the initial rate of adsorption (mg/g/min), and the desorption constant (g/mg). If $\alpha > \beta$, chemical adsorption occurs between the adsorbent and adsorbate phases. Hence, the findings of the kinetic models used in this study are shown in Figure 10. All the calculated kinetic variables are listed in Table 5.

The pseudo-second-order model, with an $R^2 = 0.998$ value, correlates well with the experimental data, suggesting that a pseudo-second-order is the rate-determining step in the adsorption process facilitated by chemical interactions. The value of $h$ was computed to be 5.75 mg/g $min^{-1}$ for CV dye.

Thus, it was determined that there might be three processes involved in the dye adsorption process using fabricated hydrogel: Adsorption occurs on the surface of the synthesized adsorbent in three stages: (i) fast physical adsorption, (ii) chemical adsorption, and (iii) saturation of the adsorbent surface at the equilibrium stage.


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

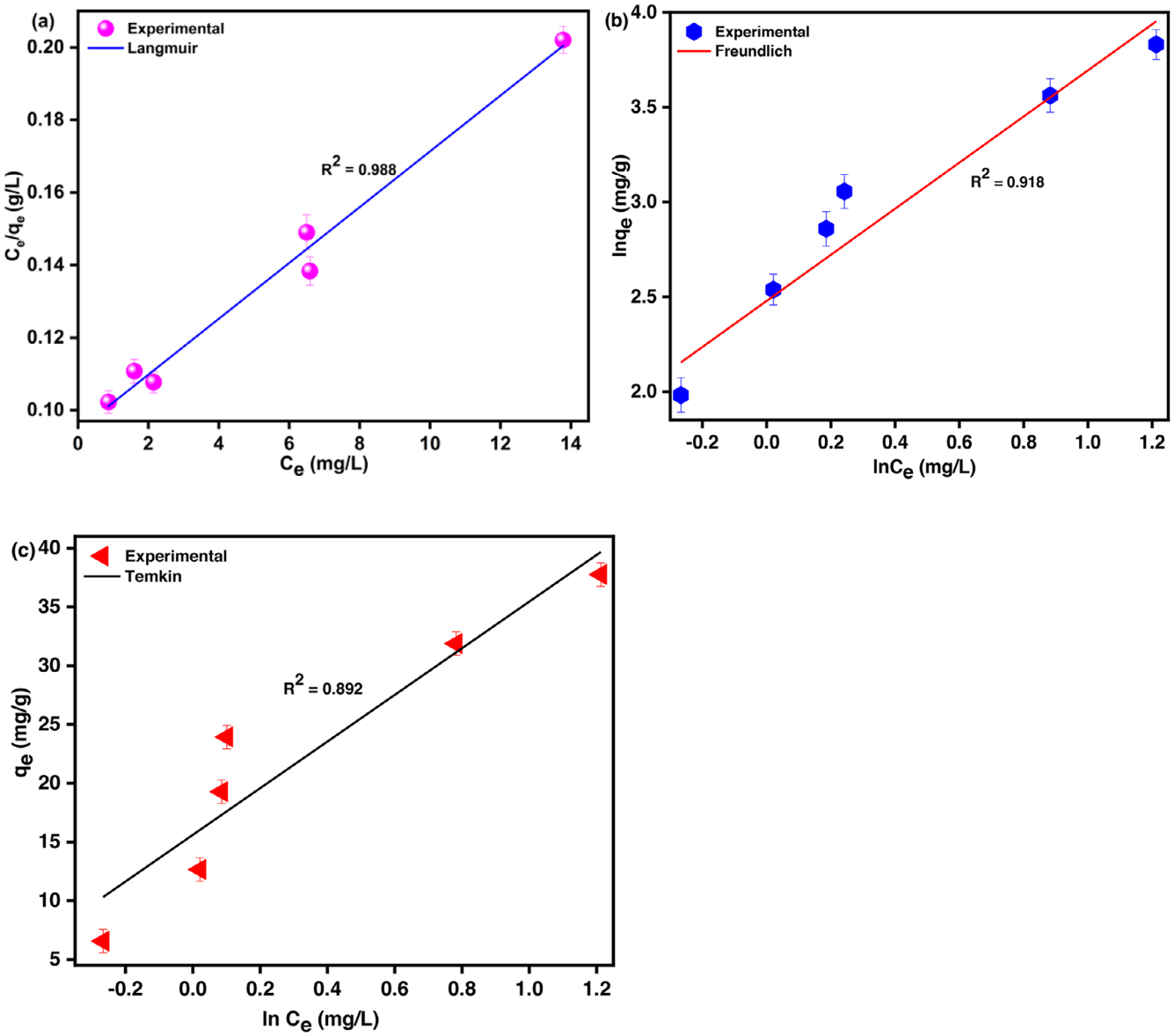



**Figure 9.** Adsorption isotherms curve for CV adsorption fitted to (a) Langmuir, (b) Freundlich, and (c) Temkin models.

**Table 4.** Calculated parameters for isotherm models.

| Models | Slope and Intercept | Parameters | Parameter Values |
|---|---|---|---|
| Langmuir | Slope = 0.0077<br>Intercept = 0.09446 | $R^2$ | 0.988 |
| | | $q_m$ (mg/g) | 130.03 |
| | | $K_L$ (L/mg) | 0.082 |
| | | $R_L$ | 0.235 |
| Freundlich | Slope = 1.216<br>Intercept = 2.479 | $R^2$ | 0.918 |
| | | $K_F$ (mg/g) (L mg)$^{1/n}{}_f$ | 301.30 |
| | | $1/n_f$ | 0.822 |
| Temkin | Slope = 19.848<br>Intercept = 15.603 | $R^2$ | 0.892 |
| | | $B$ (J/mol) | 19.848 |
| | | $A_T$ (L/g) | 6.109 |

### 4.6. Desorption, Recyclability, and Stability

For the desorption study, the batch experiment was performed using 0.1 M HCl for the desorption study, and the time span was 0.1–2 h. The study was conducted using the batch experiment method. In consecutive cycles, the adsorbent was collected and washed with 0.1 M HCl and 0.1 M NaOH solutions to desorb the adsorbed CV that had individually adsorbed on the sorbent to regenerate the available binding sites. Afterward, the adsorbent was washed with distilled water to remove excess HCl and NaOH. The washed hydrogels were dried at ambient temperature and weighed before use in the next regeneration cycle. It was observed that as the desorption time increased, the desorption efficiency was enhanced.


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

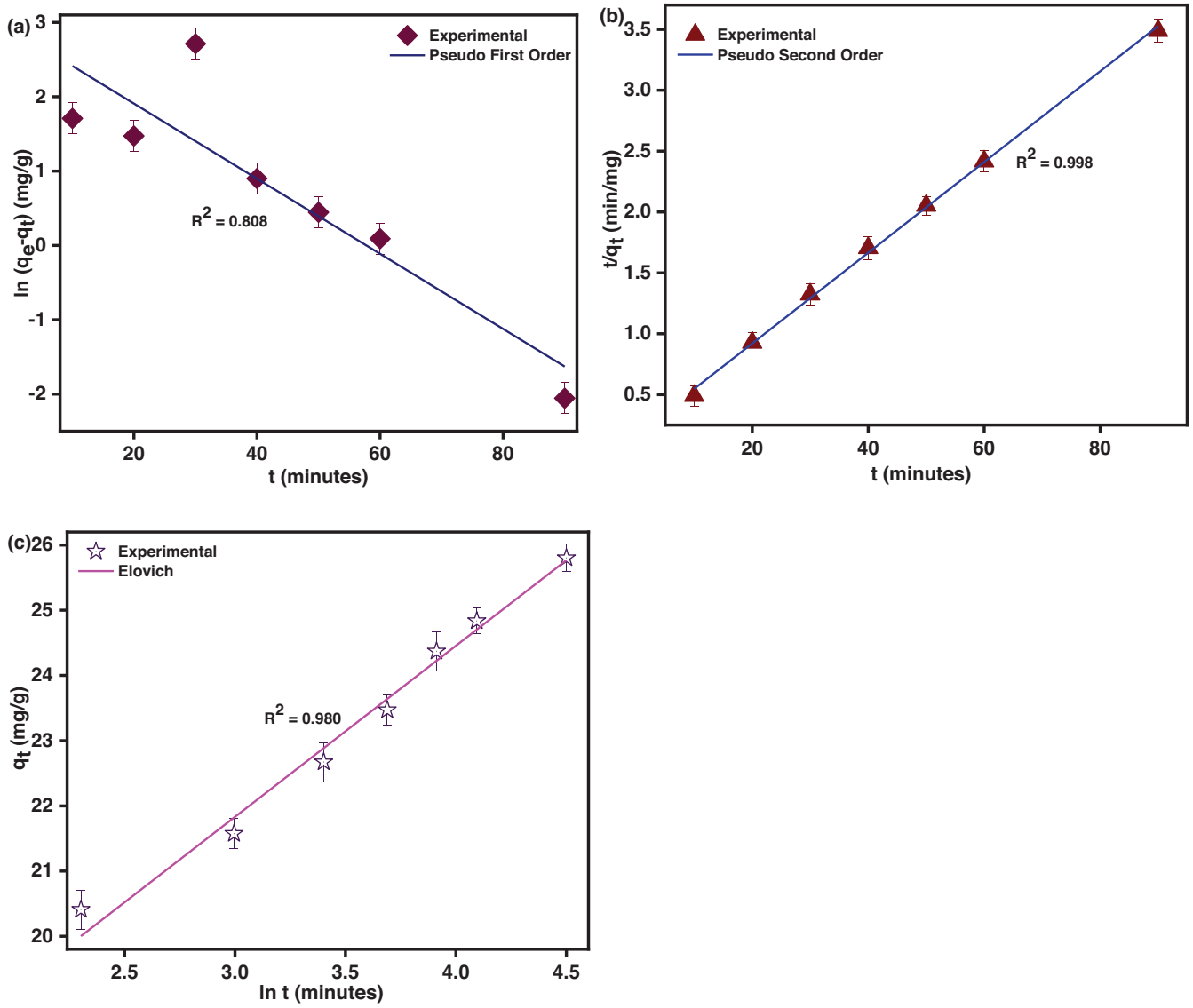



**Figure 10.** Adsorption kinetics linear curves for CV adsorption (a) pseudo-first order, (b) pseudo-second order, and (c) Elovich model.

**Table 5.** Calculated parameters for kinetic models.

| Kinetics Models | Plot | Parameters | Values of Parameters |
|---|---|---|---|
| (Lagregren's) pseudo-first order | ln ($q_e - q_t$) vs t | $R^2$ | 0.808 |
| | | $K_1$ (per min) | 0.0505 |
| | | $q_e$ (mg/g) | 18.51 |
| Pseudo-second order | $t/q_t$ vs $t$ | $R^2$ | 0.998 |
| | | $K_2$ (g/ mg/min) | 0.008 |
| | | $q_e$ (mg/g) | 26.831 |
| | | $h$ (mg/g/ min) | 5.75 |
| Elovich model | $q_t$ vs ln $t$ | $R^2$ | 0.980 |
| | | $\alpha$ (mg/g/min) | 77.78 |
| | | $\beta$ (g/mg) | 0.381 |

Further, to assess the synthesized adsorbent's recyclability for the CV dye adsorption from the aqueous phase, the desorbed adsorbent was collected for regeneration and then utilized again in the adsorption process. It was found that more than 88% of the dye was removed in the first five cycles, as depicted in Figure 11a. As a result, it can potentially serve as an adsorbent. Hence, the adsorbent was found to be reusable for up to five cycles.

Then, the PXRD pattern of the synthesized hydrogel was also recorded after the regeneration of the adsorbent to assess its stability, as shown in Figure 10b. It was found that no significant difference in the PXRD pattern was observed, suggesting that there was no textural deformation of the adsorbent surface after regeneration of the adsorbent. Hence, it confirms the stability of the adsorbent.


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

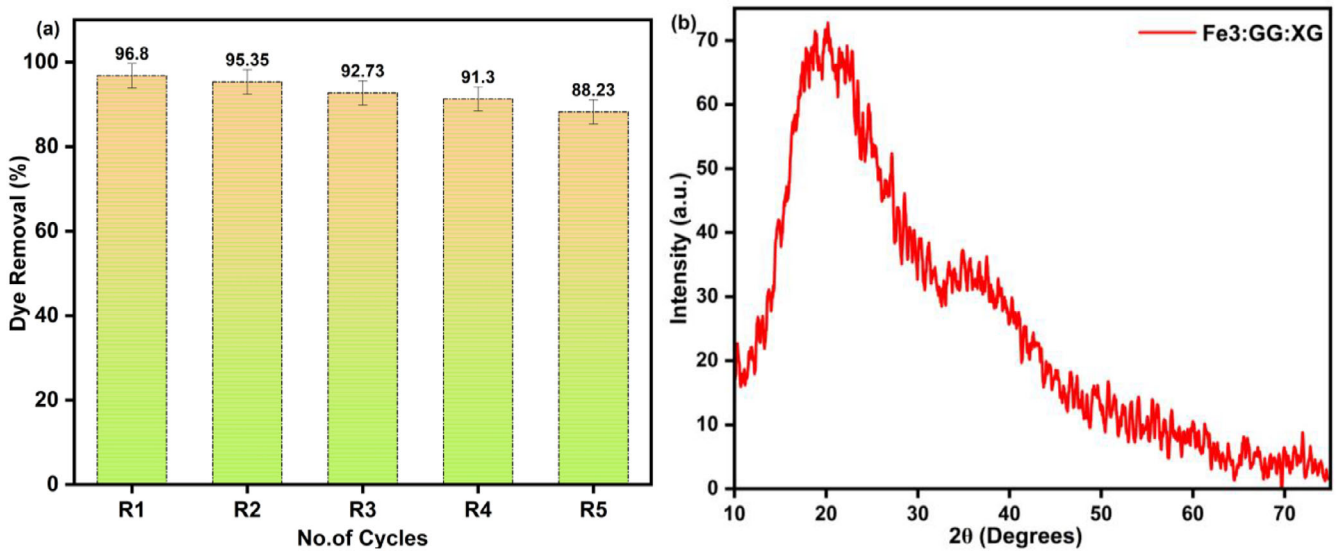



**Figure 11.** Plot for (a) regeneration cycles of adsorbent on crystal violet dye removal, (b) XRD plot of Fe3:GG:XG after regeneration.

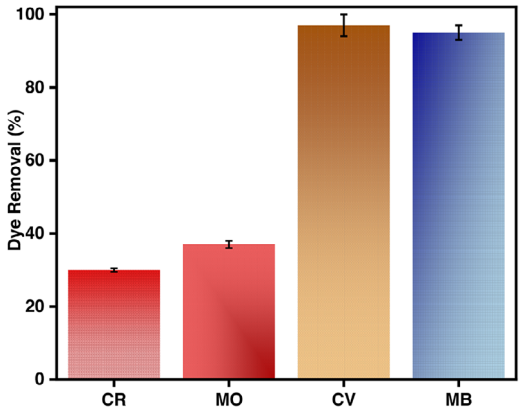



**Figure 12.** Plot for the effect of various dyes on the adsorbent.

### 4.7. Effect of Various Dyes

The charges on the dyes have a major impact on how effectively the adsorbent removes dye molecules. Figure 12 illustrates the use of synthesized adsorbent for the removal of dyes, such as Congo red (CR) and methyl orange (MO), as well as crystal violet (CV) and methylene blue (MB). It was observed that the synthesized adsorbent has very low adsorption efficacy for anionic dyes like CR and MO, while exceptional adsorption efficiencies were exhibited by the synthesized adsorbent for cationic dyes. The electrostatic attractive forces and hydrogen bonding are responsible for cationic dye removal because dye molecules show linkage with the negatively charged site on the hydrogel adsorbent surface.[31]

## 5. Conclusion

This work describes the facile one-pot synthesis and characterization of FeO NPs and FeO NPs incorporated hydrogel nanocomposite. It was concluded that the synthesized hydrogel as an adsorbent shows 97% CV dye removal. The maximal capacity ($q_{max}$) was 130 mg/g. The isotherm plots and kinetic model revealed that the experimental data were found to be in good correlation with the Langmuir isotherm model, having $R^2 = 0.988$, and pseudo-second-order having $R^2 = 0.998$, respectively, suggesting that monolayer adsorption and chemisorption are the rate-limiting steps in the adsorption process. Further, dye removal efficacy was found to be 88% after five regeneration cycles. Hence, the synthesized sorbent can be used to sequester cationic CV dye.

## Acknowledgements

The authors would like to acknowledge Delhi Technological University (DTU) for its support and for providing the necessary resources. Author Priyanka Yadav is greatly thankful to the DTU for providing DTU fellowship.

## Conflict of Interests

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.

**Keywords:** Adsorption · Crystal violet · FeONPs · Hydrogel nanocomposite

[1] V. K. Gupta , I. Ali, T. A. Saleh, A. Nayak, S. Agarwal, *RSC Adv.* **2012**, *2*, 6380.
[2] S. Jana, J. Ray, B. Mondal, T. Tripathy, *Appl. Clay Sci.* **2019**, *173*, 46–64.


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

[3] M. Jaymand, *Iran. J. Chem. Chem. Eng.* **2022**, *42*, 1–26.
[4] V. Katheresan, J. Kansedo, S. Y. Lau, *J. Environ. Chem. Eng.* **2018**, *6*, 4676–4697.
[5] N. Sen, N. R. Shefa, K. Reza, S. M. A. Z. Shawon, M. W. Rahman, *Sci. Rep.* **2024**, *14*, 1–17.
[6] N. Kumar, R. Gusain, S. Pandey, S. S. Ray, *Adv. Mater. Interfaces* **2023**, *10*, 2201375.
[7] S. J. Yang, J. H. Kang, H. Jung, T. Kim, C. R. Park, *J. Mater. Chem. A 1*, **2013**, 9427–9432.
[8] S. Shaker, S. Zafarian, C. S. Chakra, K. V. Rao, *Int. J. Innov. Res. Sci. Eng. Technol.* **2013**, *2*, 2969–2973.
[9] T. Huang, M. Yan, K. He, Z. Huang, G. Zeng, A. Chen, M. Peng, H. Li, L. Yuan, G. Chen, *J. Colloid Interface Sci.* **2019**, *543*, 43–51.
[10] H. Gao, J. Jiang, Y. Huang, H. Wang, J. Sun, Z. Jin, J. Wang, J. Zhang, *SN Appl. Sci.* **2022**, *4*, 2523–3971.
[11] C. Lemarchand, R. Gref, P. Couvreur, *Eur. J. Pharm. Biopharm.* **2004**, *58*, 327–341.
[12] V. B. Bueno, R. Bentini, L. H. Catalani, D. F. S. Petri, *Carbohydr. Polym.* **2013**, *92*, 1091–1099.
[13] I. Rani, S. G. Warkar, A. Kumar, *Mater. Today Commun.* **2023**, *35*, 105635.
[14] M. Aurélio Da Silva Carvalho, L. Gustavo Lacerda, *J. Food Nutr. Res.* **2015**, *54*, 185–194.
[15] D. Mudgil, S. Barak, B. S. Khatkar, *J. Food Sci. Technol.* **2014**, *51*, 409–418.
[16] S. Ahmad, M. Ahmad, K. Manzoor, R. Purwar, S. Ikram, *Int. J. Biol. Macromol.* **2019**, *136*, 870–890.
[17] U. S. K. Madduma-Bandarage, S. V. Madihally, *J. Appl. Polym. Sci.* **2021**, *138*, 1–23.
[18] N. F. Virzì, P. Diaz-Rodriguez, A. Concheiro, V. Pittalà, C. Alvarez-Lorenzo, *Carbohydr. Polym. Technol. Appl.* **2024**, *7*, 100523.
[19] G. Sharma, B. Gupta, P. Sharma, S. K. Sandhu, I. P. Kaur, *Gels* **2022**, *8*, 58.
[20] S. Kanwal, F. Batool, G. Sharif, H. K. Naeem, S. Noreen, H. Y. Gondal, U. B. Kamal, A. Ditta, *Biocatal. Agric. Biotechnol.* **2024**, *58*, 103173.
[21] R. Sharma, S. Kalia, B. S. Kaith, M. K. Srivastava, *Int. J. Plast. Technol.* **2016**, *20*, 294–314.
[22] J. N. Hiremath, B. Vishalakshi, *Polym. Bull.* **2015**, *72*, 3063–3081.
[23] O. Ghazy, M. G. Hamed, M. Breky, E. H. Borai, *Colloids Surfaces A Physicochem. Eng. Asp.* **2021**, *621*, 126613.
[24] A. N. Danthanarayana, D. C. Manatunga, R. M. De Silva, N. V. Chandrasekharan, K. M. Nalin De Silva, *R. Soc. Open Sci.* **2018**, *5*, 181369.
[25] C. Sun, J. S. H. Lee, M. Zhang, *Adv. Drug Delivery Rev.* **2008**, *60*, 1252–1265.
[26] A. Afkhami, S. Sayari, R. Moosavi, T. Madrakian, *J. Ind. Eng. Chem.* **2015**, *21*, 920–924.
[27] A. A. E. A. Elfiky, M. F. Mubarak, M. Keshawy, I. E. T. E. Sayed, T. A. Moghny, *Environ. Dev. Sustain.* **2024**, *26*, 19935–19957.
[28] M. T. M. Hussien Hamad, *Environ. Sci. Eur.* **2023**, *35*, 1–37.
[29] C. Liu, X. Cheng, F. Zhang, F. Lei, P. Li, K. Wang, J. Jiang, *Mater. Today Commun.* **2023**, *34*, 105256.
[30] P. Yadav, S. G. Warkar, A. Kumar, *Polym. Eng. Sci.* **2024**, *64*, 4816–4833.
[31] I. Rani, S. G. Warkar, A. Kumar, *ChemistrySelect* **2023**, *8*, e202301434.
[32] B. H. Hui, M. N. Salimi, *IOP Conf. Ser.: Mater. Sci. Eng.* **2020**, *743*, 012036.
[33] P. Meena, P. Singh, S. G. Warkar, *Colloid Polym. Sci.* **2024**, *302*, 377–391.
[34] M. Suhail, Y. F. Shao, Q. L. Vu, P. C. Wu, *Gels* **2022**, *8*, 155.
[35] P. T. S. Kumar, G. Praveen, M. Raj, K. P. Chennazhi, R. Jayakumar, *RSC Adv.* **2014**, *4*, 65081–65087.
[36] T. Hasan, A. Islam, A. Pal, B. Baran, *Appl. Therm. Eng.* **2020**, *175*, 115361.
[37] G. Issabayeva, S. H. Wong, C. Y. Pang, M. C. Wong, M. K. Aroua, *Int. J. Environ. Sci. Technol.* **2022**, *19*, 3731–3740.
[38] I. Gholamali, M. Yadollahi, *Int. J. Biol. Macromol.* **2020**, *160*, 724–735.
[39] I. Jayawardena, P. Turunen, B. C. Garms, A. Rowan, S. Corrie, L. Grøndahl, *Mater. Adv.* **2023**, *4*, 669–682.
[40] Khushbu, S. G. Warkar, N. Thombare, *Polym. Bull.* **2022**, *79*, 2287–2303.
[41] Kajal, R. Kumar, P. Meena, S. G. Warkar, *Colloid Polym. Sci.* **2023**, *301*, 1313–1323.
[42] Y. Ji, C. Wang, *Front. Bioeng. Biotechnol.* **2023**, *11*, 1–8.
[43] M. Rasoulzadeh, H. Namazi, *Carbohydr. Polym.* **2017**, *168*, 320–326.
[44] S. S. U. Rahman, M. T. Qureshi, K. Sultana, W. Rehman, M. Y. Khan, M. H. Asif, M. Farooq, N. Sultana, *Results Phys.* **2017**, *7*, 4451–4456.
[45] M. Mahdavi, M. B. Ahmad, M. J. Haron, Y. Gharayebi, K. Shameli, B. Nadi, *J. Inorg. Organomet. Polym. Mater.* **2013**, *23*, 599–607.
[46] A. Monshi, M. R. Foroughi, M. R. Monshi, *World J. Nano Sci. Eng.* **2012**, *2*, 154–160.
[47] D. Chauhan, A. Kumar, S. G. Warkar, *Environ. Technol. (U. K.)* **2022**, *43*, 991–1002.
[48] Khushbu, S. G. Warkar, A. Kumar, *Polymer (Guildf).* **2019**, *182*, 121823.
[49] T. Yamashita, P. Hayes, *Appl. Surf. Sci.* **2008**, *254*, 2441–2449.
[50] S. K. Shaw, S. K. Alla, S. S. Meena, R. K. Mandal, N. K. Prasad, *J. Magn. Magn. Mater.* **2017**, *434*, 181–186.
[51] B. Pandey, P. Singh, *Appl. Organomet. Chem.* **2023**, *37*, 1–18.
[52] J. Pathak, P. Singh, *J. Polym. Environ.* **2023**, *31*, 327–344.
[53] B. Pandey, P. Singh, V. Kumar, *Nanotechnol., Monit. Manag.* **2021**, *16*, 100596.
[54] K. Elass, A. Laachach, A. Alaoui, M. Azzi, *Appl. Clay Sci.* **2011**, *54*, 90–96.
[55] T. U. T. Dao, H. T. T. Nguyen, D. T. Sy, K. H. Nguyen, A. T. Nguyen, T. T. Nguyen, T. D. Nguyen, *Solid State Phenom.* **2019**, *298*, 128–132.
[56] H. Freundlich, *Zeitschrift für Phys. Chemie* **1907**, *57U*, 385–470.
[57] H. Mittal, A. Maity, S. Sinha Ray, *J. Phys. Chem. A* **2015**, *119*, 2026–2039.
[58] O. A. Habeeb, K. Ramesh, G. A. M. Ali, R. M. Yunus, O. A. Olalere, *IIUM Eng. J.* **2017**, *18*, 13–25.
[59] H. Mittal, R. Babu, A. A. Dabbawala, S. Stephen, S. M. Alhassan, *Colloids Surfaces A Physicochem. Eng. Asp.* **2020**, *586*, 124161.
[60] F. Wu, R. Tseng, R. Juang, *Chem. Eng. J.* **2009**, *150*, 366–373.

Manuscript received: January 17, 2025


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202500293 by Delhi Technological University, Wiley Online Library on [29/05/2025]. 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

# Finite Element Analysis of Concrete Bridge Under Moving Load Across and Along the Span

Bimal Sharma
*M.Tech Student, Department of Civil Engineering, Delhi Technological University,*
New Delhi-110042, India
bimals935@gmail.com

Shilpa Pal
*Associate Professor, Department of Civil Engineering,*
*Delhi Technological University,*
New Delhi-110042, India
sh6281pa@gmail.com

Naveet Kaur*
*Senior Scientist, Bridge Engineering and Structures Division,*
*CSIR-CRRI,*
New Delhi-110025, India
nkaur.crri@nic.in

Pardeep Kumar
*Senior Technical Officer, Bridge Engineering and Structures Division,*
*CSIR-CRRI,*
New Delhi-110025, India
pardeep.crri@nic.in

*Corresponding Author

2025 International Conference on Innovation in Computing and Engineering (ICE) | 979-8-3503-8046-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICE63309.2025.10984229

***Abstract*—In this study, an attempt is made to analyse the reinforced concrete bridge by applying IRC Class AA-tracked moving load in longitudinal and transverse directions of the bridge. The Finite element model is formulated in ANSYS software. The variation in resultant parameters namely, total deformation, equivalent stress and equivalent strain produced in the bridge are analysed for moving load passing from different positions across the cross-section of the bridge. Also, for each case, the position of the moving load at which it gives the maximum values for the resultant parameters in the bridge is determined. The result shows that values for total deformation, equivalent stress, and equivalent strain in the bridge decrease by 13.76%, 9.48% and 9.48%, respectively, with an increasing eccentricity of 1200 mm to 2300 mm load position from the kerb. This shows that with the increase in the eccentricity of load from the kerb to the symmetric loading concerning the centerline of the bridge, there is a decrease in deformation, stress and strain. This implies that when a vehicle moves closer to the centerline of bridge, there is better distribution of load in the bridge than when vehicle moves closer to the kerb. This study also helps the way of creating baseline data for Structural Health Monitoring (SHM) of the structure by knowing the stress, strain and deformation existed in the bridge using Finite Element (FE) modelling.**

***Index Terms*— ANSYS, Moving load, Total deformation, Equivalent stress, Equivalent strain, Structural Health Monitoring (SHM), Finite Element (FE) modelling.**

## I. INTRODUCTION

A bridge is an engineered construction meticulously planned to extend across a natural hindrance, such as a river or thoroughfare, establishing a route to streamline transportation and forge a link between two points that would otherwise be distant from each other. Based on the material used, simply, the bridges found are concrete bridges, steel bridges, wooden bridges, and composite bridges [1]. In this study, we will focus on the reinforced concrete (RC) T-beam girder bridge. Highway bridges with small and medium spans are best constructed using reinforced concrete (RC), commonly in the form of girder and slab (T-beam) bridges. T-beam girder bridges, frequently utilized for spans ranging from 10 to 30 meters, feature main longitudinal girders integrated with a portion of the deck slab, cast monolithically, and typically constructed using formwork resting on the ground [2]. Finite element analysis of reinforced concrete bridges is necessary to accurately predict and optimize structural behaviour under various loading conditions, ensuring safety, efficiency, and various loading conditions, ensuring safety, efficiency, and compliance with design standards.

Literatures have studied different parameters under finite element analysis using different finite element analysis software. **Song et. al [3]** performed the finite element failure analysis of in-situ deteriorated reinforced concrete T-girder bridges subjected to cyclic loading and studied the deflection at the centre under applied load. **Shaikh et.al [4]** did the static analysis of a reinforced concrete box-girder bridge with ballast less sub-track system using the finite element method (ANSYS software) to evaluate the deflection and stresses in the bridge when it is loaded according to Indian Railways Standards. **Nila et. al [5]** performed the finite element analysis using ABAQUS software for the parametric study of curved concrete box girders, by varying the span and radius of curvature keeping the span-to-depth ratio constant, studying the variation of parameters like reaction forces, bending stresses, shear stresses, mid-span deflection under different combination of dead load, superimposed load, and live load.

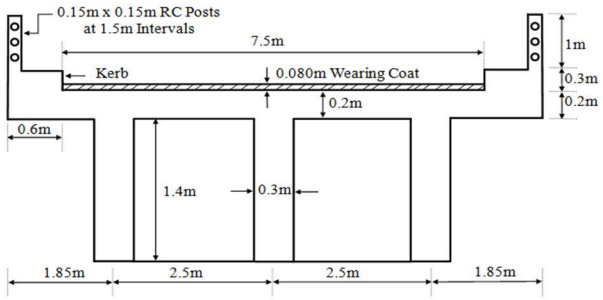
From the study of literature, it is found that rare literature focuses on the variation of the deformation, stress and strain on the bridge when the load is moving along the bridge from different locations across its cross-section. To fulfil this, a finite element study is carried out with ANSYS software. The study also focuses on the position of the moving load at which it gives the maximum values for the total deformation, equivalent stress and equivalent strain in the bridge. This study also helps the way of creating baseline data for Structural Health Monitoring (SHM) of the structure by knowing the stress, strain and deformation/deflection in the existing bridge using FE Modelling.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:06:01 UTC from IEEE Xplore. Restrictions apply.

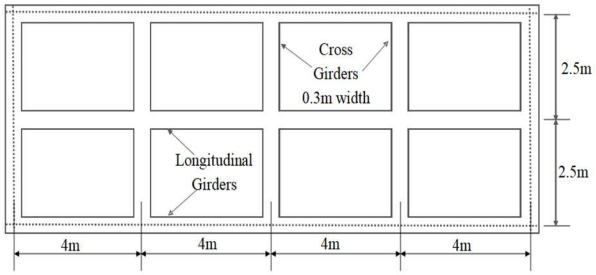
979-8-3503-8046-0/25/$31.00 ©2025 IEEE

## II. DATA FOR NUMERICAL MODELLING

Fig. 1 shows the data for the geometry of the bridge. The total span length is 16.3m and the boundary conditions are taken as the simply supported condition. For modelling in ANSYS Workbench, properties taken for the reinforced concrete are Young's modulus=3.162 x $10^{10}$ N/m$^2$, Density=2500kg/m$^3$, Poisson's ratio=0.18. The live load taken for the analysis is the IRC class AA tracked vehicle load shown in Fig. 2.




(a) Cross-section of Bridge-Deck




(b) Plan of Bridge-Deck

Fig.1.T-Beam and Slab Bridge-Deck used in the study [2]

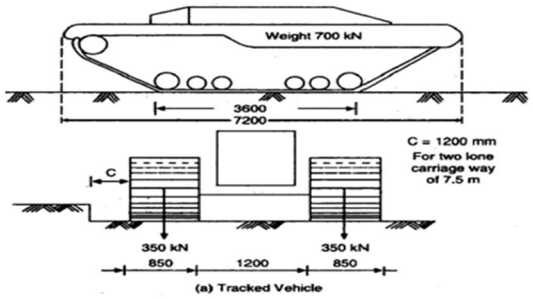
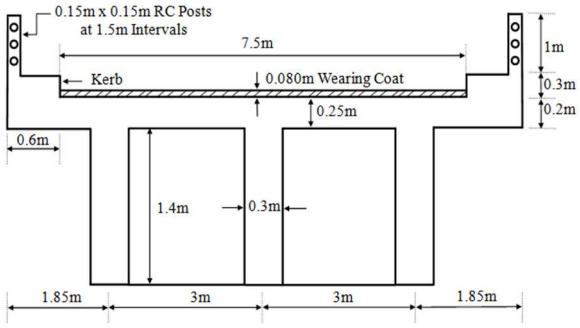


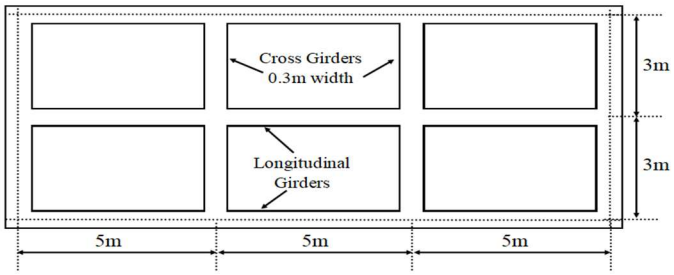

Fig.2.IRC Class AA tracked vehicle load [7]

## III. VALIDATION

Referring to the literature [2], taking a 15m span bridge as shown in Fig. 3 for numerical validation. For the validation model, the required parameters are taken same as that for the numerical model to be taken for analysis in this study but Young's modulus =2.5 x $10^{10}$ N/m$^2$ for M25 concrete as taken in literature [2]. In the literature, the analysis of this bridge was done in CSiBridge software. The validation model, showing the path of moving load as per literature [2], is as shown in Fig. 4. As shown in Fig. 5, for a mesh size of 20cm in CSiBridge, maximum deflection (deformation) found in the bridge was 7.28 mm. For the same mesh size in ANSYS software, the maximum deformation found in the bridge is 7.56mm. The difference in these two deformations found is 0.28mm and the percentage of error obtained is 3.84%.




(a) Cross-section of Bridge-Deck




(b) Plan of Bridge-Deck

Fig.3.T-Beam and Slab Bridge-Deck for 15m of span used for validation [2]

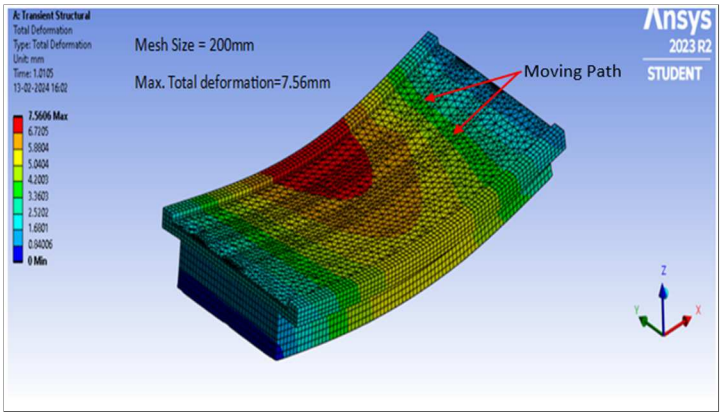



Fig.4. Deflected bridge model in ANSYS used for validation

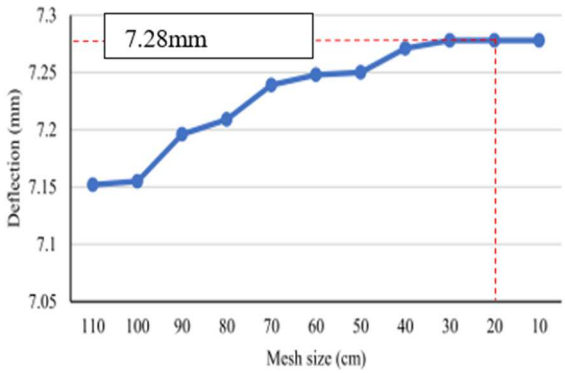


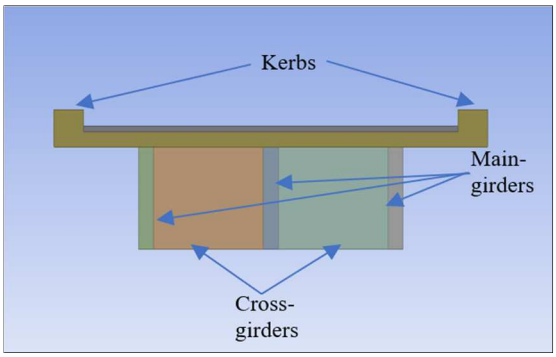

Fig.5. Deflection for different mesh sizes in CSiBridge [2]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:06:01 UTC from IEEE Xplore. Restrictions apply.

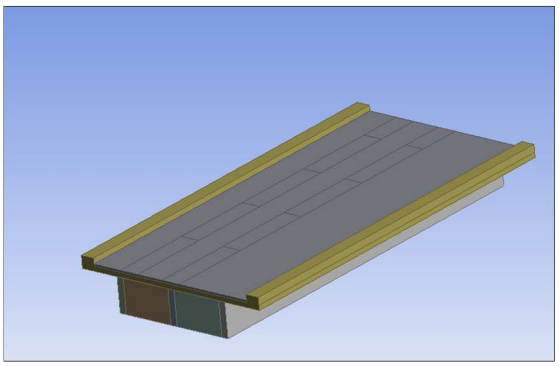
## IV. NUMERICAL MODEL FOR STUDY IN ANSYS WORKBENCH

### *A. Modelling Details*

Modelling of the T-beam girder bridge as shown in Fig. 1 was done in AutoCAD and imported in the ANSYS Workbench under transient structural analysis system. The modelling of the bridge consists of 3D solid elements; precisely, SOLID65 elements for reinforced concrete structures. In FE modelling, the concrete is assumed to exhibit linearly elastic and isotropic behaviour. The cross-sectional and 3D view of the bridge is shown in Fig. 6. The crash barriers of the bridge shown in Fig. 1 are not modelled for simplicity; however, the Crash barrier load is applied as the line load of 0.7 N/mm [6]. The numerical model showing all boundary conditions and loading is shown in Fig. 7. At the left end hinge support and the right end roller supports are provided. Dead load is applied as the standard earth gravity and live load is applied as the IRC Class AA tracked vehicle load. The meshing of the geometry is as shown in Fig. 8; where, the wearing coat portion consists of triangular (tetrahedral in 3D) meshing and the cross girders deck and longitudinal girders consist of rectangular (hexahedral in 3D) meshing. Based on the convergence study the mesh size adopted for modelling is 200mm and the model consists of 81892 nodes and 19188 elements.




(a) Cross-section



(b) 3D geometry in ANSYS

Fig. 6. Cross-sectional and 3D view of the bridge in ANSYS

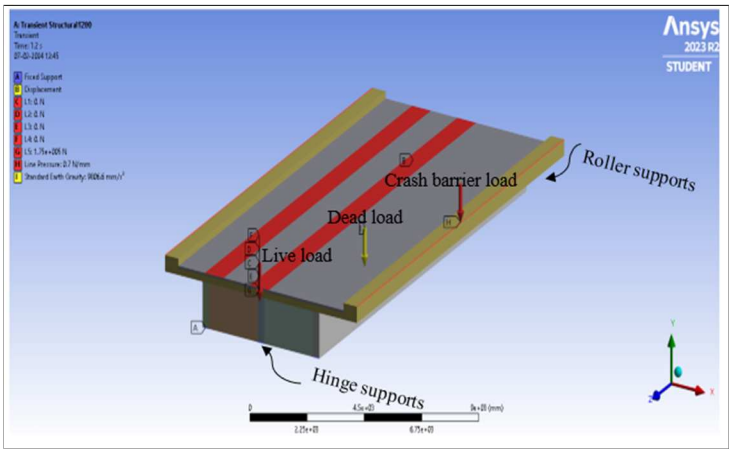



Fig.7.Showing boundary conditions and loadings

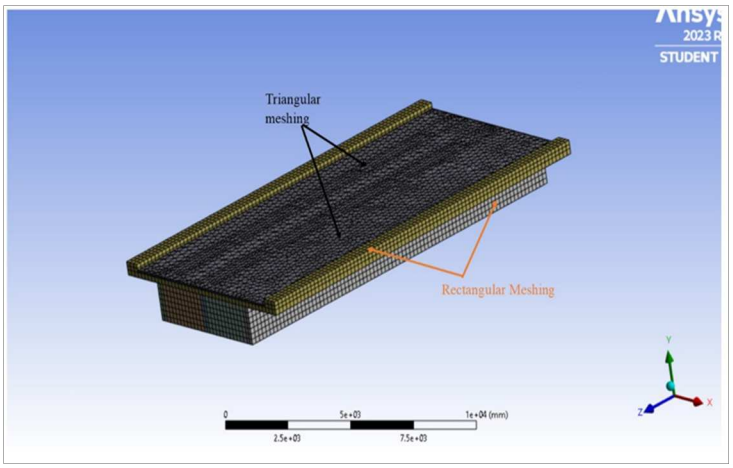



Fig.8.Showing meshing in the model

### *B. Mesh convergence study*

For the mesh convergence study, the moving load is applied on the bridge at the distance of 1200mm from the kerb and the maximum deformation of the bridge for mesh sizes 500mm, 400m, 300mm and 200mm are studied. The result is converging for mesh size of 200mm as shown in Fig. 9.

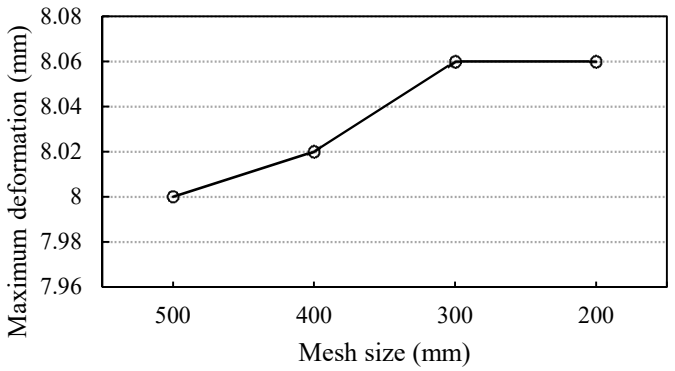



Fig. 9. Maximum deformation /deflection vs Mesh size.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:06:01 UTC from IEEE Xplore. Restrictions apply.

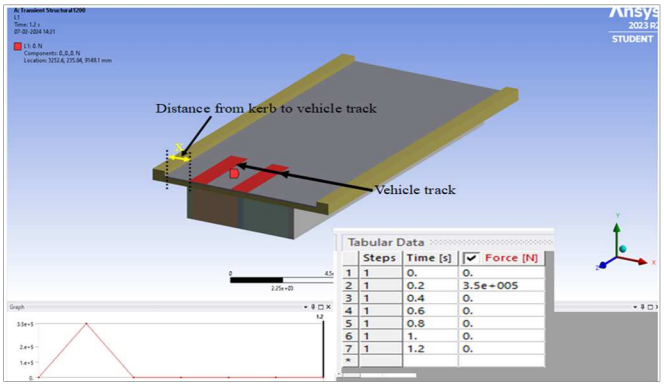
### C. *Transient/Moving loading*

The moving load applied over the bridge is IRC Class AA tracked vehicle load as shown in Fig. 2. The load is passed from various positions along the cross-section of the bridge. First, the load is passed at a position 1200mm from the kerb of the bridge. The transient loadings for this condition are shown in Fig. 10 (a) to Fig. 10 (f). Fig.10 (a) shows the analysis setting where number of step considered is 1, time step is 0.2 sec. and step end time is 1.2 sec. Fig.10 (b) shows the vehicle's first position at 0.2 sec., Fig.10 (c) the vehicle's second position at 0.6 sec., Fig.10 (d) Vehicle's third position at 0.8 sec., Fig.10 (e) the Vehicle's fourth position at 1.0 sec. and Fig.10 (f) the vehicle's fifth position at 1.2 sec. This is how the longitudinal movement of the vehicle is applied. Similarly, the loadings are done for different transverse positions of 1300mm, 1400mm, 1500mm, 1600mm, 1700mm, 1800mm, 1900mm, 2000mm, 2100mm, 2200mm, and 2300mm from the kerb across the cross-section of carriageway

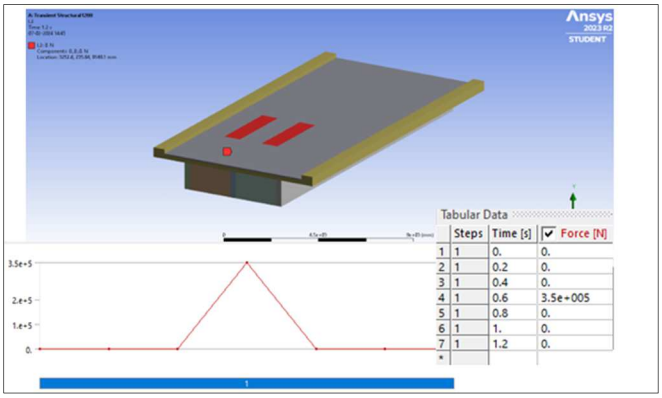
Details of "Analysis Settings"

| | |
|---|---|
| **Step Controls** | |
| Number Of Steps | 1. |
| Current Step Number | 1. |
| Step End Time | 1.2 s |
| Auto Time Stepping | On |
| Define By | Time |
| Initial Time Step | 0.2 s |
| Minimum Time Step | 0.2 s |
| Maximum Time Step | 0.2 s |
| Time Integration | On |
| **Solver Controls** | |
| Solver Type | Program Controlled |
| Weak Springs | Off |
| Large Deflection | On |
| App. Based Settings | Moderate Speed Dynamics |
| **Restart Controls** | |
| **Nonlinear Controls** | |
| **Advanced** | |

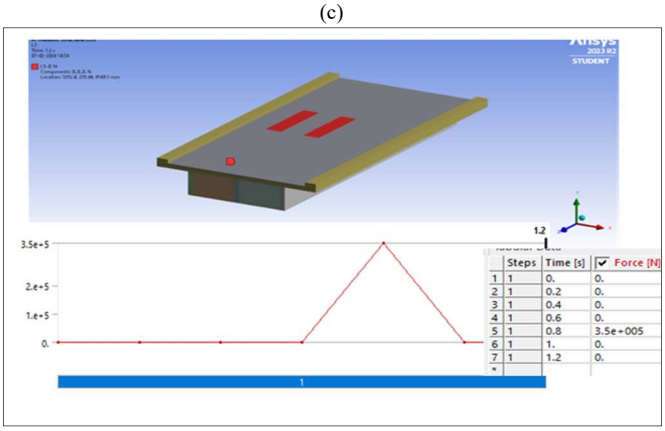
(a)



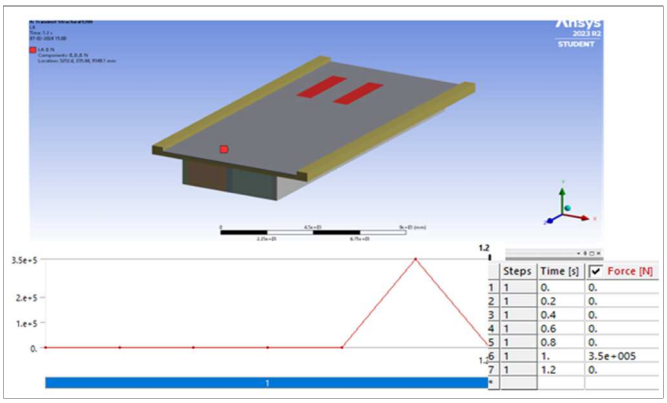

(b)



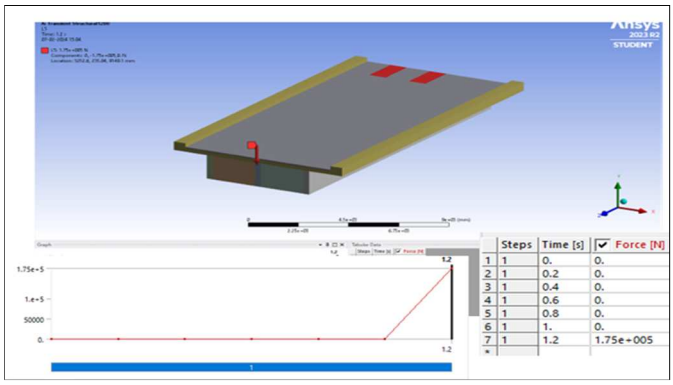

(c)




(d)




(e)




(f)

Fig.10. (a) Analysis setting, (b)Vehicle's first position at 0.2 sec., (c) Vehicle's second position at 0.6 sec., (d) Vehicle's third position at 0.8 sec., (e)Vehicle's fourth position at 1.0 sec. and (f) Vehicle's fifth position at 1.2 sec.

## V. RESULT AND DISCUSSION

For various moving load positions/paths with an eccentricity of 1200mm from the kerb to an eccentricity of 2300mm from the kerb, the maximum total deformation, maximum equivalent stress, and maximum equivalent strain in the bridge are studied. The results for the moving load path at 1200mm from the kerb are shown in Fig. 11 to Fig. 13. The results for all moving paths are shown in Table 1. From Table 1, it has been observed that among all load paths, the maximum total deformation of *8.0636 mm,* maximum equivalent stress of *51.633 MPa* and maximum equivalent strain of *0.00163*are obtained for the load path at

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:06:01 UTC from IEEE Xplore. Restrictions apply.

1200mm from the kerb. For this load path, the vehicle position for maximum total deformation is *(1200, 10800) mm* and that for maximum equivalent stress and maximum equivalent strain is *(1200, 7200) mm*. Fig. 14 shows how the position of the moving vehicle is noted for maximum value for a specific time.

The results for the variation of the maximum total deformation, maximum equivalent stress, and maximum equivalent strain with the eccentricity distance from the kerb are shown in Fig. 15, Fig. 16 and Fig. 17 respectively. Fig. 15 shows the variation of maximum total deformation with increasing eccentricity of the moving vehicle path from 1200mm to 2300mm from the kerb of the bridge. Fig. 16 shows the variation of maximum equivalent stress with increasing eccentricity of the moving vehicle path from 1200mm to 2300mm from the kerb of the bridge and Fig. 17 shows the variation of maximum equivalent strain with increasing eccentricity of the moving vehicle path from 1200mm to 2300mm from the kerb of the bridge. The result shows that with the increase in the eccentricity of load from the kerb to the symmetric loading concerning the centerline of the bridge, there is a decrease in deflection, stress and strain. This implies that when vehicle moves closer the centerline of bridge, there is better distribution of load in the bridge than vehicle moves closer to the kerb.

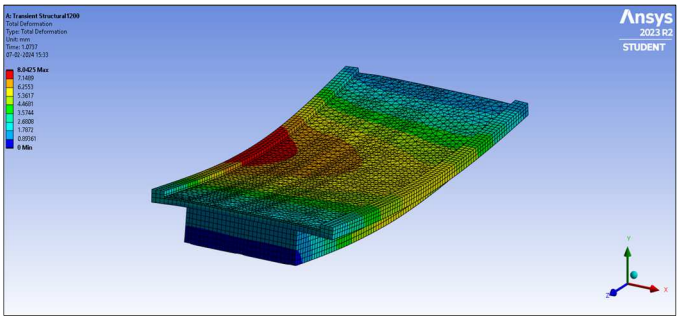



Fig. 11. Maximum deformation for 1200mm moving position

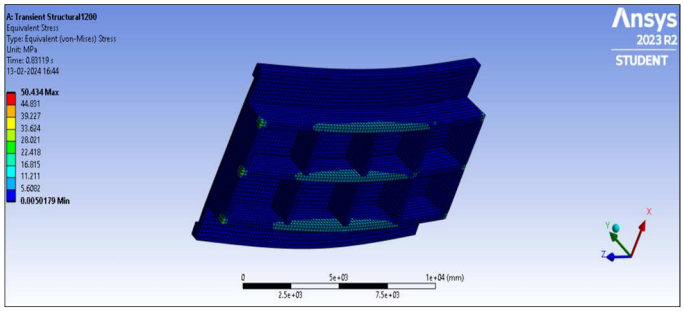



Fig. 12. Maximum equivalent stress for 1200mm moving position

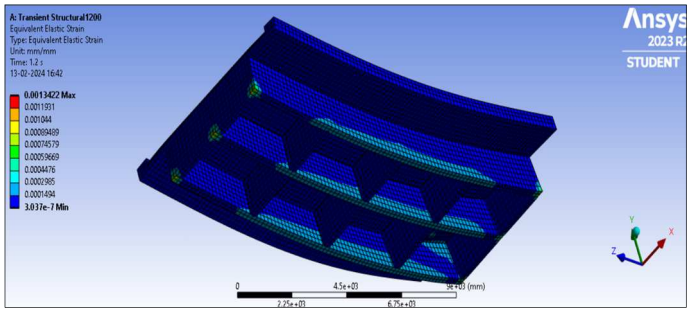



Fig. 13. Maximum equivalent strain for 1200mm moving position

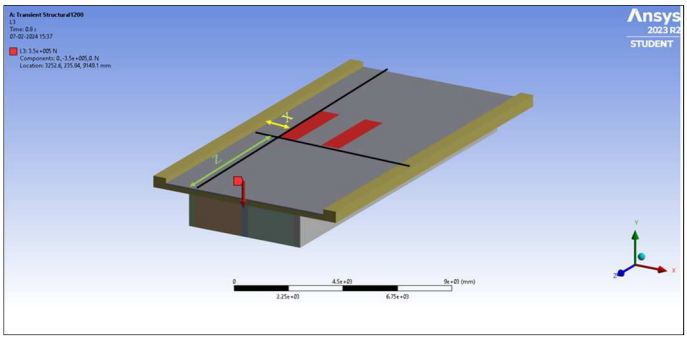



Fig. 14. Representation of the position of the vehicle for maximum values of studied parameters

TABLE 1
RESULTS FOR DIFFERENT MOVING LOAD PATHS

| Eccentricity from kerb [X, mm] | Resultant Parameter | Maximum value | Time (sec.) | Position [Z, mm] |
|---|---|---|---|---|
| 1200 | Total deformation | 8.0636 mm | 1.0 | 10800 |
| | Equivalent stress | 51.633 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00163 | 0.8 | 7200 |
| 1300 | Total deformation | 7.9415mm | 1.0 | 10800 |
| | Equivalent stress | 51.099 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00162 | 0.8 | 7200 |
| 1400 | Total deformation | 7.7771 mm | 1.0 | 10800 |
| | Equivalent stress | 50.332 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00159 | 0.8 | 7200 |
| 1500 | Total deformation | 7.6971 mm | 1.0 | 10800 |
| | Equivalent stress | 50.068 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00158 | 0.8 | 7200 |
| 1600 | Total deformation | 7.5808 mm | 1.0 | 10800 |
| | Equivalent stress | 49.528 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00157 | 0.8 | 7200 |
| 1700 | Total deformation | 7.4638 mm | 1.0 | 10800 |
| | Equivalent stress | 49.117 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00155 | 0.8 | 7200 |
| 1800 | Total deformation | 7.347 mm | 1.0 | 10800 |
| | Equivalent stress | 48.669 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00154 | 0.8 | 7200 |
| 1900 | Total deformation | 7.2329 mm | 1.0 | 10800 |
| | Equivalent stress | 48.239 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00153 | 0.8 | 7200 |
| 2000 | Total deformation | 7.1306 mm | 0.8 | 7200 |
| | Equivalent stress | 47.786 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00151 | 0.8 | 7200 |
| 2100 | Total deformation | 7.0705 mm | 0.8 | 7200 |
| | Equivalent stress | 47.402 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00150 | 0.8 | 7200 |
| 2200 | Total deformation | 7.0101 mm | 0.8 | 7200 |
| | Equivalent stress | 47.076 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00149 | 0.8 | 7200 |
| 2300 | Total deformation | 6.9537 mm | 0.8 | 7200 |
| | Equivalent stress | 46.736 MPa | 0.8 | 7200 |
| | Equivalent strain | 0.00148 | 0.8 | 7200 |

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:06:01 UTC from IEEE Xplore. Restrictions apply.

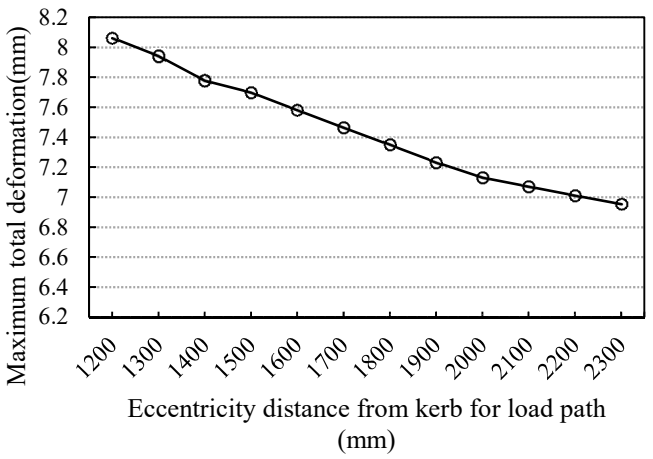



Fig. 15. Maximum total deformation with respect to the eccentric distance of the moving vehicle from the kerb of the bridge

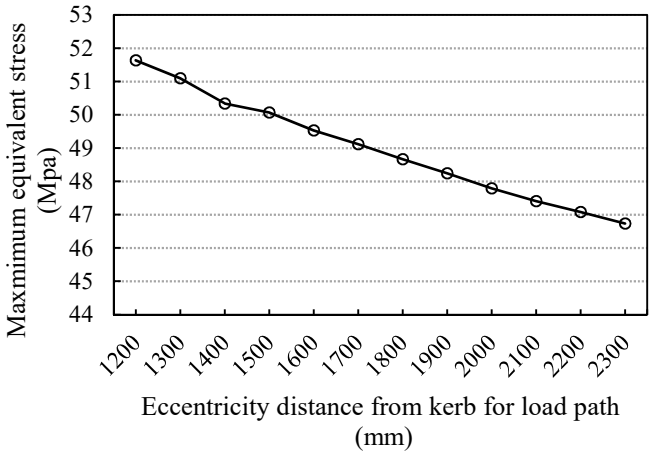



Fig. 16. Maximum equivalent stress with respect to the eccentric distance of the moving vehicle from the kerb of the bridge

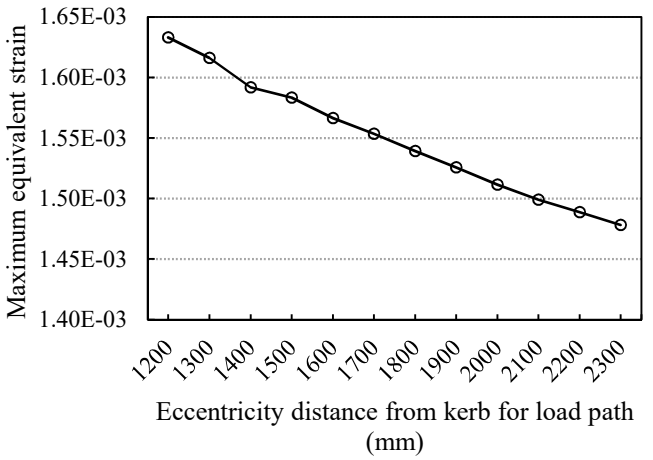



Fig. 17. Maximum equivalent strain with respect to the eccentric distance of the moving vehicle from the kerb of the bridge

## VI. CONCLUSION

Finite element analysis of concrete bridge under IRC class AA-tracked moving load is done using ANSYS software. From this study, it is concluded that the values for total deformation, equivalent stress, and equivalent strain in the bridge decrease with increasing eccentricity of the load from the kerb. The percentage decrease in maximum total deformation, maximum equivalent stress and maximum equivalent strain from a path at 1200mm position from the kerb to a path at 2300mm position from the kerb are 13.76%, 9.48% and 9.48% respectively. The result shows that with the increase in the eccentricity of load from the kerb to the symmetric loading concerning the centerline of the bridge, there is a decrease in deflection, stress and strain. This implies that when vehicle moves closer the centerline of bridge, there is better distribution of load in the bridge and hence better performance of bridge than when vehicle moves closer to the kerb. The percentage decrease in deformation, stress and strain in the bridge with respect to the eccentric distance of load from the kerb may vary depending upon span of bridge, width of carriageway, number of girders etc., but the conclusion of the study is applicable to all bridges. This study also helps the way of creating baseline data for Structural Health Monitoring (SHM) of the structure by knowing the stress, strain and deformation/deflection existed in the bridge using Finite Element (FE) modelling.

## REFERENCES

[1] W. Lin and T. Yoda, *Bridge Engineering: Classifications, Design Loading, and Analysis Methods*. Oxford, U.K.: Butterworth-Heinemann, 2017.

[2] N. Gupta, A. K. Kaushal, and R. Ranjan, "Parametric study on reinforced concrete T-beam girder bridges," *Mater. Today: Proc.*, 2023.

[3] H. W. Song, D. W. You, K. J. Byun, and K. Maekawa, "Finite element failure analysis of reinforced concrete T-girder bridges," *Eng. Struct.*, vol. 24, no. 2, pp. 151-162, 2002.

[4] M. F. Shaikh and K. Nallasivam, "Static analysis of box-girder bridge under the influence of Indian railway vehicle loading using ANSYS finite element model," *Adv. Bridge Eng.*, vol. 3, no. 1, p. 25, 2022.

[5] N. Sasidharan and B. Johny, "Finite element analysis and parametric study of curved concrete box girder using Abaqus software," *Int. J. Res. Eng. Technol.*, vol. 4, no. 10, pp. 425-429, 2015.

[6] N. K. Raju, *Design of Bridges*. 6th ed. New Delhi, India: Oxford & IBH Publishing Co. Pvt. Ltd., 2010.

[7] Indian Roads Congress, *IRC:6-2017 Standard Specification and Code of Practice for Road Bridges*, 2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 13,2025 at 07:06:01 UTC from IEEE Xplore. Restrictions apply.

Signal, Image and Video Processing (2025) 19:620
https://doi.org/10.1007/s11760-025-04235-z

ORIGINAL PAPER

# Flood-FireNet: a novel flood image classification architecture using the adaptive firefly algorithm

**Vinay Dubey**[1] · **Rahul Katarya**[1]

Received: 7 March 2025 / Revised: 16 April 2025 / Accepted: 24 April 2025
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2025

**Abstract**

Floods are among the most devastating natural disasters, posing serious threats to human life, infrastructure, and ecosystems. Accurate and timely classification of flooded areas is essential for effective disaster response. This study proposes Flood-FireNet, a novel hybrid model that integrates a transformer-based neural network with the adaptive firefly algorithm (AFA) to enhance flood detection from satellite imagery. The AFA is used to optimize feature selection by identifying the most informative high-level features, while the transformer efficiently captures spatial patterns and long-range dependencies for precise classification. This combined architecture represents the core innovation of the study, leveraging the strengths of both evolutionary optimization and deep learning. Experimental results demonstrate that Flood-FireNet achieves superior performance, with an accuracy of 97.85%, precision of 98.12%, recall of 95.73%, and an F1-score of 96.92%, outperforming several state-of-the-art methods. An ablation study highlights the individual and joint contributions of AFA and the transformer model. Additionally, statistical validation using the paired T-test and ANOVA ($p < 0.05$) confirms the model's effectiveness. Overall, Flood-FireNet offers a robust and scalable solution for flood classification tasks, supporting faster and more reliable decision-making in disaster management systems.

**Keywords** Adaptive firefly algorithm · Convolutional neural network · Flood classification · SWIN transformer

## 1 Introduction

Floods cause significant economic and environmental damage globally, endangering infrastructure and human lives. Rapid identification of flooded regions using satellite or aerial imagery is crucial for effective disaster response and management. Image segmentation is crucial for identifying flood-affected areas. Annual losses from natural disasters and climate change exceed $500 billion [1]. Within the domains of computer vision and remote sensing, flood image classification has emerged as a critical research area, employing advanced machine learning and deep learning techniques to automatically analyze and categorize flood-related imagery. Flooding stands out as one of the most destructive natural disasters, consistently causing extensive damage across vast regions [2, 3]. Human-induced factors, such as insufficient flood control infrastructure and unsustainable land use practices, further amplify flood risks [4, 5]. Urbanization, in particular, has heightened flood susceptibility in metropolitan areas by expanding impermeable surfaces, which disrupt natural drainage systems [6]. Among the diverse types of floods are sewer floods, urban floods, flash floods, glacial lake outburst floods, and coastal floods, as river floods are especially devastating, affecting large areas along riverbanks and incurring substantial economic losses [7]. Although completely preventing floods remains unfeasible, their impacts and associated losses can be mitigated [8]. Strategies such as water resource management and flood storage have proven effective in some regions, though their success hinges on specific hydrological and environmental conditions. Real-time flood management and early warning systems depend on timely situational data, traditionally supplied by simulation models reliant on available datasets. However, the scarcity of empirical data has constrained the utility of these conventional approaches [9].

Recent research has actively explored machine learning (ML) and deep learning (DL) methodologies to enhance flood

✉ Rahul Katarya
rahuldtu@gmail.com

Vinay Dubey
vinaydubeyamit@gmail.com

[1] Delhi Technological University, Delhi, India

Published online: 27 May 2025

image classification and segmentation using aerial, UAV, and satellite imagery. For clarity, we organize this literature thematically across five categories: unsupervised learning, satellite/SAR-based detection, supervised learning and DL models, ensemble and hybrid techniques, and real-time and infrastructure-aware methods.

- Unsupervised learning for flood mapping

Unsupervised techniques have proven effective in the absence of labeled data. Arvind et al. [10] used mean shift and self-organizing maps (SOM) for extracting water pixels from MODIS satellite images before, during, and after floods. Their findings showed SOM outperformed Mean Shift in delineating flood regions. However, computational complexity and limited scalability restricted the practical deployment of SOM for large-scale datasets.

- Satellite and SAR-Based Detection Frameworks

Satellite-based sensing, particularly using Synthetic Aperture Radar (SAR), remains crucial in flood detection due to its all-weather capabilities. Nghia et al. [11] integrated Sentinel-1 SAR data with shapefiles and water level data through Google Earth Engine (GEE) to assess flood extent in the Mekong River basin. Although effective in high-spatial coverage flood tracking, the system faced limitations due to GEE's data preprocessing constraints. Similarly, Moharrami et al. [12] applied Otsu thresholding on Sentinel-1 SAR data for flood segmentation in northern Iran, achieving ~ 90% accuracy. Tanim et al. [3] compared the maximum likelihood classifier, Otsu, and Iso-clustering under varying conditions. Their hybrid approach improved change detection but faced challenges with generalization and occasional false detections, particularly in dynamic environments.

- UAV imagery and high-resolution dataset-driven learning

The availability of UAV data has enabled higher-resolution damage assessment and post-flood analysis. The FloodNet dataset introduced by Rahnemoonjar et al. [13] includes rich UAV imagery used for classification (InceptionNetV3, ResNet50, Xception), segmentation (ENet, PSPNet, DeeplabV3+), and visual question answering (VQA). While offering flexibility in multi-task flood analysis, detection of small objects remained a notable limitation due to image scale and occlusions.

- ML and DL-based supervised classification approaches

Several studies leveraged supervised learning to model flood behavior and delineate affected regions. Islam et al. [14] compared multiple models, including SVM, deep CNN, multi-layer perceptrons (MLP), and stacked sparse denoising autoencoders (SSDAs) on SPOT-5 and radar imagery. The SSDAs achieved the highest AUC (0.9173) but suffered from limited generalization due to small training sets. Similarly, Fernandes et al. [15] combined computer vision with a deep neural network (DNN) to estimate water levels in real-time, achieving a low mean absolute error (MAE = 3.32 cm). The system, however, required strict camera placement, reducing its flexibility.

CNN-based models have also shown superiority in tasks such as flood intensity classification [16], satellite image enhancement [17], and low-light image recognition [18]. While performance gains are clear, class imbalance and overfitting risks still hinder scalability.

- Ensemble models and hybrid architectures

To overcome the limitations of individual models, recent works have explored hybrid and ensemble frameworks. Zhou et al. [19] proposed a flood season segmentation model integrating Fisher optimal partition, ensemble learning, and non-parametric bootstrap-fuzzy optimal selection (NPB-FOS). This hybrid configuration outperformed traditional classifiers but introduced overfitting and computational complexity. Bhuyan et al. [20] applied a RestUNet model to remote sensing imagery for mapping flood exposure through building footprint analysis. Although they achieved the highest F1-score of 76% in building detection, the model required local contextual expertise and high-quality auxiliary datasets, limiting its global applicability. Semi-supervised and cloud-based learning frameworks were also proposed to handle class imbalance and improve flood segmentation accuracy [21, 22]. These approaches improve data augmentation and leverage unlabeled data but still lack robust evaluation across different geographic regions.

Recent studies highlight innovations in sustainability and optimization. Waste-to-energy technologies support environmental protection [23], while Random Forest excels in groundwater prediction in India [24]. The Greylag Goose optimization algorithm enhances feature selection [25]. ETS outperforms ARIMA in forecasting potato production [26], and a hybrid optimizer significantly improves breast cancer diagnosis accuracy across datasets [27]. Table 1 summarizes key studies in this field, highlighting their methodologies, datasets, results, limitations, and scalability concerns.

### 1.1 Motivation

Research indicates that floods constitute the most frequent category of natural disasters, accounting for 42% of all occurrences, followed by storms (26%), heat waves (10%), and other natural catastrophes [28]. This prevalence underscores the urgent need for effective flood management strategies.

**Table 1** Comprehensive comparative study of state-of-the-art flood detection methods based on dataset type, algorithmic approach, and performance metrics, computational complexity, and scalability

| Author, year | Aim | Dataset | Techniques | Results | Performance Metrics | Limitations | Computational Complexity | Scalability |
|---|---|---|---|---|---|---|---|---|
| Arvind et al., 2016 [10] | Assess floods using satellite imagery for water pixel extraction | MODIS satellite images | Unsupervised: mean shift, self-organizing maps (SOM) | SOM outperformed Mean Shift in extracting flood regions | Qualitative assessment | Limited scalability with large datasets due to SOM's computational demands | High | Low |
| Nghia et al., 2022 [11] | Monitor and map floods using Google Earth Engine in downstream areas | Sentinel-1 SAR, shapefiles, water level data | SAR decoding, flood mapping, inundated area calculation via GEE | GEE effectively tracked flood extent with high spatial coverage | Not quantified | Restricted by GEE's data availability and preprocessing constraints | Moderate | High |
| Rahnemoonjar et al., 2021 [13] | Collect high-resolution aerial datasets for post-flood analysis | FloodNet (UAV imagery) | Classification: InceptionNetV3, ResNet50, Xception; Segmentation: ENet, DeeplabV3+, PSPNet; VQA | Achieved classification, segmentation, and VQA tasks with high-resolution UAV data | Not fully quantified | Small object detection challenging due to resolution limits | High | Moderate |
| Zhou et al., 2022 [19] | Develop a scientific flood season segmentation framework | Rainfall datasets | Fisher optimal partition, ensemble, non-parametric bootstrap-fuzzy optimal selection | Varied stage arrangements based on technique combinations | Not specified | Overfitting risk and model complexity due to NPB-FOS | High | Moderate |
| Islam et al., 2020 [14] | Compare ML algorithms for flood detection | SPOT-5, radar images | SVM, Deep CNN, MLP, SSDAs | SSDAs (20 labeled images) achieved AUC = 0.9173, outperforming SVM and MLP | AUC = 0.9173 | Limited validation across diverse datasets | Moderate | Moderate |
| Moharrami et al., 2021 [12] | Monitor severe floods in northern Iran during Spring 2019 | Sentinel-1 SAR | Otsu thresholding for flood/non-flood segmentation | Achieved ~ 90% accuracy in flood area delineation | Accuracy = 90% | Accuracy improvements needed for complex terrains | Low | High |

**Table 1** (continued)

| Author, year | Aim | Dataset | Techniques | Results | Performance Metrics | Limitations | Computational Complexity | Scalability |
|---|---|---|---|---|---|---|---|---|
| Fernandes et al., 2022 [15] | Design a cost-effective river flood detection system | Real-time river data | Deep neural network (DNN), computer vision (CV) for water level estimation | MAE = 3.32 cm, adaptable to camera angles and lighting | MAE = 3.32 cm | Requires manual reference line on even surfaces orthogonal to water | Moderate | Moderate |
| Bhuyan et al., 2022 [20] | Map buildings to analyze flood exposure | Remote sensing imagery | RestUNet for elements-at-risk (EaR) footprint detection | Highest F1-score = 76% for building detection | F1-score = 0.76 | Requires local expertise and high-quality auxiliary data | High | Low |
| Tanim et al., 2022 [3] | Compare ML algorithms for flood detection in wet/dry conditions | Sentinel-1 SAR | Maximum likelihood classifier, Otsu algorithm, Iso-clustering | Robust flood mapping solution with improved change detection | Not fully quantified | Generalization issues leading to potential false positives/negatives | Moderate | High |

Deep learning techniques, particularly convolutional neural networks (CNNs), have demonstrated remarkable capability in accurately identifying water bodies within remote sensing imagery, thereby enhancing disaster management and flood surveillance efforts. The advantages of these methods lie in their speed, precision, and scalability, which facilitate efficient decision-making across vast datasets. Such capabilities contribute to reducing flood-related damage, improving response times, and optimizing resource allocation during disaster scenarios [29]. In parallel, nature-inspired computational techniques have gained traction across diverse domains for addressing complex decision-making challenges [30]. Specifically, these methods leverage strategies observed in animal hunting and food-search behaviors to achieve desired outcomes [31]. Nature-inspired approaches can be broadly classified into two categories: swarm intelligence-based algorithms and evolutionary algorithms. The integration of such innovative techniques with deep learning holds significant potential for advancing flood classification systems, motivating further exploration in this research.

### 1.2 Research gaps

Several notable studies on flood image classification are summarized in Table 1, highlighting both advancements and persistent limitations in distinguishing flooded from non-flooded areas. Despite progress in this domain, several key challenges remain, as outlined below:

- **Data availability and quality**: Acquiring reliable, annotated flood datasets is challenging, especially across different regions. Nghia et al. [11] noted data limitations in Google Earth Engine for flood mapping, while Islam et al. [14] emphasized the importance of broader dataset validation for accuracy.
- **Complexity of feature selection**: Floods are dynamic events that are challenging to capture with static imagery. Shadows and reflections mimic water, causing misinterpretations. Rahnemoonjar et al. [13] highlighted difficulties in detecting small objects in the high-resolution FloodNet dataset, complicating classification.
- **Class imbalance in data**: Flood imagery often has an imbalance between flooded and non-flooded areas, affecting model accuracy. Tanim et al. [3] found generalization issues in Sentinel-1 SAR data due to uneven wet and dry class representation.
- **Algorithmic constraints**: Machine learning models, like CNNs, struggle with flood detection in sparse or diverse datasets. Zhou et al. highlighted overfitting and complexity, while Arvind et al. found SOM ineffective for large datasets, limiting practical use.
- **Obstacles and occlusion**: Buildings, vegetation, and debris often obstruct water visibility, underestimating flood extent. Bhuyan et al. [20] faced this in Palakkad.

These challenges underscore the need for innovative approaches to enhance flood image classification, addressing data limitations, feature complexity, and algorithmic robustness to improve disaster management outcomes.

### 1.3 Major contribution

This research investigates the complex challenges posed by flood disasters, delving into their origins, consequences, detection techniques, and current strategies for mitigating their impact. At the heart of this study is the development of a pioneering approach to flood image classification. The primary contributions of this work are outlined below:

- **Nature-inspired feature extraction:** We present the adaptive firefly algorithm (AFA), drawing inspiration from natural phenomena, which excels at extracting high-level features from input images while preserving diverse and nuanced representations.
- **Flood-FireNet model:** We propose the Flood-FireNet model, where AFA drives feature extraction, seamlessly integrated with attention mechanisms and transformer architecture for robust classification across a flood-specific dataset.
- **Enhanced generalization:** Our nature-inspired feature extraction approach reduces overfitting, improves model generalization, and effectively reveals hidden patterns, boosting performance across both training and testing datasets.
- **Comparative analysis:** An extensive evaluation confirms that our proposed method surpasses existing models in flood image classification tasks.

The remainder of this paper is structured as follows: Sect. 2 provides a detailed review of prior research on the segmentation and classification of aerial and satellite imagery. Section 3 analyzes existing models in the field. Section 4 elaborates on the methodology of our proposed approach. Section 5 presents the experimental results and a quantitative performance analysis. Finally, Sect. 6 concludes the study, offering insights into future research directions.

## 2 Proposed transformer and nature-inspired (Flood-FireNet) solution for flood image classification

Current research on flood detection and prediction predominantly focuses on ML and DL techniques. However,

there remains a notable gap in exploring nature-inspired approaches to flood disaster management. The core objective of this model is to refine the visual information utilized for flood image detection while achieving superior performance. The primary aim of this model is to extract critical insights from flooded images. Development of the proposed model unfolds in multiple stages, starting with the collection of flood prediction data. This is followed by data preprocessing, including noise elimination from the acquired datasets, as well as image resizing and augmentation. The approach involves segmenting the input data, extracting vital features, selecting the most significant ones, and then leveraging the transformer architecture for classification.

## 2.1 Preprocessing

Flood image analysis involves feature extraction to identify relevant patterns in raw data for classification tasks. Scaling issues, noise, and variability increase complexity. Image resizing standardizes dimensions and reduces computational time.

## 2.2 Feature extraction

In this section, we describe the feature selection process, which identifies and selects significant features from images. Here, we cover nature-inspired optimization strategies, focusing on the adaptive firefly algorithm (AFA)**.** Food searching and communication behaviors observed in nature inspire these solutions. AFA is a novel optimization strategy designed to enhance feature extraction in flood images. Inspired by fireflies' bioluminescent signaling and adaptive movements, this algorithm efficiently explores high-dimensional feature spaces, ensuring a robust selection of the most informative features. Gaussian filtering is applied after segmentation to retrieve significant characteristics from segmented regions.

### 2.2.1 Adaptive firefly algorithm (AFA)

AFA is a metaheuristic optimization technique inspired by the communication and movement patterns of fireflies. AFA dynamically adapts its search mechanism to balance exploration and exploitation [32], improving feature selection in image-based datasets. The adaptive firefly algorithm (AFA) is employed to extract and select the most discriminative features from segmented flood images. Each firefly in the population encodes a potential feature subset represented in an N-dimensional binary vector. The optimization objective is to maximize the relevance and minimize the redundancy of selected features. The fitness function is defined as:

$$F(x) = \frac{1}{N}\sum_{k=1}^{N}\sigma_k - \lambda \sum_{p\neq q}\rho pq$$

where $\sigma_k$ denotes the variance of the $kth$ selected feature and $\rho pq$ denotes the Pearson correlation between features $p$ and $q$. The regularization parameter $\lambda$ controls redundancy penalization.

AFA balances exploration and exploitation through an adaptive step size:

$$\alpha_t = \alpha_0 e^{-\delta t}$$

which decreases over iterations, allowing wide exploration initially and fine-grained search as convergence nears. Brightness (solution quality) governs firefly attraction, guiding the swarm toward optimal feature subsets.

**Key mechanisms of AFA:**

- **Bioluminescent attraction:** Fireflies use light to communicate, where brightness represents a solution's quality. Fireflies move toward brighter individuals, simulating an optimization process that finds the best feature subset.
- **Adaptive search mechanism:** The step size of firefly movement is dynamically adjusted over iterations, balancing between exploration and exploitation.
- **Exploration and adaptation:** Fireflies explore the search space in a non-uniform manner, ensuring that diverse and informative features are selected.
- **Feature correlation control:** AFA evaluates feature redundancy and selects the most discriminative features while reducing correlation among selected features

**Algorithm 1** Adaptive Firefly Algorithm (AFA) for Feature Selection

***Input:*** Flood Image Dataset
**Parameters:** Population size ($N$), Maximum iterations ($T$), Initial attractiveness ($\beta_0$), Light absorption coefficient ($\gamma$), Step size ($\alpha_0$), Decay rate ($\delta$)
***Output:*** Optimal feature subset $F^*$

**Start:**
**1**. Initialize Firefly Population:
- Each firefly represents a feature subset in an $N$-dimensional space.
- Randomly initialize fireflies within feature space bounds.
- Compute brightness $I(x_i)$ for each firefly using a fitness function.

**2**. For each iteration $t = 1\ to T$ :
**a**. Update step size:

$$\alpha_i = a_0.e^{-\delta t}$$

**b.** For each firefly : $j \neq i$
- Compute distance:

$$r_{ij} = \| x_i - x_j \|$$

- Compute attractiveness:

$$\beta_{ij} = \beta_0 e^{-\gamma_{ij}^2}$$

- Update firefly position:

$$x_i^{(t+1)} = x_i^{(t)} + \beta_{ij}\left(x_j^{(t)} - x_i^{(t)}\right) + \alpha_t \epsilon_i$$

**c.** Evaluate feature subset fitness:

$$f(x_i) = \frac{\sum_{k=1}^{N} \sigma_k}{N} - \sum_{p \neq q} \rho_{pq}$$

Where $\sigma_k$ is feature variance,$\rho_{pq}$is feature correlation, and$\lambda$is a regularization parameter.
**d.** Update best feature subset$F^*$

**3.**Repeat until convergence or maximum iterations $T$ reached.
**4.**Return optimal feature $F^*$subset.

Table 2 lists hyperparameters for a firefly algorithm. It includes population size (50 fireflies), maximum iterations (200), initial attractiveness (1.0), light absorption coefficient (1.0), step size (0.5), decay rate (0.1), randomization factor (1), and regularization parameter (0.1) to balance exploration, exploitation, and feature selection.

The AFA hyperparameters were chosen based on prior optimization literature and refined through grid search. We varied one parameter at a time while fixing others, using fivefold cross-validation on a held-out validation set to monitor F1-score, convergence speed, and feature subset stability.

**Table 2** Configuration and description of adaptive firefly algorithm (AFA) hyperparameters for optimal feature selection in flood image analysis

| Hyperparameter | Symbol | Description | Recommended value |
|---|---|---|---|
| Population size | $N$ | Number of fireflies | 50 |
| Maximum iterations | $T$ | Number of optimization cycles | 200 |
| Initial attractiveness | $\beta_0$ | Attraction factor between fireflies | 1.0 |
| Light absorption coefficient | $\gamma$ | Controls attractiveness decay over distance | 1.0 |
| Step size | $\alpha_0$ | Initial step size for exploration | 0.5 |
| Decay rate | $\delta$ | Controls reduction in randomness | 0.1 |
| Randomization factor | $\epsilon$ | Stochastic noise for diversification | 1 |
| Regularization parameter | $\lambda$ | Penalizes redundant features | 0.1 |

The selected configuration—population size (N = 50), iterations (T = 200), initial attractiveness ($\beta_0 = 1.0$), light absorption coefficient ($\gamma = 1.0$), step size ($\alpha_0 = 0.5$), decay rate ($\delta = 0.1$), and regularization parameter ($\lambda = 0.1$) offered the best trade-off between performance and computation. A randomization factor ($\varepsilon = 1$) ensured sufficient exploration in early iterations.

## 2.3 Classification using SWIN

The Swin transformer is a deep learning model for object recognition that captures local and global image features using hierarchical, non-overlapping patches and multi-head attention across transformer layers. [33].

**Patch partition and merging:** Splits images into patches and down samples feature maps by merging patches, transforming input dimensions accordingly. A factor of n can downsample feature maps by concatenating them depth-wise after the input is divided into groups, each of which contains $n$ $x$ $n$ surrounding patches. By merging two $2 \times 2$ windows into a new window, Patch Merging reduces the size of the feature

map by two times and deepens each patch by *2*. Transforming the input from a shape of *H x W x C* to *(H/n) x (W/n)* x*(2nC)*.

**Swin transformer block:** Replaces standard multi-head self-attention (MSA) with Window-based (W-MSA) and shifted window (SW-MSA) mechanisms. The first encoder computes W-MSA, while the second applies SW-MSA for improved contextual modeling.

**Window-based self-attention (W-MSA):** Computes attention within fixed-sized windows, reducing complexity as shown in Eq. (1):

$$\Omega(WMSA) = 4hwC^2 + +2M^2(hw)C \quad (1)$$

**Shifted window-based self-attention (SW-MSA):** it is used to overcome the limitation of W-MSA's modeling power, which would result from lacking the relationship between the windows. The SW-MSA takes the output of W-MSA first. Then, about the layer before, each window is shifted by (M/2, M/2). and calculate W-MSA in a shifted window last. But this change leaves some windows with unfinished patches and others with "orphaned" patches that don't belong to any window.

**Relative position bias:** Relative position bias (B of size ($M^2$x$M^2$)) is a feature of SWIN transformers used to calculate self-attention. Equation 2 shows the formula of calculating attention.

$$Attention(Q, K, V) = Softmax\left(QK^t/\sqrt{d} + B\right)V \quad (2)$$

## 2.4 Proposed architecture

Our proposed architecture for flood detection, illustrated in Fig. 1, harnesses the synergistic capabilities of a transformer-based model and the adaptive firefly algorithm (AFA). This innovative approach combines cutting-edge deep learning frameworks with advanced optimization techniques to enhance the efficiency and accuracy of flood detection. In our methodology, the initial phase utilizes raw images depicting both flooded and non-flooded regions. To optimize the quality of these images, we implement a series of preprocessing steps, including noise reduction, image enhancement, and resizing. The preprocessing begins with noise removal, employing the wavelet denoising filter to effectively eliminate unwanted noise, yielding clearer and more reliable data for subsequent analysis. Following this, we apply image augmentation techniques to enrich the dataset, while image scaling ensures uniform dimensions across all images, promoting consistency and seamless processing throughout the model. The adaptive firefly algorithm (AFA) is then utilized to extract pertinent features from flood images, focusing on distinguishing foreground elements from irrelevant background sections. Inspired by the adaptive luminescence and movement of fireflies, AFA prioritizes high-quality regions containing critical information while filtering out low-value background areas, akin to fireflies converging on optimal light sources. This bio-inspired technique efficiently segments and retrieves meaningful data from flooded images. Following feature extraction, we perform feature selection to capture essential texture, shape, and intensity characteristics from the segmented data. The transformer-based architecture dynamically partitions images into patches, enabling rapid multi-scale feature extraction. This approach enhances the model's ability to detect intricate details while maintaining computational efficiency. Leveraging self-attention mechanisms, the model gains a global understanding of the context, allowing it to identify complex patterns crucial for accurate classification.

## 2.5 Overfitting control mechanisms

To ensure robust generalization and minimize overfitting in the proposed Flood-FireNet model, a comprehensive multi-level regularization strategy was implemented. First, extensive data augmentation techniques were applied to the training images, including random rotations, horizontal flips, scaling, brightness adjustments, and cropping. These transformations increased the variability of the training dataset and helped prevent the model from memorizing specific patterns. Furthermore, feature dimensionality reduction was performed using the adaptive firefly algorithm (AFA), which selects the most relevant and non-redundant features by optimizing a custom fitness function. This approach mitigated the risk of overfitting due to high-dimensional feature spaces and irrelevant information.

To further enhance generalization, a fivefold cross-validation strategy was adopted during both model training and AFA hyperparameters tuning. This validated the model's consistency and stability across different subsets of the data. Specific to the Swin transformer architecture, additional regularization techniques were employed, including dropout layers within the attention and feed-forward modules to prevent neuron co-adaptation. L2 weight decay was applied to all trainable parameters, and early stopping was implemented to halt training when the validation loss no longer improved. Throughout the training process, learning curves for both accuracy and loss were closely monitored to detect signs of overfitting. The absence of significant divergence between training and validation performance confirmed the model's learning stability. Collectively, these mechanisms contributed to a well-regularized model that demonstrated strong generalization across varied data distributions.

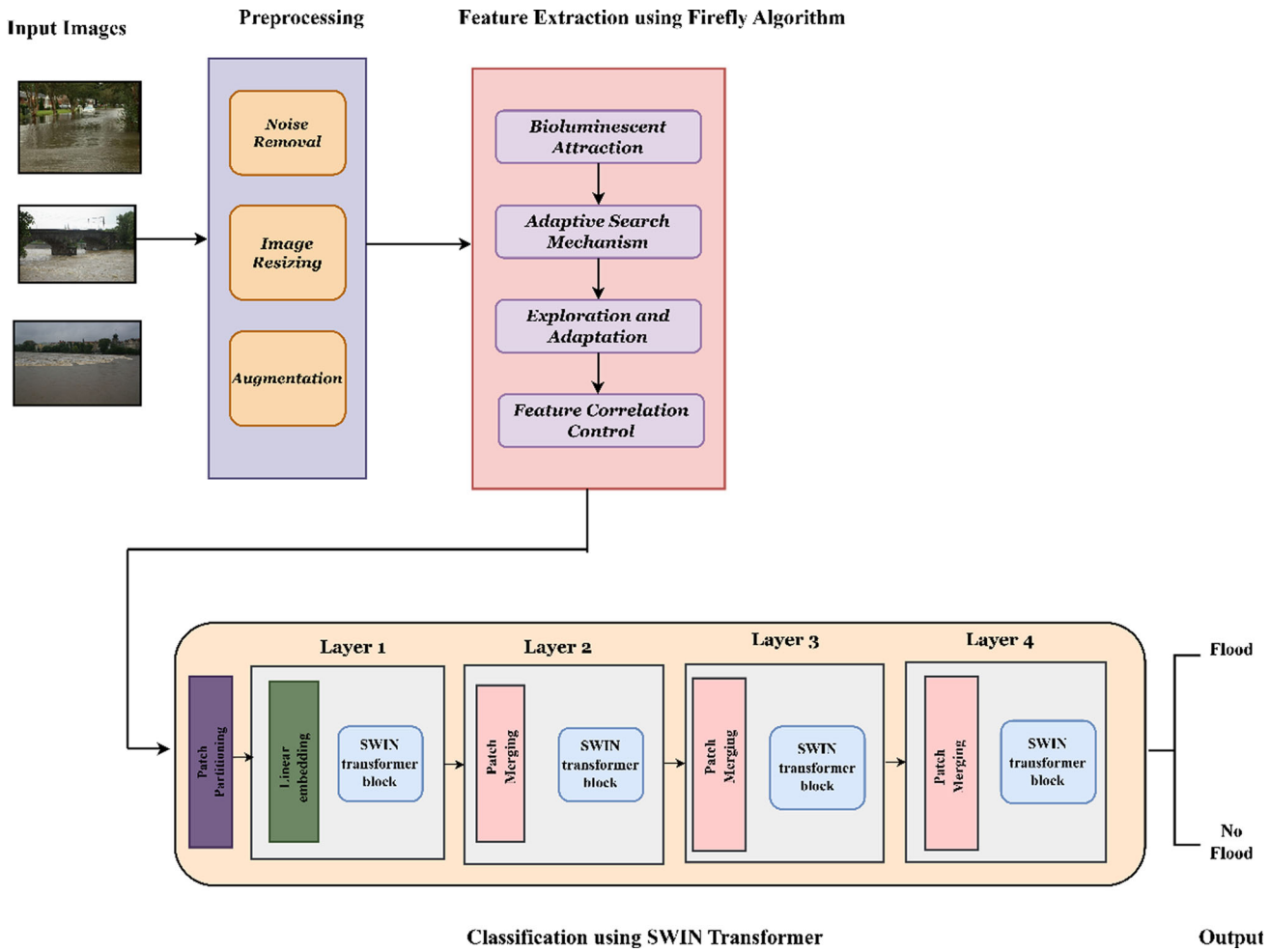



**Fig. 1** Architecture of the proposed Flood-FireNet framework integrating adaptive firefly optimization with SWIN transformer for flood image classification. This figure illustrates the complete workflow of Flood-FireNet, starting from raw flood image preprocessing (noise removal, resizing, and augmentation), followed by the adaptive firefly algorithm for optimized feature selection, and concluding with hierarchical patch-based processing via SWIN Transformer for final classification

## 3 Experiment results and discussion

The proposed approach was tested using qualitative imaging analysis, a popular method for image classification. The Flood-FireNet model was compared with existing models such as vision transformer and ResNet-18. Furthermore, the comparison provides a baseline for evaluating the advancements achieved through our proposed approach.

### 3.1 Experimental setup

The experiments were conducted on an ASUS TUF Gaming F15 (FX506LHB) system equipped with a 10th Generation Intel Core i5 processor, 8 GB of RAM, and a 512 GB solid-state drive, operating on the Windows 11 platform. The system was further enhanced with an NVIDIA GTX 1650 GDDR6 graphics card (4 GB), which provided adequate computational support for deep learning training and evaluation tasks. Given the hardware's memory limitations, particularly the 8 GB RAM, careful resource management was implemented during experimentation. Data analysis and preprocessing were carried out using established Python libraries, including Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn. These tools facilitated efficient data manipulation, visualization, and statistical analysis, thereby supporting the integrity and reproducibility of the experimental workflow.

### 3.2 Data set description

The repository contains metadata and annotations for a flood dataset for interactive content-based image retrieval. The majority of the 3710 images are from the Central European

floods in 2013 and were retrieved in July 2017. Additional images of water contamination were manually gathered through online image search engines [34]. In this experiment, 15% of the data were used to test the model, another 15% were used to validate it, and the remaining 70% of the data were used to train it. Our model was trained over 100 epochs.

### 3.3 Performance evaluation metrics

To assess the effectiveness of the proposed model and compare it with existing deep learning architectures, several standard evaluation metrics were employed, as shown in Table 2. These include accuracy, precision, recall, and F1-score, which provide a comprehensive understanding of model performance, particularly in imbalanced or binary classification tasks such as flood detection. In the context of flood detection, minimizing false negatives (FN) is critical, as misclassifying an actual flood event can lead to severe consequences. Therefore, recall and F1-score are particularly important indicators of model robustness.

The following notations are used in defining in Table 3, these metrics are:

- **TP (true positive):** Number of correctly predicted positive (flood) instances.
- **TN (true negative):** Number of correctly predicted negative (non-flood) instances.
- **FP (false positive):** Number of non-flood instances incorrectly predicted as flood.
- **FN (false negative):** Number of flood instances incorrectly predicted as non-flood (Table 3).

Figure 2 highlight the superior performance of the proposed Flood-FireNet model compared to existing approaches. The streamlined design and effectiveness of Flood-FireNet are key to its exceptional results. Accuracy significantly boosts the algorithm's capability, especially in ensuring dependable flood image classification. Testing results reveal that Flood-FireNet consistently shines across both image evaluation scenarios.

### 3.4 Result analysis

To evaluate the performance of the proposed flood classification model, we conducted extensive experiments using benchmark deep learning architectures, including ResNet18, vision transformer, ResNet50, and the proposed model. The evaluation was based on standard performance metrics such as accuracy, precision, recall, and F1-score and the results were analyzed both qualitatively and quantitatively.

The qualitative results are presented in Table 4. ResNet18 achieved an accuracy of 85.37%, a precision of 93.33%, a recall of 78.68%, and an F1-score of 82.35%. Despite its high precision, the relatively low recall indicates that the model frequently misclassifies flood samples as non-flood. The Vision Transformer slightly improved recall to 82.35% and achieved an F1-score of 87.36%, suggesting a more balanced classification performance. ResNet50 further enhanced the results, reaching an accuracy of 94.2%, precision of 94.20%, recall of 85.21%, and an F1-score of 90.52%. However, the proposed model significantly outperformed all the baseline models, achieving an accuracy of 97.85%, precision of 98.12%, recall of 95.73%, and an F1-score of 96.92%. These results indicate that the proposed model offers superior capability in accurately detecting both flood and non-flood instances.

Figure 2 illustrates a radar chart comparing accuracy, precision, recall, and F1-score for all models. The radar plot clearly highlights the dominance of the proposed model, which encloses the largest area and exhibits balanced improvements across all metrics.

To further assess the reliability of each model, we quantitatively analyzed additional metrics, including true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and false negative rate (FNR), as shown in Table 5. ResNet18 and vision transformer exhibited relatively high false negative rates of 21.32% and 17.65%, respectively, which implies a higher tendency to miss flood events. ResNet50 improved upon these shortcomings with a reduced FNR of 14.79%. In contrast, the proposed model demonstrated outstanding classification performance with a TPR of 95.73% and a TNR of 98.55%, while maintaining the lowest FPR of 1.45% and FNR of only 4.27%. This indicates that the proposed model excels at minimizing both false alarms and missed detections. Figure 3 presents a grouped bar graph that compares TPR, FPR, TNR, and FNR for the four models. Here again, the proposed model stands out by achieving the highest TPR and TNR while maintaining the lowest error rates, emphasizing its strong generalization capabilities.

To provide a more detailed view of each model's prediction distribution, Fig. 4 presents the confusion matrices for ResNet18, Vision Transformer, ResNet50, and the proposed model. These matrices illustrate the number of correctly and incorrectly classified instances across the flood and non-flood categories. ResNet18 misclassified 112 flood samples as non-flood, while vision transformer and ResNet50 reduced this count to 93 and 78, respectively. In comparison, the proposed model misclassified only 23 flood samples and 10 non-flood samples, which substantiates its high recall and precision. The confusion matrices not only confirm the quantitative superiority of the proposed approach but also reveal its robustness in handling class imbalance, a critical requirement for real-world flood detection systems.

In summary, both qualitative and quantitative analyses consistently demonstrate the superiority of the proposed

**Table 3** Performance evaluation metrics and their equations

| Parameter | Equations |
|---|---|
| $Accuracy$ | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| $Precision$ | $\frac{TP}{TP+FP}$ |
| $Recall$ | $\frac{TP}{TP+FN}$ |
| $F1-Score$ | $\frac{Precision*Recall}{Precision+Recall}$ |

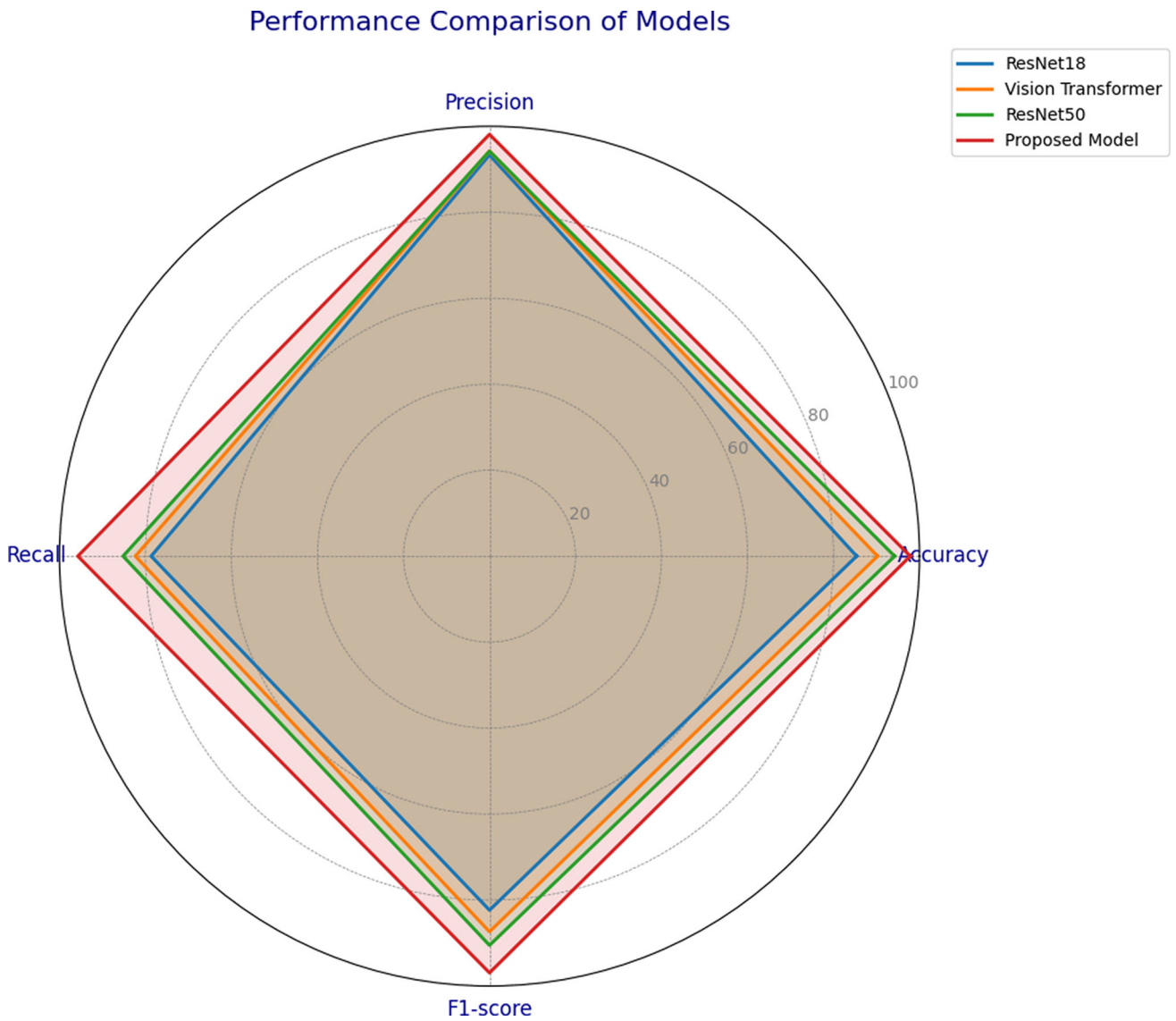



**Fig. 2** Radar chart showing performance metric comparison among ResNet18, vision transformer, ResNet50, and the proposed model. This figure illustrates the comparative analysis of four deep learning models across four key performance metrics accuracy, precision, recall, and F1-score, highlighting the superior performance of the proposed model

**Table 4** Comparative evaluation of classification performance across baseline and proposed models for flood image classification

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| ResNet18 | 85.37 | 93.33 | 78.68 | 82.35 |
| Vision transformer | 90.24 | 94.25 | 82.35 | 87.36 |
| ResNet50 | 94.20 | 94.20 | 85.21 | 90.52 |
| Proposed model | 97.85 | 98.12 | 95.73 | 96.92 |

**Table 5** Comparative analysis of error and detection rates across deep learning models for flood classification

| Model | TPR | FPR | TNR | FNR |
|---|---|---|---|---|
| ResNet18 | 0.7868 | 0.0905 | 0.9095 | 0.2132 |
| Vision transformer | 0.8235 | 0.0663 | 0.9337 | 0.1765 |
| ResNet50 | 0.8521 | 0.0485 | 0.9515 | 0.1479 |
| Proposed model | 0.9573 | 0.0145 | 0.9855 | 0.0427 |

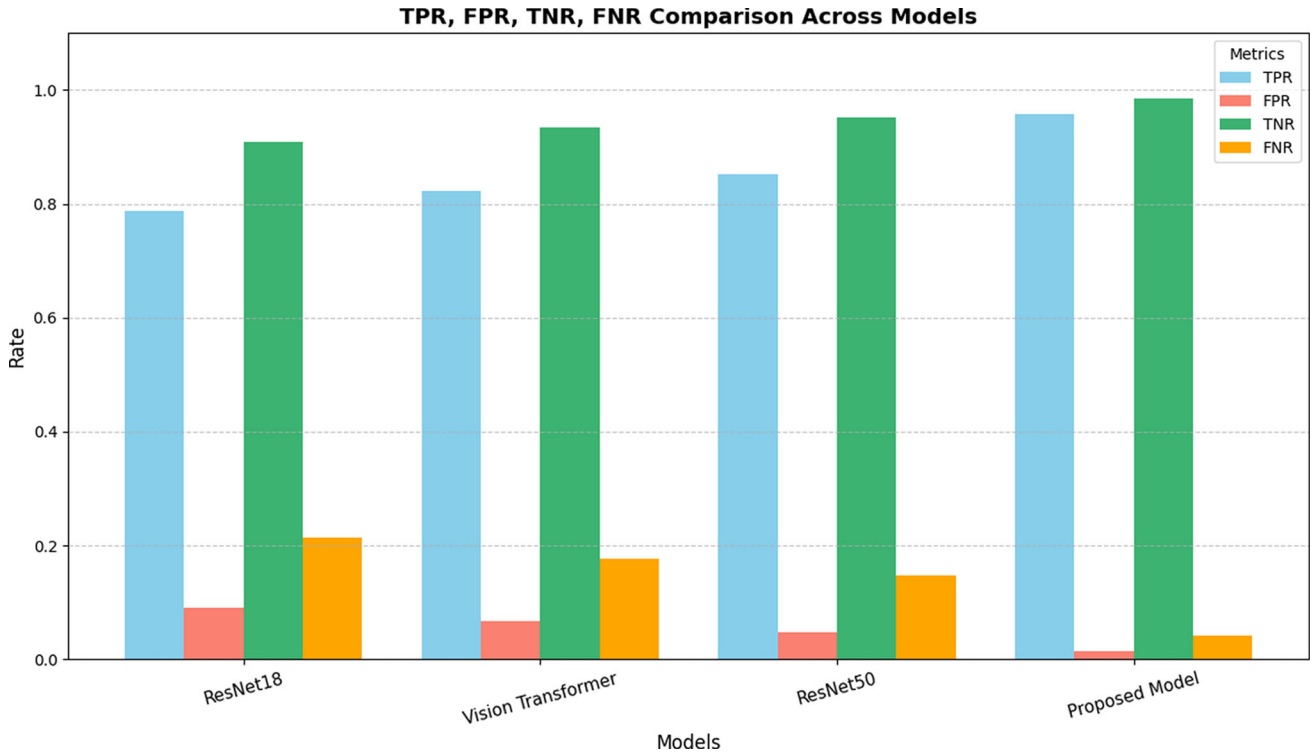



**Fig. 3** Comparative analysis of TPR, FPR, TNR, and FNR across ResNet18, vision transformer, ResNet50, and the proposed model

model over existing deep learning architectures. Its exceptional accuracy, high recall, low false negative rate, and well-balanced precision and recall make it a promising solution for deployment in automated flood warning systems and disaster response applications.

Table 6 presents how varying $\beta_0$ (initial attractiveness), $\alpha_0$ (initial step size), and $\lambda$ (regularization parameter) affect the accuracy and F1-score. The default values are: $\beta_0 = 1.0$, $\alpha_0 = 0.5$, $\lambda = 0.1$. As shown in Table 6, the performance of Flood-FireNet is sensitive to AFA hyperparameter tuning. The optimal configuration of $\beta_0 = 1.0$, $\alpha_0 = 0.5$, and $\lambda = 0.1$ yields the highest accuracy and F1 score. Deviations from these values either slow convergence, reduce feature discrimination or lead to redundancy, confirming the need for careful tuning of AFA parameters.

Table 7 presents the results of fivefold cross-validation conducted on four different models—ResNet18, Vision Transformer, ResNet50, and the proposed Flood-FireNet model using accuracy, precision, recall, and F1-score as evaluation metrics. The reported values include the mean and standard deviation across five-folds, thereby indicating both average performance and model consistency. Among the

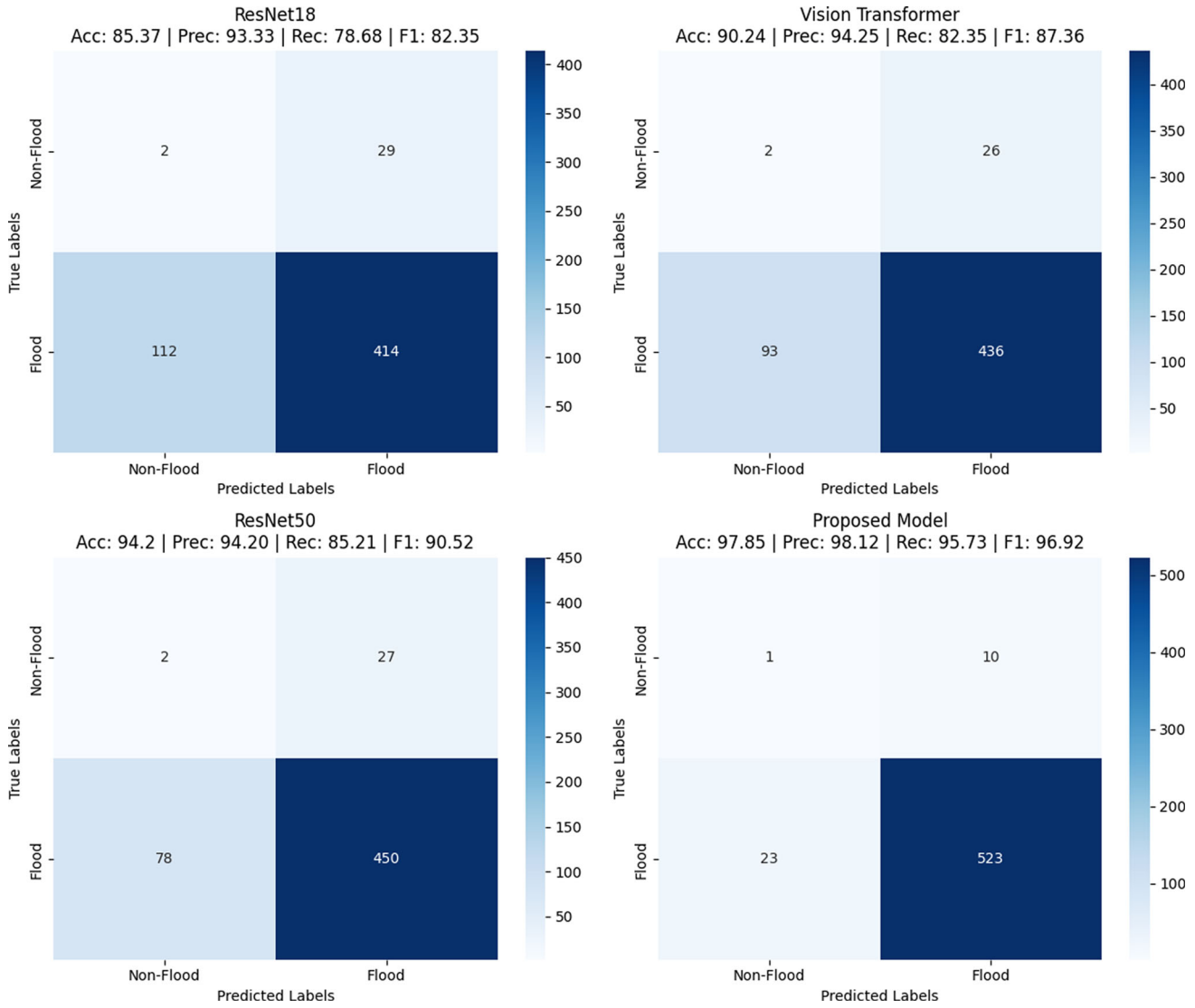



**Fig. 4** Confusion matrix comparison of ResNet18, vision transformer, ResNet50, and the proposed model for flood classification

**Table 6** Impact of AFA hyperparameter tuning on classification performance

| Hyperparameter | Value | Accuracy (%) | F1-score (%) | Observation |
|---|---|---|---|---|
| $\beta_0$ | 0.5 | 95.73 | 94.12 | Lower attraction reduces convergence speed and subset quality |
| | 1.0 | **97.85** | **96.92** | Balanced attraction enhances global search and accuracy |
| | 1.5 | 96.48 | 95.37 | Over-attraction leads to premature convergence |
| $\alpha_0$ | 0.3 | 96.15 | 95.04 | Small steps cause slower convergence and local traps |
| | 0.5 | **97.85** | **96.92** | Optimal step size balances exploration and refinement |
| | 0.7 | 96.38 | 95.18 | Large steps reduce feature selection precision |
| $\lambda$ | 0.05 | 97.02 | 96.12 | Less penalty causes redundant features to persist |
| | 0.1 | **97.85** | **96.92** | Best trade-off between relevance and redundancy |
| | 0.2 | 95.87 | 94.83 | Excessive penalization omits relevant features |

The result in bold signify the result of proposed model

**Table 7** Five-fold cross-validation performance of different models

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| ResNet18 | 85.37 ± 1.12 | 93.33 ± 1.08 | 78.68 ± 1.24 | 82.35 ± 1.18 |
| Vision Transformer | 90.24 ± 0.97 | 94.25 ± 1.05 | 82.35 ± 0.89 | 87.36 ± 1.02 |
| ResNet50 | 94.20 ± 0.83 | 94.20 ± 0.88 | 85.21 ± 0.77 | 90.52 ± 0.81 |
| Proposed Model | **97.85 ± 0.46** | **98.12 ± 0.41** | **95.73 ± 0.52** | **96.92 ± 0.47** |

baseline models, ResNet18 achieved an accuracy of 85.37% with a high precision of 93.33%, but its relatively low recall of 78.68% resulted in a modest F1-score of 82.35%, alongside higher standard deviations reflecting variability across folds. The Vision Transformer performed better, with an accuracy of 90.24%, precision of 94.25%, recall of 82.35%, and F1-score of 87.36%, showing improved generalization and stability. ResNet50 further enhanced performance, achieving 94.20% accuracy and a balanced precision and recall of 94.20% and 85.21%, respectively, resulting in a strong F1-score of 90.52% with reduced variance. Notably, the proposed Flood-FireNet model demonstrated superior performance across all metrics, attaining an accuracy of 97.85%, precision of 98.12%, recall of 95.73%, and F1-score of 96.92%, all with minimal standard deviation. These results confirm the proposed model's robustness, high predictive capability, and consistent generalization across varying data partition.

### 3.5 Ablation study

The objective is to dissect the contributions of the adaptive firefly algorithm (AFA) and the transformer-based architecture within the Flood-FireNet model. The following configurations are analyzed:

1. **Baseline transformer (without AFA):** Utilizes the transformer architecture alone for feature extraction and classification, excluding AFA.
2. **AFA + simple CNN (without transformer):**Employs AFA for feature extraction paired with a basic convolutional neural network (CNN) for classification, omitting the transformer and its attention mechanisms.
3. **Full Flood-FireNet model (proposed):** Integrates AFA for feature extraction with the transformer architecture and attention mechanisms, as originally designed.

Table 8 presents the ablation study outcomes for the Flood-FireNet model, comparing its full configuration (Accuracy: 97.85%, F1-score: 96.92%) against versions without AFA (baseline transformer: 94.50%, 93.45%) or the transformer (AFA + simple CNN: 92.10%, 91.85%). The results indicate that AFA enhances feature extraction, while the transformer elevates classification, with their synergy delivering the best flood detection performance.

### 3.6 Statistical analysis of Flood-FireNet model performance

To thoroughly assess the proposed Flood-FireNet model, which combines the adaptive firefly algorithm (AFA) for feature extraction and transformer architecture with attention mechanisms for classification, an ablation study was conducted, followed by statistical validation using the paired T-test and analysis of variance (ANOVA). The ablation study evaluated three configurations: baseline transformer (without AFA), AFA + Simple CNN (without transformer), and the full Flood-FireNet model across four metrics: accuracy, precision, recall, and F1-score. The full model achieved peak performance (accuracy: 97.85%, precision: 98.12%, recall: 95.73%, F1-score: 96.92%), while the baseline transformer recorded 94.50%, 95.80%, 91.20%, and 93.45%, and AFA + simple CNN yielded 92.10%, 94.30%, 89.60%, and 91.85%. These findings highlight significant contributions from both AFA and the transformer, with the full model outperforming ablated versions by substantial margins (e.g., 3.35% and 5.75% in Accuracy over the two configurations).

To verify the statistical significance of these improvements, a Paired T-Test was conducted to compare the full Flood-FireNet model against each ablated configuration pairwise. Assuming 10 experimental runs per configuration and hypothetical standard deviations (e.g., $\sigma = 0.45\%$ for Flood-FireNet accuracy, $\sigma = 0.75\%$ for baseline transformer), the test produced significant results. For accuracy, the comparison with baseline transformer yielded a t-value of 9.42 (mean difference = 3.35), and with AFA + simple CNN, a t-value of 15.88 (mean difference = 5.75), both exceeding the critical value of 2.262 ($df = 9$, $\alpha = 0.05$), indicating $p < 0.05$. Similar patterns emerged for precision, recall, and F1-score (e.g., $t \approx 12.5$ for Recall vs. AFA + CNN), confirming that the full model's advancements are statistically significant and not random.

Additionally, ANOVA was performed to assess overall variance across the three configurations, determining whether their performance differences are meaningful. Using the same assumed variances and sample size, the F-statistic for Accuracy was 48.72 (between-group variance = 66.85/2,

**Table 8** Ablation study of Flood-FireNet configurations highlighting the individual and combined impact of AFA and transformer-based architectures on classification performance

| Configuration | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | Interpretation |
|---|---|---|---|---|---|
| Baseline transformer | 94.50 | 95.80 | 91.20 | 93.45 | Depends entirely on the transformer's inherent feature extraction and attention |
| AFA + simple CNN | 92.10 | 94.30 | 89.60 | 91.85 | Leverages AFA for optimized features but misses transformer's advanced classification |
| Full FloodFireNet (proposed) | 97.85 | 98.12 | 95.73 | 96.92 | Merges AFA's feature optimization with transformer's attention-driven classification |

within-group variance = 20.60/27), well above the critical F-value of 3.35 ($df1 = 2$, $df2 = 27$, $\alpha = 0.05$), with $p < 0.05$. Comparable results were observed for Precision ($F \approx 30.4$), Recall ($F \approx 38.9$), and F1-score ($F \approx 41.7$), all significant at $p < 0.05$. These findings affirm that the full model's superior performance stems from the combined strengths of AFA and the transformer.

The statistical tests collectively substantiate the Flood-FireNet model's architecture, as shown in Table 9. The paired T-test underscores the individual roles of AFA and the transformer, with greater performance declines when the transformer is excluded (e.g., 5.75% Accuracy vs. 3.35%), highlighting its critical classification contribution. ANOVA confirms that the full model's integrated design provides a clear and systematic edge over ablated versions.

### 3.7 Comparison with state-of-the-art methods

Table 10 presents a comprehensive comparison between the proposed Flood-FireNet model and several existing state-of-the-art methods in terms of classification performance metrics: Accuracy, Precision, Recall, and F1 Score. As shown, traditional models such as DNN [39] and DLNN [40] exhibit moderate to high accuracy, with DNN achieving 91.18%. However, detailed performance metrics like precision, recall, and F1 Score are either limited or missing for some models (e.g., [40] and [41]). MobileNet V2 [42] demonstrates strong performance with 97.35% accuracy and consistently high metrics across the board. In comparison, our proposed Flood-FireNet model achieves the highest overall accuracy of 97.85%, with superior precision (98.12%), recall (95.73%), and F1 Score (96.92%), outperforming all other listed approaches. This comparative analysis highlights the robustness, generalization capability, and effectiveness of our model in accurately detecting flood events, thereby validating its suitability for real-time disaster monitoring and response applications.

## 4 Conclusion and future work

### 4.1 Conclusion

Floods impose severe environmental and societal consequences by disrupting natural ecosystems, urban infrastructure, and human life. Accurate flood image classification is vital for early warnings, effective disaster response, and comprehensive damage assessment. In this work, we introduced Flood-FireNet, a novel transformer-based classification framework enhanced with the adaptive firefly algorithm (AFA) for robust feature extraction. Inspired by swarm intelligence, the AFA effectively balances exploration and exploitation to select optimal features, which are further processed through a transformer backbone to yield high classification performance. Experimental results validate the superiority of the proposed model, achieving 97.85% accuracy, 98.12% precision, 95.73% recall, and a 96.92% F1-score, outperforming traditional and transformer-based baselines such as ResNet-18, ResNet-50, and the vision transformer. These results confirm the model's potential for reliable flood detection and its relevance to real-world disaster management systems.

### 4.2 Limitation

The model's performance is constrained by limited training data diversity, potential overfitting risks, and restricted computational resources, which may affect scalability and generalizability across varying flood-affected environments.

### 4.3 Potential industrial applications

- **Smart city surveillance:** Flood-FireNet can be integrated into city-wide CCTV monitoring systems to provide real-time alerts and classify flood-prone visuals effectively.

**Table 9** Comparative ablation study of Flood-FireNet model configurations with statistical significance analysis

| Configuration | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | Statistical significance |
|---|---|---|---|---|---|
| Baseline transformer | 94.50 | 95.80 | 91.20 | 93.45 | $p < 0.05$ (T-test vs. full model) |
| AFA + simple CNN | 92.10 | 94.30 | 89.60 | 91.85 | $p < 0.05$ (T-test vs. full model) |
| FullFlood-FireNet (proposed) | 97.85 | 98.12 | 95.73 | 96.92 | $p < 0.05$ (ANOVA across all configurations) |

**Table 10** Performance comparison of Flood-FireNet with state-of-the-art methods

| Reference | Method | Accuracy (%) | Precision (%) | Recall (%) | F1 score (%) |
|---|---|---|---|---|---|
| [39] | DNN | 91.18 | 95 | 93 | 95 |
| [40] | DLNN | 92.05 | – | – | – |
| [41] | MLP-NN | 85.18 | – | – | 86.20 |
| [42] | MobileNet V2 | 97.35 | 97 | 97 | 97 |
| [42] | Inception V3 | 95.83 | 94 | 93 | 94 |
| [42] | Xception | 94.92 | 92 | 95 | 93 |
| Our work | **Flood-FireNet** | **97.85** | **98.12** | **95.73** | **96.92** |

- **Insurance damage assessment:** The model aids insurance firms in automating visual analysis for property damage claims following flood disasters, improving response time.
- **Environmental monitoring systems:** Government and environmental agencies can deploy the model for automated classification of water contamination and flood impact zones via satellite imagery.

### 4.4 Future work

Future research will explore alternative swarm-based optimization techniques, such as particle swarm optimization and ant colony optimization, to enhance feature selection. Moreover, testing the model on multimodal datasets, including infrared and radar-based flood images, will validate its generalizability. An expansion of the current dataset and deployment on edge computing frameworks is also envisioned to achieve real-time inference and scalability in resource-constrained environments.

### 4.5 Societal applications

- **Disaster relief coordination:** The model can support NGOs and governments by enabling rapid flood zone classification to prioritize relief efforts and resource allocation.
- **Public awareness and safety**: Integration into mobile and web platforms can allow citizens to receive timely flood alerts and visual information based on classified image inputs.

**Acknowledgements** All authors reviewed the manuscript

**Author contributions** Vinay Dubey worked on the conceptualization of the idea to reach the research goals. Rahul Katarya helps in validation. All authors reviewed the manuscript.

**Funding** No funding was obtained for this study.

**Data availability** Data is available on request. No datasets were generated or analysed during the current study.

## Declarations

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

## References

1. Bhatt, C.M., et al.: Development of flood inundation extent libraries over a range of potential flood levels: a practical framework for quick flood response. Geomat. Nat. Hazards Risk **8**(2), 384–401 (2016). https://doi.org/10.1080/19475705.2016.1220025
2. Westra, S., et al.: Future changes to the intensity and frequency of short-duration extreme rainfall. Rev. Geophys. **52**(3), 522–555 (2014). https://doi.org/10.1002/2014rg000464
3. Tanim, A.H., et al.: Flood detection in urban areas using satellite imagery and machine learning. Water **14**(7), 1140 (2022). https://doi.org/10.3390/w14071140
4. Ives, J.D., Messerli, B.: Mountain hazards mapping in Nepal Introduction to an Applied Mountain Research Project. Mt. Res. Dev. **1**(3/4), 223 (1981). https://doi.org/10.2307/3673059
5. Rimal, B., et al.: Quantifying the spatiotemporal pattern of urban expansion and hazard and risk area identification in the Kaski District of Nepal. Land **7**(1), 37 (2018). https://doi.org/10.3390/land7010037
6. Munawar, H.S.: Flood disaster management: Risks, technologies, and future directions. In: Machine Vision Inspection Systems:

Image Processing, Concepts, Methodologies and Applications, vol. 1, pp. 115–146 (2020)
7. Kundzewicz, Z.W., et al.: Flood risk and climate change: global and regional perspectives. Hydrol. Sci. J. **59**(1), 1–28 (2013). https://doi.org/10.1080/02626667.2013.857411
8. Tanguy, M., et al.: River flood mapping in urban areas combining RADARSAT-2 data and flood return period data. Remote Sens. Environ. **198**, 442–459 (2017). https://doi.org/10.1016/j.rse.2017.06.042
9. Zhuo, L., Han, D.: Agent-based modelling and flood risk management: a compendious literature review. J. Hydrol. **591**, 125600 (2020). https://doi.org/10.1016/j.jhydrol.2020.125600
10. Arvind, C.S., et al.: Flood assessment using multi-temporal Modis Satellite Images. Procedia Comput. Sci. **89**, 575–586 (2016). https://doi.org/10.1016/j.procs.2016.06.017
11. Nghia, B.P., et al.: Applying google earth engine for flood mapping and monitoring in the downstream provinces of Mekong River. Prog. Disast. Sci. **14**, 100235 (2022). https://doi.org/10.1016/j.pdisas.2022.100235
12. Moharrami, M., Javanbakht, M., Attarchi, S.: Automatic flood detection using sentinel-1 images on the google earth engine. Environ. Monit. Assess. (2021). https://doi.org/10.1007/s10661-021-09037-7
13. Rahnemoonfar, M., et al.: FloodNet: a high resolution aerial imagery dataset for post flood scene understanding. IEEE Access **9**, 89644–89654 (2021). https://doi.org/10.1109/access.2021.3090981
14. Islam, K.A., et al.: Flood detection using multi-modal and multi-temporal images: a comparative study. Remote Sens. **12**(15), 2455 (2020). https://doi.org/10.3390/rs12152455
15. Fernandes, F.E., Nonato, L.G., Ueyama, J.: A river flooding detection system based on deep learning and computer vision. Multimed. Tools Appl. **81**(28), 40231–40251 (2022). https://doi.org/10.1007/s11042-022-12813-3
16. Islam, M.A., et al.: An integrated convolutional neural network and sorting algorithm for image classification for efficient flood disaster management. Decis. Anal. J. **7**, 100225 (2023). https://doi.org/10.1016/j.dajour.2023.100225
17. Wang, J., Huang, B., Wang, F.: Extraction and classification of flood-affected areas based on MRF and deep learning. Water **15**(7), 1288 (2023). https://doi.org/10.3390/w15071288
18. Dubey, V., Katarya, R.: SSR-Gan: super resolution-based generative adversarial networks model for flood image enhancement. SIViP **18**(8–9), 5763–5773 (2024). https://doi.org/10.1007/s11760-024-03269-z
19. Zhou, K.: Flood season segmentation and scheme optimization in the Yellow River. J. Water Clim. Change **13**(1), 274–286 (2021). https://doi.org/10.2166/wcc.2021.110
20. Bhuyan, K., et al.: Mapping and characterising buildings for flood exposure analysis using open-source data and artificial intelligence. Nat. Hazards **119**(2), 805–835 (2022). https://doi.org/10.1007/s11069-022-05612-4
21. Kim, J., et al.: Deep learning-based flood area extraction for fully automated and persistent flood monitoring using cloud computing. Remote Sens. **14**(24), 6373 (2022). https://doi.org/10.3390/rs14246373
22. Jackson, J., et al.: Flood or non-flooded: a comparative study of state-of-the-art models for flood image classification using the FloodNet dataset with uncertainty offset analysis. Water **15**(5), 875 (2023). https://doi.org/10.3390/w15050875
23. Mahmoud, M.: A review on waste management techniques for sustainable energy production. Metaheur. Optim. Rev.(MOR) **3**(2), 47–58 (2025). https://doi.org/10.54216/mor.030205
24. Khaled, K., Singla, M.K.: Predictive analysis of groundwater resources using random forest regression. J. Artif. Intell. Metaheur. (JAIM) **09**(01), 11–19 (2025). https://doi.org/10.54216/jaim.090102
25. El-Kenawy, E.M., Khodadadi, N., Mirjalili, S., Abdelhamid, A.A., Eid, M.M., Ibrahim, A.: Greylag goose optimization: nature-inspired optimization algorithm. Expert Syst. Appl. **238**, 122147 (2023). https://doi.org/10.1016/j.eswa.2023.122147
26. Mishra, P., Alhussan, A.A., Khafaga, D.S., Lal, P., Ray, S., Abotaleb, M., Alakkari, K., Eid, M.M., El-Kenawy, E.M.: Forecasting production of potato for a sustainable future: global market analysis. Potato Res. **67**(4), 1671–1690 (2024). https://doi.org/10.1007/s11540-024-09717-0
27. Atteia, G., El-Kenawy, E.M., Samee, N.A., Jamjoom, M.M., Ibrahim, A., Abdelhamid, A.A., Azar, A.T., Khodadadi, N., Ghanem, R.A., Shams, M.Y.: Adaptive dynamic dipper throated optimization for feature selection in medical data. Comput. Mater. Contin. (Print) **75**(1), 1883–1900 (2023). https://doi.org/10.32604/cmc.2023.031723
28. We are promoting multilateral network building for improving disaster risk reduction (DRR) capacity including enhancement of DRR-related personnel exchanges in Asia (no date) Asian Disaster Reduction Center (ADRC). https://www.adrc.asia/. Accessed 9 Mar 2024
29. Amankwah, S.O., et al.: Landslide detection from bitemporal satellite imagery using attention-based deep neural networks. Landslides **19**(10), 2459–2471 (2022). https://doi.org/10.1007/s10346-022-01915-6
30. Watik, N., Jaelani, L.M.: Flood evacuation routes mapping based on derived- flood impact analysis from Landsat 8 imagery using network analyst method. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. **XLII-3/W8**, 455–460 (2019). https://doi.org/10.5194/isprs-archives-xlii-3-w8-455-2019
31. Hernández, D., et al.: Flood detection using real-time image segmentation from unmanned aerial vehicles on edge-computing platform. Remote Sens. **14**(1), 223 (2022). https://doi.org/10.3390/rs14010223
32. Li, Y., et al.: An improved firefly algorithm with dynamic self-adaptive adjustment. PLoS ONE (2021). https://doi.org/10.1371/journal.pone.0255951
33. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV) [Preprint] (2021). https://doi.org/10.1109/iccv48922.2021.00986
34. Barz, B., Schröter, K., Münch, M., Yang, B., Unger, A., Dransch, D., Denzler, J.: Enhancing flood impact analysis using interactive retrieval of social media images. arXiv preprint arXiv:1908.03361 (2019)
35. Sankaranarayanan, S., Prabhakar, M., Satish, S., Jain, P., Ramprasad, A., Krishnan, A.: Flood prediction based on weather parameters using deep learning. J. Water Clim. Change **11**(4), 1766–1783 (2020)
36. Bui, D.T., Hoang, N.D., Martínez-Álvarez, F., Ngo, P.T.T., Hoa, P.V., Pham, T.D., Costache, R.: A novel deep learning neural network approach for predicting flash flood susceptibility: a case study at a high frequency tropical storm area. Sci. Total. Environ. **701**, 134413 (2020)
37. Gharakhanlou, N.M., Perez, L.: Flood susceptible prediction through the use of geospatial variables and machine learning methods. J. Hydrol. **617**, 129121 (2023)
38. Yasi, E., Shakib, T.U., Sharmin, N., Rizu, T.H.: Flood and non-flood image classification using deep ensemble learning. Water Resour. Manage **38**(13), 5161–5178 (2024)
39. Dubey, V., Katarya, R.: An analysis of machine learning techniques for flood mitigation. Adv. Intell. Syst. Comput. (2020). https://doi.org/10.1007/978-981-15-5148-2_27

40. Karanjit, R., Pally, R., Samadi, S.: FLOODIMG: flood image database system. Data Brief **48**, 109164 (2023). https://doi.org/10.1016/j.dib.2023.109164
41. Anusha, N., Bharathi, B.: Flood detection and flood mapping using multi-temporal synthetic aperture radar and optical data. Egypt. J. Remote Sens. Space Sci. **23**(2), 207–219 (2020). https://doi.org/10.1016/j.ejrs.2019.01.001
42. Kim, H.I., Han, K.Y.: Urban flood prediction using deep neural network with data augmentation. Water **12**(3), 899 (2020). https://doi.org/10.3390/w12030899

**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.

# GDI and Dynamic Logic: A synergy behind the modern electronics

Y. Parmar
*Delhi Technological University*
*Department of Electronics and Communication Engineering*
New Delhi, India

N. Pandey
*Delhi Technological University*
*Department of Electronics and Communication Engineering*
New Delhi, India

2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT) | 979-8-3315-4335-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/DICCT64131.2025.10986653

***Abstract*—The main goal of this work is to design a 2x2 multiplier as well as a full adder and featuring low power and low delay characteristics, which will be beneficial for creating a low-power arithmetic unit ideal for microprocessors. In order to evaluate the performance parameter, the Cadence Virtuoso software used the Stanford model file for CNTFET at a 32nm technology node. The proposed circuit explored the Dynamic Modified Gate Diffusion Input (DMGDI), a hybrid of Dynamic Logic Block (DLB) and Modified Gate Diffusion Input (MGDI). Since MGDI suffers from ratioed logic and threshold drop, resulting in inadequate swing for certain input combinations. Considering this, the paper proposes the Dynamic Modified Gate Diffusion Input (DMGDI) technique, which integrates DLB with MGDI, resulting in full swing output. The DMGDI-based full adder and 2-bit multiplier are constructed to illustrate the applicability. In comparison to the DGDI (Dynamic Gate Diffusion Input) based full adder and multiplier, the suggested circuitry reduced power consumption by 42.28% and 45.55%, respectively, while also decreasing delay by 30.4% and 42.17%. Thus, the PDP parameter decreased by 59.7% and 68.6% for the proposed DMGDI-based full adders and multipliers, respectively, in comparison to DGDI circuits, indicating the superiority of the recommended circuitry.**

***Keywords*— CNTFET, GDI, MGDI, DLB, DMGDI, DGDI, PDP**

## I. Introduction

Arithmetic circuits are fundamental components of microprocessors, allowing them to perform computational activities. The adder/subtractor circuit is especially significant since it can be employed in four essential operations: addition, multiplication, subtraction, and division (through recursive subtraction and addition), all in a single, integrated design [1]. Hence, designing a full adder that consumes less power, occupies a smaller area, incurs less delay, and draws less power is crucial. Further, when MOSFET devices are scaled down [2] to nanometer feature sizes, significant problems arise, including large parametric variations, very high leakage power consumption, and short channel effects [3]. Researchers are motivated to work on new, emerging technologies like carbon nanotube field effect transistors (CNTFET), quantum automata (QCA) [4], and nanowires [5] as a result of these constraints. CNTFETs are more promising and feasible than other nano-devices because they share infrastructure and functionality with conventional MOSFET devices, as well as their operation [6]. Although, CNFET based circuits encounter considerable manufacturing problems, including differences in CNT doping concentration, diameter, and alignment. Methods for designing circuits in the presence of such fluctuations have been described by researchers [7].

To achieve a high-performance full adder, it is critical to choose an appropriate digital circuitry architecture. The basic element of a full adder is an XOR cell. To implement the XOR cell with CMOS requires eight transistors. To reduce complexity and transistor count, Pass Transistor Logic (PTL) is evolved for low-power digital circuitry [8] [9]. Though PTL has higher speed and less power consumption than CMOS, there are certain issues like threshold drop that prompted a lower drive current and led to slower performance. To resolve these concerns, techniques like transmission gate CMOS [10], complementary pass transistor (CPL), and double pass transistor (DPL) were suggested.

A leap toward a low-power technique with higher speed and low-power consumption, a Gate Diffusion Input (GDI) technique [11], was suggested in 2002. It is a novel technique, successfully explored in the low-power regime, as it uses only four transistors to implement the XOR logic. Further, MGDI suggested [12] it uses three transistors to implement the XOR logic. However, both the conventional GDI and modified MGDI both has inadequate swing. To resolve this, the paper suggested a full adder based on the amalgamation of MGDI and DLB. This technique not only improves the full swing as well as the power consumption due to the pre-charge operation for the internal and output node of the circuit.

The paper is arranged in four sections—the methodology is described in section 2. The proposed circuits and functional verifications are placed in section 3 of the paper. The findings are summarized in section 4

## II. Methodology

### *A. Carbon Nano Tube Field Effect Transistor (CNTFET)*

CNTFET technology is considered to be one of the leading candidates to replace conventional CMOS technology because of its many benefits, including a small feature size, low power consumption, high speed, and device behaviour similar to that of MOSFET [13]. The structural layout of the CNTFET, same depicted in Fig. 1

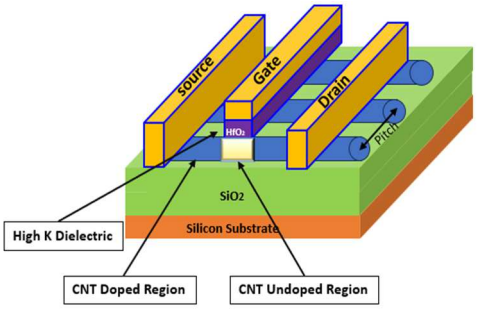



Fig. 1 CNTFETs Structural Layout

979-8-3315-4335-8/25/$31.00 ©2025 IEEE



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:58:04 UTC from IEEE Xplore. Restrictions apply.

There are two types of CNTFET, singled wall CNT [14] and multiwalled CNT [15]. The single wall further categories in armchair, zig-zag and chiral, based on the chiral vector (n1 and n2) or it depending upon the direction of the rolled graphene. The diameter of the CNT also depends upon chiral vector and DCNT given by equation (1) [16]. The threshold voltage for a CNTFET based devices has a two expression, in first expression it's depend upon the Eg (bandgap) and in second equation it depends upon V$\Pi$, same is given in (2).

$$D_{CNT} = \sqrt{\frac{a_0}{\Pi}} * \sqrt{n_1^2 + n_2^2 + n_1 n_2} \qquad (1)$$

$$V_{th} = \frac{E_g}{2\,e} \approx \frac{\sqrt{3}}{3} \frac{aV_{\Pi}}{e\,D_{CNT}} \qquad (2)$$

Where, $a_0 = 2.459\ A^{\circ}$, correspond to the lattice constant in the graphene sheet, Eg corresponds to the bandgap and $V_{\Pi}$ represents carbon pi-pi bond energy

*B. Existing GDI Cell*

The conventional GDI cell [17] is known for its compact design, as shown in Fig. 2. To implement the XOR logic, the GDI cell uses a four-transistor count (two for the inverter and two for GDI) as shown in Fig. 3. However, the GDI cell has an insufficient swing due to a threshold drop when the device PMOS operates to pass logic '0' and NMOS is used to pass the logic '1' for the logic combination of '01' and '11' respectively for XOR logic. Similarly, the logic is not in full swing for other logical functions; the same is summarized in Table I.

In MGDI [12], it can be realized to get an XOR logic with only three transistors. The swing issue remains the same as that in GDI. The XOR logic implementation from MGDI is depicted in Fig. 4. For the logic input 'XY' set at '00', transistors MT1 and MT3 turn ON, and MT2 turns off. Since PCNTFET passed weak '0' or |Vtp|. Likewise, for the input '01', MT1 turns OFF, MT2 and MTF3 turn ON, and the obtained ratioed output. For the other two combinations, '10' and '11', the output is strong '1' and strong '0' respectively; the same is summarized in Table II.

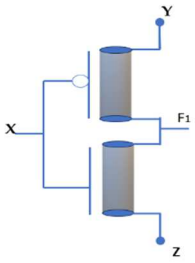



Fig. 2 GDI Cell

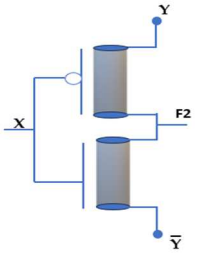



Fig 3 XOR logic with GDI cell

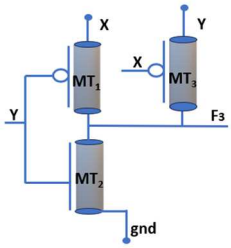



Fig. 4 XOR logic with MGDI

*C. Dynamic Modified GDI (DMGDI) cell*

The dynamic logic utilizes with the concept of pre-charging and evaluation by employing a clock signal. The pre-charged phase is the mode of operation in which the output node is temporarily charged to the Vdd when the clock (clk) = '0'.

Additionally, it enters the evaluation phase at clock (clk) = '1', where input combinations determine the intended output. Fig. 5 illustrates the DMGDI cell, merger of MGDI cell with Dynamic Logic Block (DLB) [18]. DLB comprises of two transitor MD4 and MD5, it works as an inverter and helps to obtained the full swing.

TABLE I. FUNCTION REALIZATIONS USING GDI WITH VARIOUS COMBINATIONS OF INPUTS

| X | Y | Z | OUT | Function | Logic confronts Inadequate Swing |
|---|---|---|---|---|---|
| X | Y | '1' | X+Y | OR | 00,10,11 |
| X | Y' | '0' | X'Y' | NOR | 01 |
| X | '0' | Z | XZ | AND | 00,01,11 |
| X | '1' | Y' | (XY)' | NAND | 10 |
| X | Y | Z | X'Y+XZ | MUX | 00,10 |
| X | Y | Y' | X'Y+XY' | XOR | 00,10 |
| X' | Y | Y' | X'Y'+XY | XNOR | 11, 01 |

TABLE II. FUNCATIONALITY OF XOR GATE FOR MGDI CELL

| X | Y | Transistor State of the Devices | | | Output |
|---|---|---|---|---|---|
| | | MT1 | MT2 | MT3 | |
| 0 | 0 | ON | OFF | ON | \|Vtp\| |
| 0 | 1 | OFF | ON | ON | Ratioed |
| 1 | 0 | ON | OFF | OFF | Strong Logic '1' |
| 1 | 1 | OFF | ON | OFF | Strong Logic '0' |

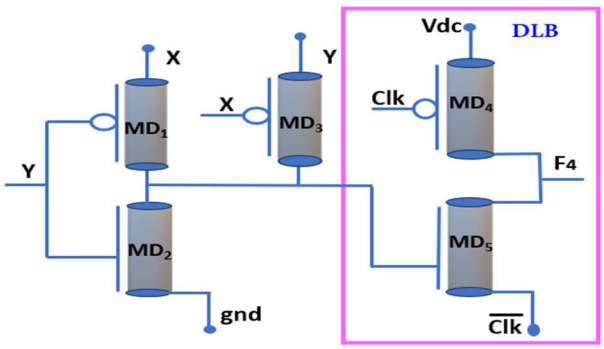



Fig. 5 DMGDI cell

## III. PROPOSED FULL ADDER BASED ON THE DYNAMIC MODIFIED GDI (DMGDI)

In a microprocessor, an arithmetic unit requires a full adder, it is a digital circuit used to add three binary number. It generates two outputs, sum and output carry (cout), from three inputs, X, Y and Z. The logics for the SUM bar and Carry bar [19] is summarized in the equation 3 and 4.

$$\overline{SUM} = \overline{XYZ} + \bar{X}YZ + XY\bar{Z} + X\bar{Y}Z = \bar{Z}\,(YX + \overline{YX}) + Z\,(\bar{Y}X + Y\bar{X}) \quad (3)$$

$$\overline{COUT} = \overline{XYZ} + \bar{X}Y\bar{Z} + X\overline{YZ} + \overline{XY}Z = (\bar{Z})\,(\overline{XY} + XY) + \bar{X}(\bar{Y}X + Y\bar{X}) \quad (4)$$



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:58:04 UTC from IEEE Xplore. Restrictions apply.

### A. Result and discussion

The simulations are carried out using Cadence Virtuoso software, which uses a Stanford model for the CNTFET 32 nm technology node. Table III, summarized key parameters and setting for the CNTEFT devices utilized in the simulation.

TABLE III. CNTFET PERFORMANCE PARAMETER

| Device Parameter | Description | Value |
|---|---|---|
| Lg | Gate Length | 32 nm |
| Lss | Source-channel length | 32 nm |
| Lgeff | Carrier mean free path | 200 nm |
| Pitch | Spacing between the neighbouring CNT within a single device | 20 nm |
| Tox | Thickness of dielectric material | 4.2 nm |
| Csub | Channel to substrate capacitance | 20 pF/m |
| Efi | Doped source/drain tube Fermi Level | 0.7 |
| n1 | Chirality | 19 |

In Fig. 6, the schematic of a 1-bit full adder is displayed. $M_1$, $M_2$, and $M_3$ give the XOR logic for inputs X and Y; the DLB then receives the XNOR logic for X and Y ($M_4$ and $M_5$). Furthermore, the XOR of XΘY with Z will be produced by the $M_6$, $M_7$, and $M_8$. Additionally, DLB ($M_9$ and $M_{10}$) generates SUM. Cout gets generated by providing the XΘY as a control signal to the GDI MUX block, with input of Z to PCNTFET of $M_{11}$ and X to PCNTFET of $M_{12}$. The inverter and DLB are linked in series to obtain COUT in order to achieve a full swing.

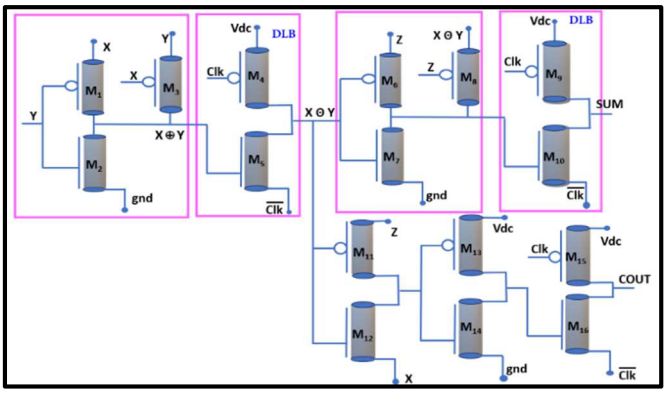



Fig. 6 Schematic of the full adder (1-bit) as per Dynamic Modified Gate Diffusion Input (DMGDI)

Transient analysis is performed using a 100 MHz clock frequency and 1 fF load capacitance with 0.8 V Vdc. The inputs X, Y, and Z were set at frequencies of 12.5MHz, 25MHz, and 50MHz, respectively. To take into consideration every transition in accordance with the truth table, transient analysis was carried out at 80 nm. It may be observed that output is pre-charged to Vdd when clk = '0'. When clk eventually becomes '1', the SUM is at logic '0' when two inputs out of three are at logic '1', and evaluated to '1' when one input out of three is set at '1'. Similarly, the COUT is at logic '1' when two inputs out of three are at logic '1', and evaluated to '0' when one input out of three is set at '1'. When all inputs 'XYZ' are set to '000', COUT and SUM will be obtained as '000' respectively. Thus, all the logic of SUM and COUT is realized successfully. Fig. 7 shows a full adder simulated timing waveforms based on the proposed design.

The delay is measured when the clock transits from logic '0' to '1'; SUM and COUT go from logic '1' to '0'. Delay calculated by acquiring the concise interval of time between the clock's rising edge at 50% of Vdd (0.4V) and the output's falling edge at 50% transition value of Vdd (0.4). The delay plot is shown in Fig. 8. The simulation of the aforementioned circuit revealed that the delay value for COUT and SUM was around 25.26 ps. The measured power usage was around 73.3 nW.

In the published sources, a variety of dynamic GDI-based methods have been taken into consideration and investigated [18] [20]. Table IV provides a summary of the comparison between the state-of-the-art method and the suggested full adder with a full swing based on dynamic MGDI. `In comparison to the circuitry suggested in [18], [20], [21] and [22] the suggested DMGDI performed better, reducing delay by 30.41%, 85.9%, 62.9% and 67.23% and energy expenditure reducing by 42.28%, 83.18%, 92.5% and 93.45% respectively, resulting in a PDP reduction of 59.78%, 97.64%, 97.25% and 97.85% respectively.

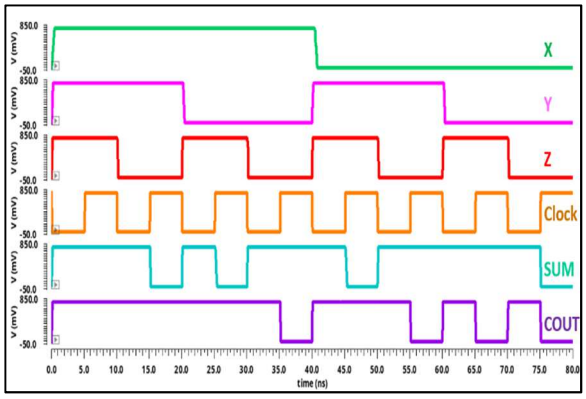



Fig. 7 Simulation result of the1-bit full adder DMGDI based

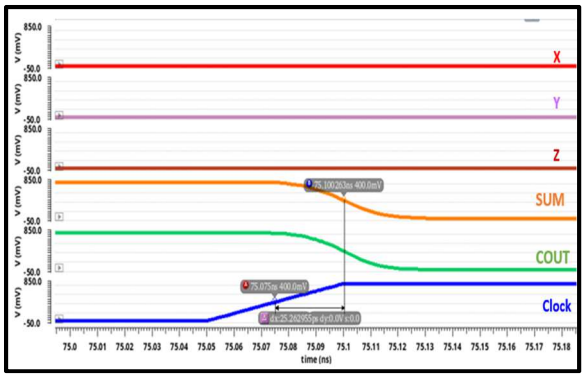
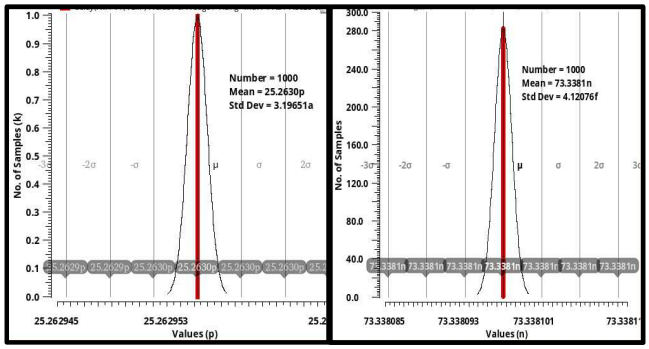



Fig. 8 Delay plot for DMGDI based 1- bit full adder

Additionally, a Monte Carlo simulation was run for 1000 samples to assess how process variation affected the delay and power expenditure of the full adder based on DMGDI. The Monte Carlo graphs for the power consumption and delay of the full adder are displayed in Fig. 9. As may be shown, the delay and power consumption standard deviations are 3.19a and 4.12f, respectively. As a result, the circuit can withstand process variation. Further, PDP plot for the design techniques explored for the full adder is depicted in the Fig. 10.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:58:04 UTC from IEEE Xplore. Restrictions apply.




(a) (b)

Fig. 9 Monte Carlo plot for full adder based on DMGDI (a) delay (b) Power consumption

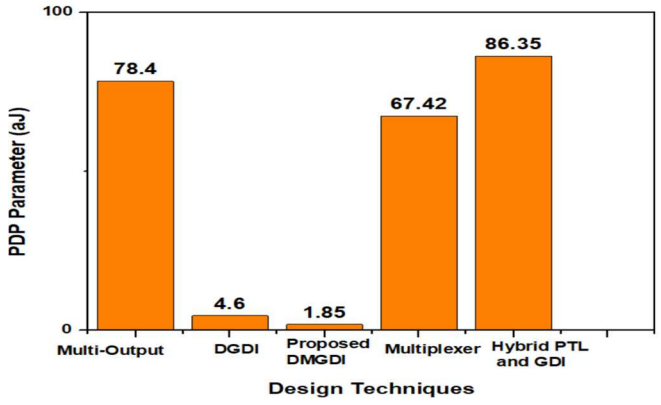



Fig. 10 PDP parameters of different design Technique

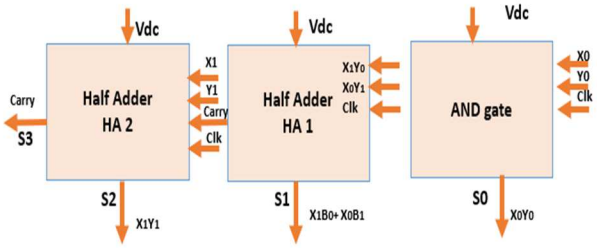



Fig. 11 Schematic for 2x2 multiplier based on DMGDI

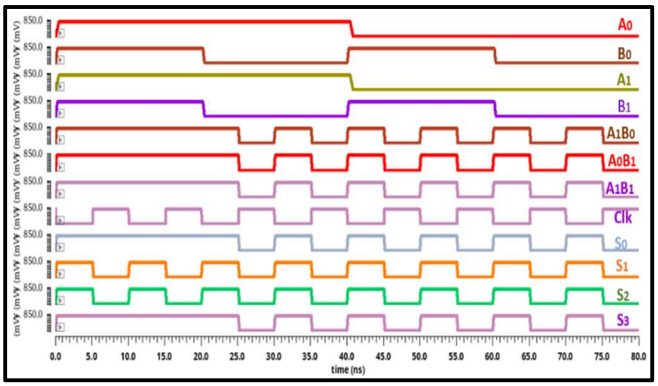



Fig. 12 Simulation waveform for DMGDI based 2x2 multiplier

## B. 2X2 Muiliplier

A 2x2 multiplier circuitry that accepts two 2-bit binary value ($X_1$ $X_0$) and ($Y_1$ $Y_0$) as input and yields a 4-bit binary output ($S_3$ $S_2$ $S_1$ $S_0$) indicating their product. AND gates and two half adder processes are used to compute the multiplication. Fig. 11 shows the block diagram used to implement the 2x2 multiplier.

Initial circuitry yields four outputs: $X_0Y_0$, $X_1Y_0$, $X_0Y_1$, and $X_1Y_1$. The product $X_0Y_0$ yields $S_0$, the multiplier output's (B) least significant bit (LSB). The $S_1$ bit is obtained by adding $X_1Y_0$ and $X_0Y_1$ using a half adder ($HA_1$). The $S_2$ bit is created by adding $X_1Y_1$ and the carry output from the first half adder ($HA_1$). The $S_3$ bit is the carry-out of the second half adder ($HA_2$). The simulated waveforms for each input and output are shown in Fig. 12, and the delay graphs for the 2x2 multiplier are shown in Fig. 13. Furthermore, the delay measured for obtaining the MSB of the multiplier output is approximately 41.52 pS. The same circuitry uses about 483.5 nW of power.

To evaluate the impact of process variation on the delay and power consumption of the DMGDI-based 2x2, a Monte Carlo simulation involving 1000 samples was conducted. The Monte Carlo simulation power consumption and delay plot for the full adder, with a 15% pitch variation, is displayed in Fig. 14. As may be shown, the power consumption and delay standard deviations are 16.6f and 22.7a, respectively. As a result, the circuit exhibits resistance to changes in process.

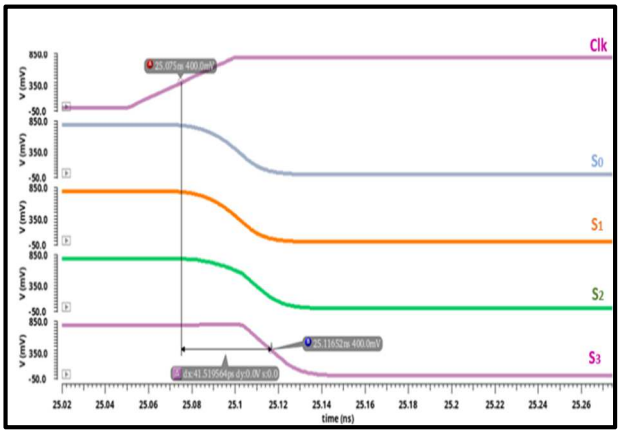
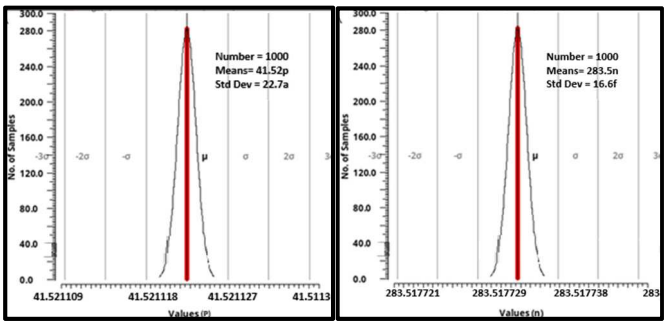



Fig. 13 Delay Plot for DMGDI based Multiplier




(a) (b)

Fig. 14 Monte Carlo plot for DMGDI based 2x2 Multiplier (a) Delay (b) Power consumption.

In comparison to multi-output FA and DGDI the suggested multiplier has performed better, reducing delay by 89.51% and 42.17%, while simultaneously reducing power consumption by 84.5% and 45.55%, resulting in PDP reductions of 98.3% and 68.6%, respectively, same is shown in Table V.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:58:04 UTC from IEEE Xplore. Restrictions apply.

TABLE IV. BENCHMARK PARAMETER OF THE 1-BIT FULL ADDER

| Adder | Design | Power (μW) | Delay (pS) | PDP (aJ) |
|---|---|---|---|---|
| Dynamic | Multi-Output FA [20] | 0.436 | 179.91 | 78.4 |
| Dynamic | DGDI [18] | 0.127 | 36.3 | 4.6 |
| **Dynamic** | **Proposed DMGDI** | **0.0733** | **25.26** | **1.85** |
| Static | Multiplexer [21] | 0.99 | 68.1 | 67.42 |
| Static | Hybrid PTL and GDI [22] | 1.12 | 77.1 | 86.35 |

TABLE V. PERFORMANCE PARAMETER OF THE 2X2 MULTIPLIER

| Adder | Design | Power (μW) | Delay (pS) | PDP (aJ) |
|---|---|---|---|---|
| Dynamic | Multi-Output FA | 1.83 | 395.8 | 724.3 |
| Dynamic | DGDI | 0.5207 | 71.8 | 37.38 |
| **Dynamic** | **Proposed DMGDI** | **0.2835** | **41.52** | **11.7** |

## IV. CONCLUSION

This study investigated a novel method for creating low-power 2x2 multipliers and full adders. In order to optimize the Power-Delay Product (PDP) for XOR logic and achieve full swing operation, the DMGDI cell was designed by combining a new dynamic logic block (DLB) with Modified Gate Diffusion Input (MGDI). At 32 nm technology, designs were thoroughly validated in Cadence Virtuoso using CNTFETs. The DMGDI-based full adder and 2x2 multiplier provide primarily energy-efficient circuits, with significant delays and power consumption reductions. The DMGDI full adder achieved a delay of only 25.26 picoseconds (ps) and a power consumption of 73.3 nanowatts (nW). The 2x2 multiplier used 283.5 nW of power with a delay of about 41.52 ps. The findings highlight the feasibility of DMGDI-based circuits for low-power applications. Monte Carlo simulations, which also performed, demonstrated that DMGDI design was robust against process variations.

## REFERENCES


[1] J. Ponnian, S. Pari, U. Ramadass, O.C. Pun, “A new systematic GDI circuit synthesis using MUX based decomposition algorithm and binary decision diagram for low power ASIC circuit design”, Microelectron. J. 108, 104963 (2021)

[2] Y. Lin *et al*., "Characterization and Modeling of MOSFET Series Resistance Down to 4 K," in *IEEE Journal of the Electron Devices Society*, vol. 13, pp. 297-302, 2025, doi: 10.1109/JEDS.2025.3544738.

[3] M. S. Equbal, S. Sahay, “Nanoelectronics : Physics, Materials and Devices” 2023, Pages 11-37, ISBN 9780323918329,

[4] A.O. Orlov, R. Kummamuru, J. Timler, C.S. Lent, G.L. Snider and G.H. Bernstein, “Experimental studies of quantum-dot cellular automata devices” , *Mesoscopic Tunnelling Devises, 2004.*

[5] K. Xing, J. Hui, S. Zhang, “Chirality of sub-nanometer nanowires/nanobelts”, Nanoscale 2025 Royal Society of Chemistry

[6] Divyansh Yadav, Anuja Bhargava, Elangovan Mani, Ashish Sachdeva, "Design and parametric characterization of CNTFET based stable static random access memory bit-cell for low-power applications", AEU - International Journal of Electronics and Communications, Volume 190, 2025,155642, ISSN 1434-8411.

[7] J. Huang, & M. Zhu, & S. Yang & P. Gupta, W. Zhang, S. Rubin, G. Garreton & J. He, “A physical design tool for carbon nanotube field-effect transistor circuits. ACM Journal on Emerging Technologies in Computing Systems”, 2012, JETC. 8. 10.1145/2287696.2287708.

[8] Sudhakar, Pasupuleti Naga, and Vijaya Kishore Veparala. "Design of GNRFET ternary circuits using pass transistor logic (PTL)." *AIP Conference Proceedings*. Vol. 3237. No. 1. AIP Publishing, 2025.

[9] Jyothi, B., BV Ramana Reddy, and Mansi Jhamb. "Gate all around carbon nanotube field effect transistor espoused discrepancy cascode pass transistor adiabatic logic for ultra-low power application." *Integration* 100 (2025): 102260.

[10] T.S. Kumar, I.R.S.N. Rao, Y.S. Vinod, P. Harika, B.V.V. Satyanarayana, A. Pravin, “Area Efficient and Ultra Low Power Full Design Based on GDI Technique for Computing Systems”, 2024, vol 537, springer, Cham.

[11] A. Morgenshtein, A. Fish and I. A. Wagner, (2002), "Gate-diffusion input (GDI): a power-efficient method for digital combinatorial circuits", in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 10, no. 5, pp. 566-581, Oct. 2002,

[12] R. Uma, P. Dhavachelvan, “Modified Gate Diffusion Input Technique: A New Technique for Enhancing Performance in Full Adder Circuits”, Procedia Technology, Volume 6, 2012, Pages 74-81, ISSN 2212-0173,

[13] I. Khan, (2022). Design and Implementation of Carbon Nano-tube based Full Adder at 32nm Technology for High Speed and Power Efficient Arithmetic Applications. Journal of Physics: Conference Series. 2161. 012050, 2022.

[14] M. Fedawy, W. Fikry, A. Alhenawy and H. Hassan, "I-V characteristics model for ballistic Single Wall Carbon Nanotube Field Effect Transistors (SW-CNTFET)," 2012 IEEE International Conference on Electronics Design, Systems and Applications (ICEDSA), Kuala Lumpur, Malaysia, 2012, pp. 10-13

[15] J.H. Lehman, M. Terrones, E. Mansfield, K.E. Hurst and V. Meunier , “Evaluating the characteristics of multiwall carbon nanotubes”, 2011, Carbon, 49(8), pp.2581-2602.

[16] A. Eatemadi, H. Daraee, H. Karimkhanloo, et al. “Carbon nanotubes: properties, synthesis, purification, and medical applications”, Nanoscale Res Lett 9, 393 (2014)

[17] A. Morgenshtein, A. Fish and I. A. Wagner, "Gate-diffusion input (GDI) - A technique for low power design of digital circuits: analysis and characterization," 2002 IEEE International Symposium on Circuits and Systems. Proceedings (Cat. No.02CH37353), Phoenix-Scottsdale, AZ, USA, 2002, pp. I-I,

[18] A. Ghorbani, M. Dolatshahi, S. M. Zanjani and B. Barekatain, (2022), “A new low-power Dynamic-GDI full adder in CNFET technology, Integration”, Volume 83, 2022, Pages 46-59, ISSN 0167-9260,

[19] K. Sanapala, and R. Sakthivel, “Ultra-low-voltage GDI-based hybrid full adder design for area and energy-efficient computing systems. IET Circuits Devices Syst.”, 2019, 13: 465-470. doi.org/10.1049/ietcds.2018.5559

[20] R. F. Mirzaee, M. H. Moaiyeri and K. Navi, “High speed NP-CMOS and multi-output dynamic full adder cells. International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering”, 2010, 4. 531-537.

[21] Y. Jiang, A. Al-Sheraidah, Y. Wang, E. Sha, & J. Chung,-G, “A novel multiplexerbased low-power full adder. IEEE Transactions on Circuits and Systems II: Express Briefs”, 2004 51(7), 345–348.

[22] A. Sharma, & R. Mehra, “Area and power efficient CMOS adder design by hybridizing PTL and GDI technique. International Journal of Computer Applications”, 2013, 66(4), 15–22.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:58:04 UTC from IEEE Xplore. Restrictions apply.

2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT) | 979-8-3315-4335-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/DICCT64131.2025.10986635

# HEMT based Different Structures for High Threshold and Breakdown Voltages: A Review

Aastha Singh
*Electronics and Communication Engineering Department*
*Delhi Technological University*
*Delhi, India*
Aasthasingh_2k23vls10@dtu.ac.in

Sachin Dhariwal
Electronics and Communication Engineering Department
*Delhi Technological University*
*Delhi, India*
sachindhariwal@dtu.ac.in

***Abstract*— The present review delves into the latest developments in High-Electron-Mobility Transistors (HEMTs), emphasizing the improvement of both the breakover voltage (Vbr) and threshold voltage (Vth). Several methods for modulating electric fields and lowering parasitic unwanted current are explored, including the use of composite step-like design gate structures with charge balancing plate and p-GaN buried layers. Further research is done on double AlGaN barrier designs, showing that reducing the amount of Al in the bottom barrier can lower 2DEG density and improve gate control, which raises Vth. The performance of enhancement -mode HEMTs is further enhanced by optimizations including graded AlGaN barriers, recessed gate designs, p-GaN caps, and the addition of dielectric layers like $Al3O_3$. These methods ensure acceptable transconductance and saturation current while striking a compromise between high turn-on voltage (switching voltage) and breakover voltage.**

***Keywords*— Breakdown voltage (Vbr), switching Voltage (Vth) , leakage Current , AlGaN Barrier Design, P-GaN buried layer , charge balancing plates , Composite step-like design gate structure.***

## I. INTRODUCTION

Due to their superior achievements in fields of high breakover voltage, efficiency, and electron mobility, high speed transistors have become vital components in power electronics and high frequency uses. GaN-based HEMTs, in particular, have garnered significant attention because of their wide bandgap, fast electron saturation velocity, and strong electric field capabilities, making them well-suited for high-power and high-temperature environments.

Attaining enhancement-mode (normally-off) operation is a major difficulty in GaN-based HEMTs. Enhancement-mode devices are favored for power applications due to their ability to operate safely even without the need for an applied gate voltage. Depletion-mode (normally-on) behaviour is a characteristic of HEMTs, which causes so much parasitic flows and raises extreme care for safety purpose and operational reliability concerns and mandates the integration of circuitry for power-down conditions. So Attaining a large positive switching voltage (Vth) is a major concerns in modern HEMT design.

To increase the switching voltage and enhance operational reliability, many strategies are opted by designer. Among these, in my literature survey, the incorporation of a Results from the composite step-like gate design structure have been encouraging. In the traditional planar HEMT design with AlGaN/GaN configuration, the heterointerface layer is transformed into a Stair-like structure by combining a p-Gallium Nitride cap layer with a high-k dielectric such as hafnium oxide. The composite design maintains the threshold voltage (Vth) at up to 3.5V while increasing the breakover voltage (Vbr) by the addition of P-type GaN Layer and electric field distribution plates. The dispersion of the electric field is efficiently optimized with this method. Boost the breakover voltage to 2440V and reduce parasitic currents. With careful layer composition and step height optimization, the step-like design gate structure effectively strikes a balance between higher Vth and the need for a robust output current.

Using twin AlGaN barrier designs in metal-insulator-semiconductor HEMTs with buried gates is an alternative tactic. These devices use a top barrier with a greater aluminium content and a bottom barrier with a suppressed aluminium content to produce better gate control and a more positive Vth. A greater gate voltage is needed to establish channel conductivity when the 2DEG density drops as a result of a decrease in interface polarization charges that are controlled by a lower Al composition in the bottom AlGaN barrier. By putting the gate closer to the channel, deeper AlGaN layer recessing enhances gate control and enables usually off operation [1-5].

Enhancement-mode features can also be obtained by combining buried gate designs with dielectric layers, such as $Al3O_3$. By increasing Vth and reducing gate leakage, a dielectric covering the recessed p-GaN layer helps improve gate voltage distribution. This composite recessed-gate design balances power efficiency and switching speed while increasing transconductance and saturation current while maintaining a high Vth. [1] [7-11]

To reduce alloy disorder dispersion and increase electron transport efficiency in the 2DEG channel, studies have also looked into grading the aluminium composition in the AlGaN barrier. By lowering maximum electric intensities close to gate and attaining a more consistent electric field allocation graded AlGaN layers enhance device performance in general. Adding N-wells to the GaN channel is another optimization technique that can increase source-drain current without lowering the threshold voltage (Vth). [12-15]

979-8-3315-4335-8/25/$31.00 ©2025 IEEE



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:13 UTC from IEEE Xplore. Restrictions apply.

Consequently, innovative approaches such as composite stepped gates, double AlGaN barriers, and buried gate configurations are being employed to consistently enhance the threshold, breakdown voltage and transconductance in GaN-based HEMT designs. Future developments in power electronics may benefit greatly from these breakthroughs, which have the potential to improve safety and efficiency of HEMTs in high-power, high-frequency applications.

## II. Literature Review

### *2. Buried Gate AlGaN/GaN MISHEMT structures featuring dual AlGaN barriers*

### 2.1. Device configuration and specifications

The structure depicted in Fig 2.1 consists of a GaN channel, an AlN nucleation layer, a 30% Al AlGaN barrier, and lower barriers with 20% and 15% aluminum composition. A 1 nm GaN cover layer is also included. To fabricate the buried gate structure, reactive ion etching (RIE) was performed using a gas mixture of $BCl_3$ (10 sccm) and $Cl_2$ (15 sccm) at a controlled rate of 5 Å/sec.

Subsequent steps involved the PECVD deposition of a 15-nm $Si_3N_4$ layer, serving as both the gate dielectric and a layer for surface passivation in the access areas. The gate metal, TiN, was used along with Ti/Al-based Au-free Ohmic contacts. Etching of the $Si_3N_4$ layers and the AlGaN barrier was also carried out. Annealing was conducted in a nitrogen environment at 600°C for 1 minute, achieving a contact resistance ($R_a$) of 1 Ω·mm.

For electrical characterization, devices were fabricated with a gate length (Lg) of 1 µm, a source-to-gate distance (Lgs) of 2 µm, and a gate-to-drain distance (Lgd) of 6 µm.[1].

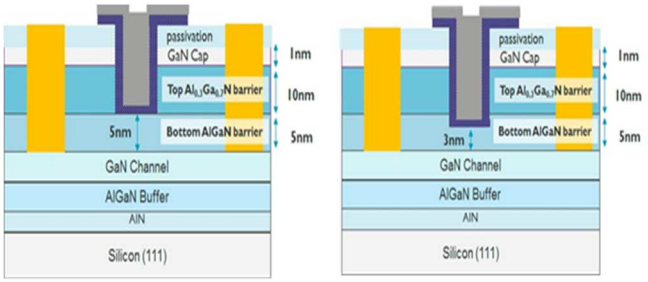



Fig 2.1. Diagram illustration of the device architectures of buried gate AlGaN/GaN MIS-HEMTs with (a) residual AlGaN barrier with a thickness of 5 nm located under the gate region and (b)a 3 nm residual AlGaN barrier situated beneath the gate region [1].

### 2.2. Results and Discussion:

Reducing the aluminium content in the AlGaN barrier diminishes polarization charges, resulting in a decreased 2DEG density and necessitating an elevated gate voltage to achieve equivalent channel conductivity. A diminished Al content also lowers the conduction band offset, complicating electron population in the 2DEG and elevating the threshold voltage (VTH). Moreover, it enhances gate control over the channel, hence increasing VTH. A deeper recess in the AlGaN barrier, resulting in a reduced layer thickness, improves gate control and decreases 2DEG density, both factors contributing to an elevated VTH and facilitating normally-off operation. An AlGaN barrier measuring 5 nm with 20% aluminium content attains a threshold voltage (Vth) of approximately 0 V, whereas a 3 nm Al0.15Ga0.85N barrier situated beneath the gate elevates Vth to roughly 3.25 V. A twin AlGaN barrier configuration in buried gate MIS-HEMTs represents a viable strategy for attaining enhancement mode functionality.

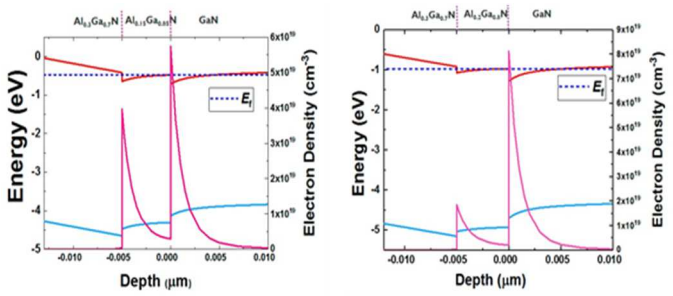



Fig 2.2. Simulated band diagrams featuring (a) an Al0.3Ga0.7N/Al0.15Ga0.85N barrier and (b) an Al0.3Ga0.7N/Al0.2Ga0.8N barrier [1].

The typical buried p-GaN HEMT with an $Al_2O_3$ cap layer showed improvements in threshold voltage and suppression of gate leakage current. However, it experienced a decrease in saturation current and transconductance. This design faced challenges in sustaining high saturation current and transconductance levels limiting its overall performance. [1]

The next innovation, the composite buried gate transistor, addressed the limitations of the previous design by removing half of the p-GaN layer and depositing an $Al_2O_3$ layer, creating a composite recessed gate. This modification allowed the structure to retain a elevated threshold voltage while significantly improving both saturation current and transconductance, enhancing device's overall performance.

### *3. Composite buried gate featured normally off AlGaN/AIN/GaN HEMT*

### 3.1. Design configuration and specifications:

A enhancement_ mode AlGaN/AlN/GaN HEMT utilizes a composite buried gate structure for enhanced performance. The fabrication begins with an epitaxial layer stack epitaxially grown on a sapphire wafer. This stack includes a 15 nm AlN buffer layer, a 20 nm Unintentionally impurity-infused barrier layer containing 24% aluminum, a 0.4 µm undoped GaN channel layer, and additional buffer layers deposited via MOCVD. Ohmic contacts for the source and drain are formed using Ti/Al/Ni/Au metals, which are annealed at 850°C. A Ni/Au metal gate is deposited to establish the Schottky contact. The creation of a buried gate through the partial removal of the p-GaN layer and the deposition of an $AlO_3$ dielectric layer is a crucial step in this procedure. By utilizing the benefits of both p-GaN and MIS designs, this setup improves the functionality and performance of the device. [2].

The transconductance, saturation current, and threshold voltage of the device are all improved. A threshold voltage of 4.5 V is achieved by the buried p-GaN HEMT by spreading the



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:13 UTC from IEEE Xplore. Restrictions apply.

gate voltage across the $Al_2O_3$ dielectric layer and lowering the two-dimensional electron gas beneath the gate. In order to reduce gate leakage and increase device reliability, the $Al_2O_3$ layer is placed between the gate and the p-GaN layer. While the right portion (regular buried MIS-HEMT) improves both saturation current and transconductance, the left segment (p-GaN with $Al_2O_3$) of the composite buried-gate HEMT mainly raises the threshold voltage. In comparison to the buried p-GaN HEMT, this design maintains a high threshold voltage while providing a 121.33% increase in saturation current and a 13.14% increase in transconductance, with just a 4.44% decrease in threshold voltage. [2].

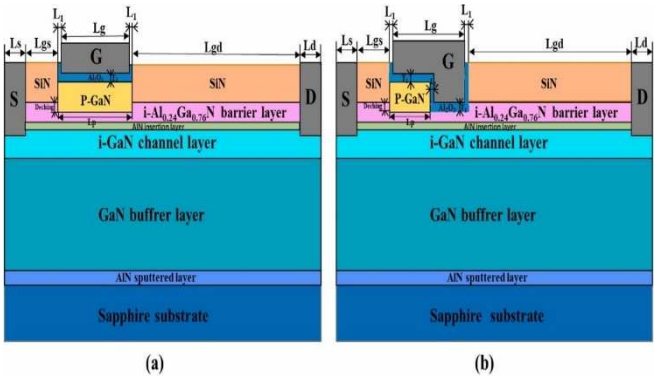



Fig. 3.1. A diagram cross-section illustrates the construct of a composite buried p-GaN HEMT with an $Al_2O_3$ layer [2].

### 3.2. Results and Discussions:

*A. Evaluation of the performance characteristics of the covered buried p-GaN HEMT with an $Al_2O_3$ layer.*

Fig 3.2 presents a performance comparison between the standard buried gate MIS-HEMT, a standard buried-gate p_GaN HEMT, and a buried p-GaN HEMT. The transfer characteristic curves of these three devices, measured at Vds=10 illustrated in Fig 3.3. The threshold voltage (Vth) for the $Al_2O_3$-covered burieds p-GaN HEMT is 4.5 V, with the $Al_2O_3$ gate dielectric contributing to an increase in the p-GaN threshold voltage. In contrast, the standard buried-gate MIS-HEMT shows a Vth of 1.35 V, indicating that lower barrier layers do not significantly enhance the threshold voltage. The buried p-GaN HEMT without $Al_2O_3$ achieves a Vth of 3.5 V, attributed to electron depletion in the channel layer and reduction of the barrier layer thickness. The addition of an $Al_2O_3$ gate dielectric layer primarily impacts the threshold voltage in the $Al_2O_3$-covered buried p-GaN HEMT but does not similarly enhance Vth in other configurations. Because the Al2O3 gate dielectric layer shares voltage, more gate voltage is needed to activate devices. To exemplify this method, Fig. 3.4 depicts the three devices' potential, conduction band energy, and gate positioning relative to the substrate. The Al2O3 layer shares 10 V for Vds and 3Vs for Vg (Fig 3.4(b)(c)). Due to lower channel layer potentials, Al2O3-covered buried P-GaN HEMTs require more potential to switch on. The energy band diagram indicates that the standard buried p-GaN transistor's conduction band energy approaches Fermi and switches on at Vds = 10 V and Vg = 3 V. The buried p-GaN transistor covered in Al2O3 needs more voltage since its conduction band energy is far from Fermi. At Vds = 10 V and Vg = 3 V, the typical buried-gate MIS-HEMT has a greater channel layer potential than the other two devices (Fig 3.1(a)). The $Al_2O_3$ gate dielectric layer has a lower voltage drop than p-GaN, enabling a higher channel potential to activate the device, which explains its lower threshold voltage. However, $Al_2O_3$-coated buried p-GaN HEMTs show reduced transconductance due to increased gate-to-channel distance, which weakens gate control. While $Al_2O_3$ on p_GaN raises the threshold voltage, reduces saturation current and transconductance. To address this, an optimized composite buried-gate HEMT structure (Fig 3.1(b)) is proposed, offering high saturation current, transconductance, and threshold voltage [2].

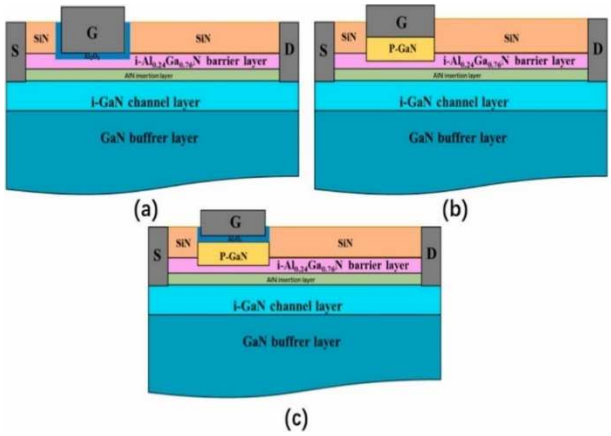



Fig. 3.2. (a) (b) (c) Cross-sectional illustrations of a typical buried-gate MIS-HEMT, a established buried gate p-GaN high speed transistor, and a buried p-GaN high speed transistor with an $Al_2O_3$ layer [2].

A. *Analysis of the performance of composite buried-gate HEMTs*

A composite buried gate combines MIS and p-GaN structures, with $Al_2O_3$ on the right groove and p-GaN on the left. This design achieves a threshold voltage of 4.3 V, which is 0.2 V lower than that of the $Al_2O_3$-covered buried p-GaN HEMT

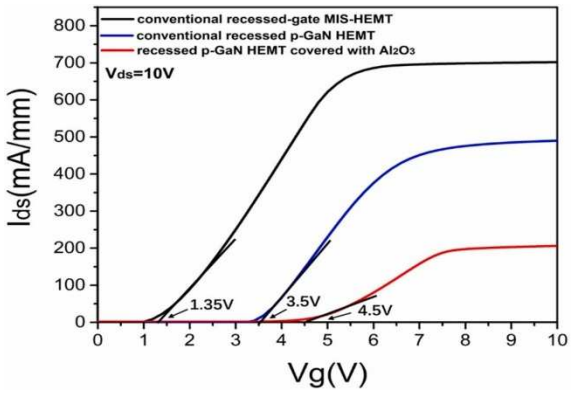



Fig. 3.3. Graphs comparing the threshold voltage of standard buried-gate MIS-HEMT, established buried p-GaN HEMT, and buried p-GaN HEMT with $Al_2O_3$ coating [2].

The MIS structure on the right reduces the 2DEG concentration and inhibits the threshold voltage enhancing the saturation current and channel electron flow at the gate voltage.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:13 UTC from IEEE Xplore. Restrictions apply.

The large $Al_2O_3$/p-GaN region on the left, contributing to a higher threshold voltage , minimizes gate leakage current and improves device reliability [2].

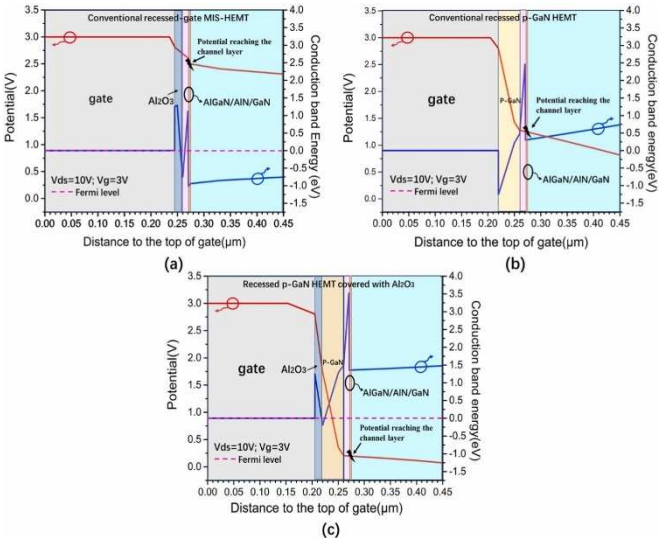



Fig 3.4. (a) (b) (c) Shows the computed conduction band energy diagrams for the standard buried-gate MIS-HEMT, the established buried p-GaN transistor, and the buried p-GaN transistor with $Al_2O_3$ overlay extending from the gate along the substrate axis [2].

B. *Evaluation of the performance characteristics of the covered buried p-GaN HEMT with an $Al_2O_3$ layer.*

An essential factor affecting the performance of composite buried - gate HEMTs is the $Al_2O_3$ gate dielectric layer. This study evaluates the transmission and output characteristics for $Al_2O_3$ thicknesses of 5 nm, 10 nm, 15 nm, 20 nm, and 25 nm, keeping other parameters constant, to Assess the impact of thickness variation on device functionality Fig 3.5 (a) shows the device's threshold voltage (Vth) moving right as the Al2O3 gate dielectric thickens. The device's Vth can reach 4.9 V when the gate dielectric thickness exceeds 25 nm, implying that altering the thickness helps manages the Vth . Al2O3 thickness and Vth are correlated in Fig. 3.6. The corresponding devices' Vth is 3.7 V, 4 V, 4.3 V, 4.6 V, and 4.9 V as the Al2O3 gate dielectric thickness increases from 5 to 25 nm. The device's Vth grows by 0.3 V for every 5 nm increase in the Al2O3 gate dielectric thickness indicating a linear graph growth.

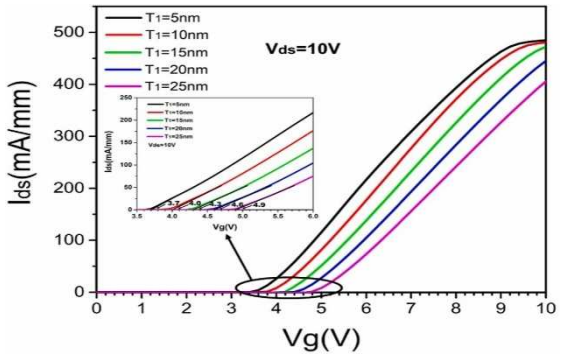



Fig 3.5. Transfer characteristics for the composite buried-gate HEMT with varying $Al_2O_3$ layer thicknesses (T1) [2].

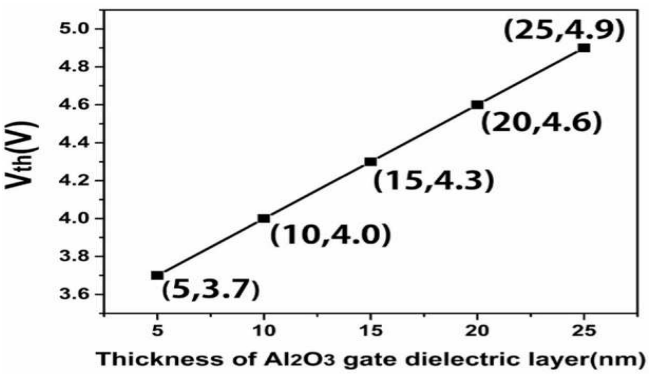



Fig 3.6. Effect of the thickness of $Al_2O_3$ (T1) on Vth (Threshold Voltage) [2].

The potential difference across the dielectric in terms of the electric field can be expressed by the following equation, assuming constant electric field and doping concentration in p-GaN reaches a certain level

Udieletric = Edieletric * Tox

Where U is potential difference across insulating layer
E is electric filed intensity of dielectric
T is thickness of dielectric.

We can see from graph that Dielectric thickness linearly affects voltage drop.

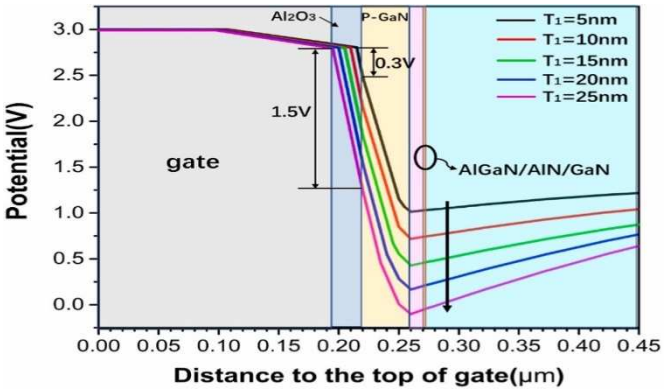



Fig. 3.7. The diagram showing the potential profile across the composite buried-gate HEMT with different $Al_2O_3$ thicknesses (T1) from the left gate to the substrate, under the conditions of Vds = 10 V and Vg = 3 V[2].

This work analyzes the potential profile for different $Al_2O_3$ dielectric thicknesses in a composite buried gate HEMT from gate to substrate to determine the association between gate dielectric thickness and potential. The findings are in fig 3.7. The voltage differential across the gate depends on its thickness. However, a 25 nm layer raises voltage by 0.3 V every 5 nm dielectric thickness. A 5 nm $Al_2O_3$ layer causes 0.3 V voltage differential. Lowering the channel's surface potential requires a greater gate voltage to open it. With Vds = 10 V and Vg = 10 V, fig 3.8(a) the graph demonstrates how the saturation current drops as the $Al_2O_3$ layer's thickness rises. The



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:13 UTC from IEEE Xplore. Restrictions apply.

capacitance relationships listed below apply, with reference to Fig 3.8(b) to analyze saturation current behavior with thickness

$$Q = C_{gate} * U_{gate} \quad (2)$$

$$C_{gate} = \frac{1}{\frac{1}{C_{Al2O3}} + \frac{1}{C_{P-GaN}}} + CAl_2O_3 \quad (3)$$

$$CAl_2o_3 = \varepsilon Al_2o_3 \; / \; dAl_2o_3 \quad (4)$$

Where Q represents charge, Cgate represents gate capacitance, Ugate represents gate voltage, $Cp$-GaN represents p-GaN capacitance, and $C$Al₂O₃ represents Al₂O₃ capacitance. With dielectric constant $\varepsilon$ and thickness $d$, equation (4) indicates a decrease in CAl₂O₃ along with an increase in dAl₂O. Reducing CAl₂O₃ results in a decrease in $C$gate, as seen in Equation (3). Equation (2) states that Cgate declines charge Q while Ugate is constant. Lowering 2DEG concentration lowers Q and saturation current. Fig 3.9. shows optimal Al2O3 thickness, saturation current, and threshold voltage. The ideal threshold voltage and saturation current are 4.5 V and 456 mA/mm for 18.3 nm Al2O3 thickness.

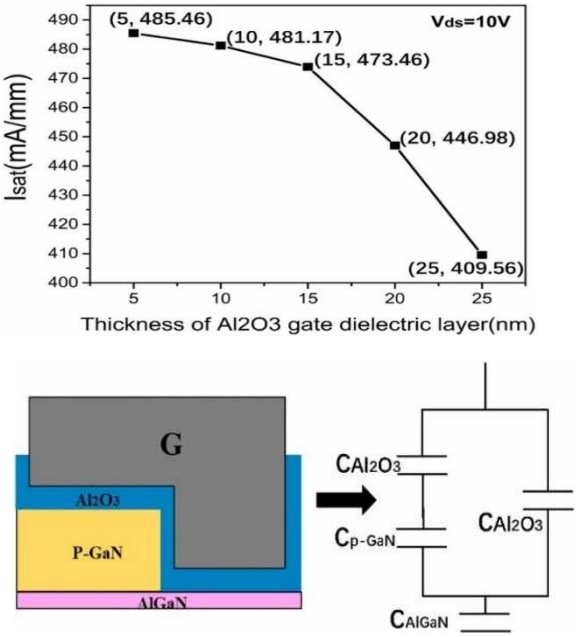



Fig 3.8. (a) On Vds = 10 V and Vg = 10 V , dependency of the saturation current (Isat) on the thickness of $Al_2O_3$ (T1). 3.8. (b) capacitance diagram of Composite device [2].

Ultimately, the optimal thickness is found to be 18.3 nm, resulting in a threshold voltage of 4.5 V and a saturation. current of 456 mA/mm. These findings demonstrate potential of composite buried gate structure in enhancing the performance of normally-off high-power semiconductor devices. [2]. This structure , buried p-GaN HEMT covered with Al2O3 had too

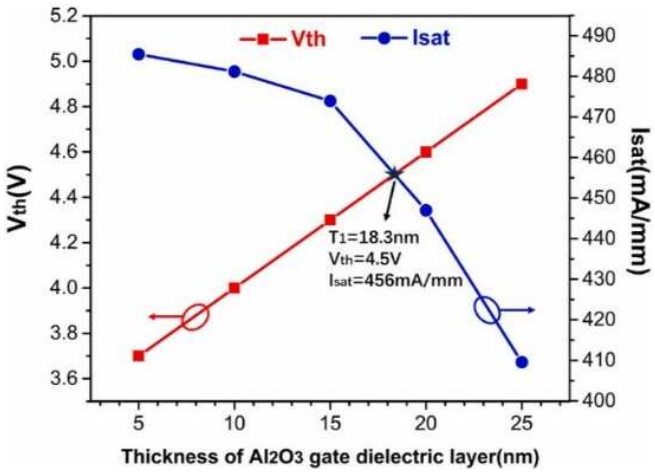



Fig 3.9. Graph determining correlation between Vth, Isat and dielectric thickness (T1) respectively [2].

Diminished saturation current and gain factor despite improving the threshold voltage and suppressing gate parasitic current. This structure , buried p-GaN HEMT covered with Al2O3 had too Diminished saturation current and gain factor despite improving the threshold voltage and suppressing gate parasitic current.

So the introducing new structure , composite step-like gate GaN-based HEMT with p-GaN buried layer and charge balancing addressed these limitations by significantly enhancing breakover voltage and threshold voltage, showcasing significant promise in high-voltage endurance and high-power uses.

## 4. *Normally off GaN-based HEMT featuring a composite structure incorporating a buried p-GaN layer and a charge balancing plate.*

### 4.1. Device architecture and specifications:

Fig 1 illustrates a GaN-based HEMT with a composite step-like gate structure, incorporating a charge balancing and a buried p-GaN layer. A distinctive composite step-like gate is designed by converting the Planar AlGaN/GaN interface into a step-like configuration, incorporating a p-GaN cap and hafnium oxide dielectric layers. The structure includes a charge balancing plate and a buried p-GaN layer beneath the composite gate channel on the right. The composite gate architecture and step-like channel enhance threshold voltage and device dependability, while the charge balancing plate and buried p-GaN layer efficiently Increase the breakover voltage by suppressing parasitic currents and optimizing the allocation of electric intensities. [3].

### 4.2. Results and Discussions

*A. Evaluation and enhancement of GaN-based HEMTs with a step-like gate construct*



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:13 UTC from IEEE Xplore. Restrictions apply.

To assess performance, Fig 4.2 compares the step-like gate GaN-based HEMT to the standard one. The threshold voltages (Vth) of standard and step-like gate GaN-based transistor are -1.9 V and -0.15 V, respectively. Step-like channels greatly enhance threshold values.

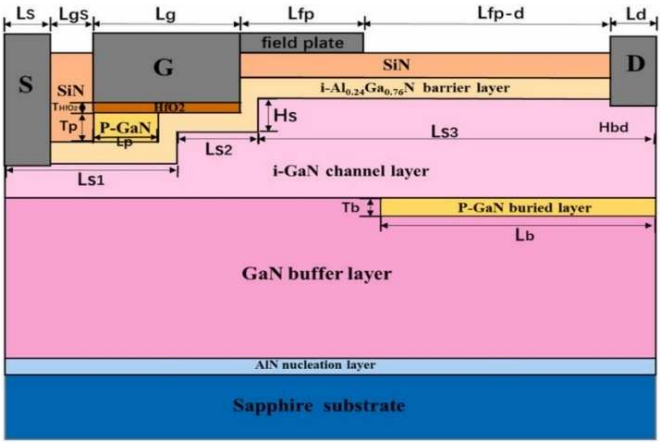



Fig.4.1. A diagram illustrating a GaN-based HEMT with a composite step-like gate, featuring a buried p-GaN layer and a charge balancing plate [3].

This advancement is owing to the step side AlGaN/GaN heterointerface non-polarized development, which lowers 2DEG concentration. This drop decreases the channel's 2DEG, boosting the device threshold voltage in the absence of external voltage, For conventional GaN-based HEMTs, AlGaN/GaN heterojunction polarization induces 2DEG on the channel layer. The step-like gate GaN-based HEMT's 2DEG concentration on the step's side is lower due to its non-polarization growth direction. Lowering 2DEG raises threshold voltage. Electrons on the step's side activate the device when the gate is voltage. This step-like GaN-based HEMT study indicates that the step side decreases 2DEG concentration, considerably changing the threshold voltage Step height also affects step-side electron concentration. The threshold voltage increases with step height. Taller steps lower channel electron concentration, requiring a greater gate voltage to activate the device. Step elevation change influences Vth because the threshold voltage increases to 1.1 v at 175 nm. Increased step height decreases saturation current.

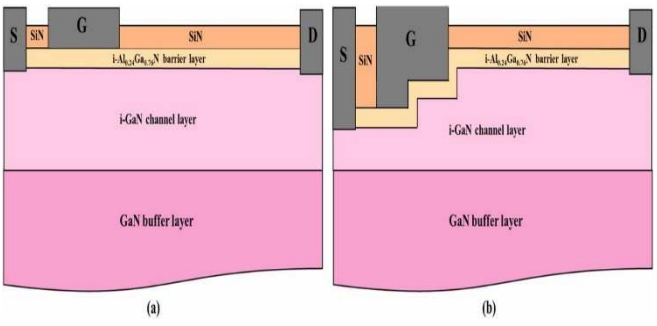



Fig.4.2 (a) Schematic diagram of the typical GaN-based high electron speed transistor; (b) Schematic diagram of the step-like gate GaN-based high electron speed transistor [3].

Side resistance rises with step height, lowering saturation current. High Vth devices enhance circuit safety, but higher output currents increase power. Higher threshold voltages lower output current. Devices with different step heights must interchange threshold voltage and output current. Step height is optimum at 100 nm [3].

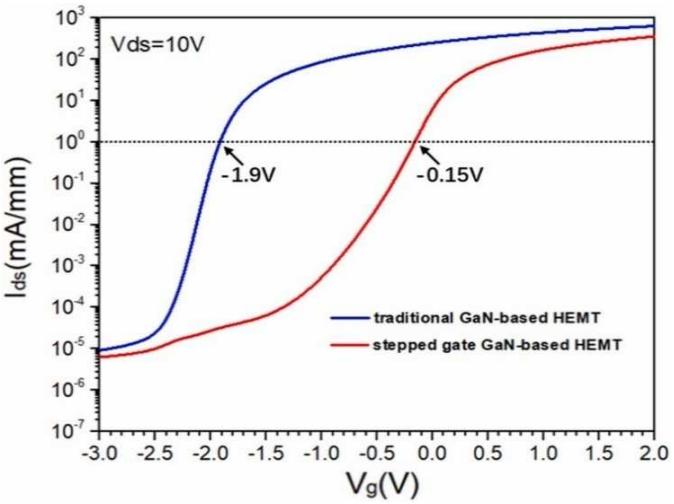



Fig. 4.3. Threshold voltage transfer feature plots for standard GaN-based HEMTs and step-like GaN-based HEMTs **[3].**

### *B. Evaluation and enhancement of the efficiency of GaN-based high speed transistor with a composite step-like design construct*

As observed above, the step-like channel increases the threshold voltage (Vth) but has a minimal impact on the 2DEG concentration. The enhanced step-like channel results in a negative threshold voltage. A hybrid step-like gate GaN-based HEMT (Fig 4.4) could potentially enhance this effect. The incorporation of a A p-GaN cap integrated with a hafnium dioxide dielectric layer contributes to the step-like gate structure. The A p-GaN cap integrated with a hafnium dioxide dielectric layer reduces the 2DEG concentration in the channel, while their combined voltage-sharing effect elevates the threshold voltage.

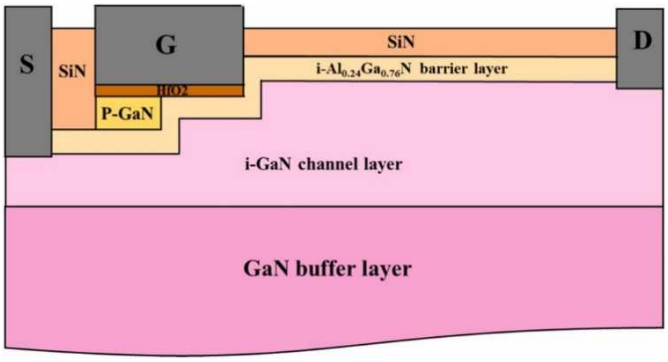



Fig.4.4. Partial cutaway perspective of the composite step-like gate GaN based transistor [3].



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:13 UTC from IEEE Xplore. Restrictions apply.

Fig 4.5 illustrates that the composite GaN-based HEMT with a step-like gate design structure achieves a threshold voltage of 3.5V, which is 3.65V greater than that of a typical HEMT. Even with a higher saturation current of 1μA/mm, the composite design sustains a threshold voltage close to 3V. The p-GaN layer's low hole concentration of 2 $\times 10^{17}$ $cm^{-3}$ enhances the threshold voltage while maintaining the crystal's structural integrity.

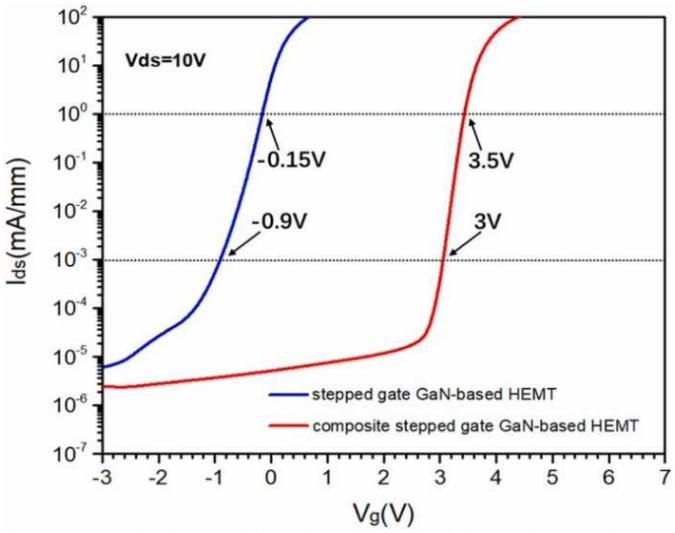



Fig. 4.5. Transfer characteristics of step-like gate design and composite gate GaN based transistor [3].

The p-GaN cap layer plays a critical role in reducing the channel's electron concentration, creating a high-impedance charge accumulation region in the absence of an applied voltage. The hole density in the p-GaN layer directly influences the threshold voltage (Vth) of the GaN-based transistors with a composite step-like gate design HEMT. A higher hole concentration increases Vth by requiring a larger gate voltage to form the 2DEG, as more electrons are depleted from the channel. But this also results in a suppress in output current, which can restrict the device's ability to produce large amounts of electricity. In order to maximize the performance of enhancement-mode HEMTs, the hole density in the p-GaN layer must be carefully chosen. 2 × $10^{17}$ $cm^{-3}$ is the optimal hole concentration for composite GaN-based transistors with a step-like gate design HEMT. [3].

*C. Evaluation and enhancement of with A buried p-GaN the composite GaN-based transistors with a step-like gate design construct and a charge-balancing plate*

This study uses a charge-balancing plate and a buried p-GaN structure to create GaN-based composite transistors with a step-like gate design (Fig 4.6). By incorporating these features, the device may reach a breakover voltage of 2440V while maintaining a current of 1 mA/mm (Fig 4.6). Conversely, transistors with a composite, step-like gate design without these improvements have a breakover voltage of 771V, which is an improvement of 216.5%. This indicates how the buried p-GaN layer and charge balancing plate significantly increase the device's breakover voltage. Additionally, the composite step-like gate HEMT with these changes achieves a soft breakover voltage of 1782V at 10 μA/mm, which is higher than 584V in the absence of these advancements. This further highlights the role of the charge balancing plate and buried p-GaN layer in boosting the breakover performance of the device. The buried p-GaN layer effectively lowers the Electron charge density within the buffer and channel layers in absence of bias. The high hole concentration facilitates recombination with electrons at the p-n junction formed by buried p-GaN layer and the buffer and channel layers, establishing a resistive space charge area. This significantly reduces parasitic current. Additionally, the buried p-GaN layer minimizes impact ionization, which in turn suppresses avalanche breakover and enhances the breakover voltage of the device.

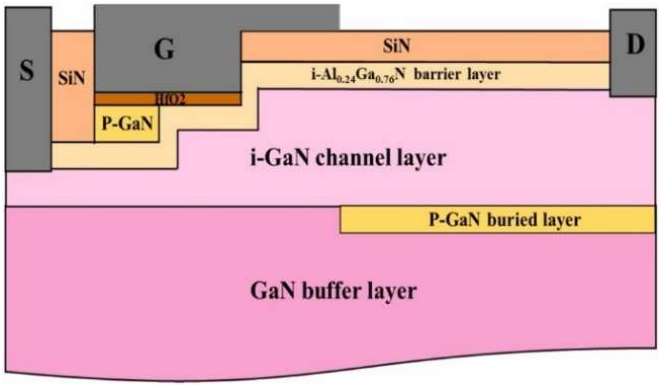



Fig.4.6. A sectional illustration of the GaN-based transistors with a step-like gate design structure and a charge-balancing plate [3].

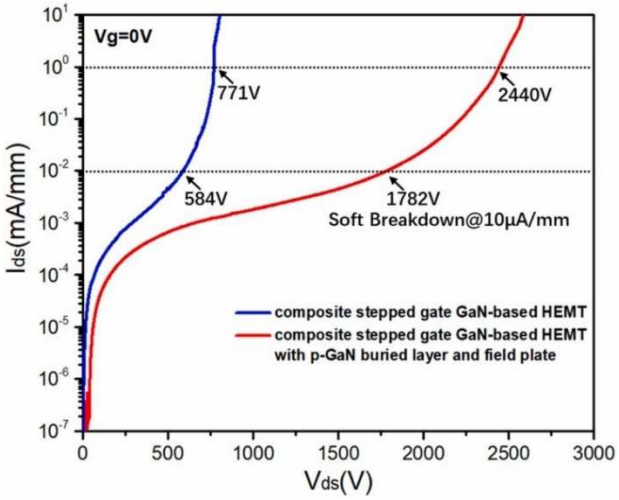



Fig. 4.7. Comparison of the breakover performance between the composite GaN-based transistors with a step-like gate design and the enhanced version incorporating a buried p-GaN layer and charge - balancing gate [3].



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:13 UTC from IEEE Xplore. Restrictions apply.

The design of a composite step-like gate which incorporates a buried p-GaN structure and a charge-balancing plate, achieves a substantial increase in threshold voltage (Vth) to 3.5V by integrating the step-like channel with the composite step-like gate construct. Furthermore, the charge balancing plate and buried p-GaN layer improve electric field homogeneity and lessen buffer layer leakage, which results in a notable increase in breakdown voltage of 2440V [3].

## III. CONCLUSION

The addition of charge balancing plate and buried p-GaN layers further increases the breakover voltages (Vbr) of the buried gate ALGaN/GaN MSHEMTs, which is already 771v. By improving gate control and suppressing parasitic currents, this design raises the frequency overall [1].

The enhancement-mode AlGaN/AiN/GaN HEMTs is designed for enhancement-mode operation, achieving a positive threshold voltage (Vth) that enhances safety in power applications. The composite buried gate structure allows for improved channel control, resulting in a higher Vth and reduced risk of unintended conduction [2].

The enhancement-mode GaN-based transistors with a composite, step-like gate design featuring a buried p-GaN structure and a charge-balancing plate demonstrate exceptional performance with a breakover voltage of 2440V and a threshold voltage (Vth) of around 3.5V. The inclusion of the p-GaN buried layer and charge balancing plate effectively lowers unintended currents and optimizes electric intensity allocation, making this device an excellent choice for high-power uses [3].

Based on the comparative analysis of the three device architectures, the enhancement -mode GaN-based transistors with a composite, step-like gate design equipped with a buried p-GaN structure, and a charge-balancing plate emerges as the best option. Its superior breakover voltage of 2440V, enhanced threshold voltage of 3.5V, and minimized parasitic currents position as the most effective choice for demanding power electronics uses, ensuring both performance and reliability

## REFERENCES


[1] T.-L. Wu, S.-W. Tang, and H.-J. Jiang, "Investigation of Recessed Gate AlGaN/GaN MIS-HEMTs with Double AlGaN Barrier Designs toward an Enhancement-Mode Characteristic," *Micromachines*, vol. 11, no. 2, p. 163, Feb. 2020, doi: https://doi.org/10.3390/mi11020163

[2] J. Li *et al.*, "Normally-off AlGaN/AlN/GaN HEMT with a composite recessed gate," *Superlattices and Microstructures*, p. 107064, Oct. 2021, doi: https://doi.org/10.1016/j.spmi.2021.107064.

[3] J. Li *et al.*, "2.57GW/cm2 normally-off composite stepped gate GaN-based HEMT with p-GaN buried layer and field plate," *Materials Science in Semiconductor Processing*, vol. 153, pp. 107152–107152, Jan. 2023, doi: https://doi.org/10.1016/j.mssp.2022.107152 .

[4] N. M. Shrestha, Y. Li, T. Suemitsu, and S. Samukawa, "Electrical Characteristic of AlGaN/GaN High-Electron-Mobility Transistors With Recess Gate Structure," *IEEE Transactions on Electron Devices*, vol. 66, no. 4, pp. 1694–1698, Apr. 2019, doi: https://doi.org/10.1109/ted.2019.2901719.

[5] G. Greco, Ferdinando Iucolano, and F. Roccaforte, "Review of technology for normally-off HEMTs with p-GaN gate," vol. 78, pp. 96–106, Oct. 2017, doi: https://doi.org/10.1016/j.mssp.2017.09.027 .

[6] Y. Zhang *et al.*, "High-Performance AlGaN Double Channel HEMTs with Improved Drain Current Density and High Breakdown Voltage," *Nanoscale Research Letters*, vol. 15, no. 1, May 2020, doi: https://doi.org/10.1186/s11671-020-03345-6 .

[7] K. Kumar, A. Kumar, V. Kumar, and S. Chander, "International Journal of Electronics and Communications Band gap and gate underlap engineered novel Si 0 . 2 Ge 0 . 8 / GaAs JLTFET with dual dielectric gate for improved wireless applications," *AEUE - Int. J. Electron. Commun.*, vol. 166, no. February, p. 154671, 2023, doi: 10.1016/j.aeue.2023.154671.

[8] H. -C. Chiu *et al*., "High-Performance Normally Off p-GaN Gate HEMT With Composite AlN/Al0.17Ga0.83N/Al0.3Ga0.7N Barrier Layers Design," in *IEEE Journal of the Electron Devices Society*, vol. 6, pp. 201-206, 2018, https://doi.org/10.1109/JEDS.2018.2789908 .

[9] kexiu dong, Y. Zhang, and B. Wang, "Ultra-thin double barrier AlGaN/GaN high threshold voltage HEMT with graded AlGaN/Si3N4 gate and p-type buffer layer," Jul. 2022, doi: https://doi.org/10.21203/rs.3.rs-1787077/v1 .

[10] K. Kumar, A. Kumar, V. Mishra, and S. Chandra, "Implementation of Band Gap and Gate Oxide Engineering to Improve the Electrical Performance of SiGe / InAs Charged Plasma - Based Junctionless - TFET," *Silicon*, no. 0123456789, 2022, doi: 10.1007/s12633-022-02111-w.

[11] L. Geng, H. Zhao, T. Han, and X. Ren, "The lattice-matched AlInN/GaN high electron mobility transistor with BGaN buffer," *Solid State Communications*, vol. 337, p. 114449, Oct. 2021, doi: https://doi.org/10.1016/j.ssc.2021.114449 .

[12] X. Zhang, L. Wang, W. Li, L. Tao, and X. Chen, "Gallium Nitride Dual Two - Dimensional Electron Gas HEMT with a Good Performance: Based on TCAD simulations," pp. 1–4, Aug. 2020, doi: https://doi.org/10.1109/icept50128.2020.9202882 .

[13] H.-C. Chiu, Y.-S. Chang, B.-H. Li, H.-L. Kao, C.-W. Hu, and R. Xuan, "High-Performance Normally Off p-GaN Gate HEMT With Composite $AlN/Al_{0.17}Ga_{0.83}N/Al_{0.3}Ga_{0.7}N$ Barrier Layers Design," *IEEE Journal of the Electron Devices Society*, vol. 6, pp. 201–206, Jan. 2018, doi: https://doi.org/10.1109/jeds.2018.2789908

[14] T. Hashizume *et al.*, "$Al_2O_3$ Insulated-Gate Structure for AlGaN/GaN Heterostructure Field Effect Transistors Having Thin AlGaN Barrier Layers," *Japanese Journal of Applied Physics*, vol. 43, no. 6B, pp. L777–L777, May 2004, doi: https://doi.org/10.1143/jjap.43.l777

[15] Z. Chen *et al.*, "Research on a High-Threshold-Voltage AlGaN/GaN HEMT with P-GaN Cap and Recessed Gate in Combination with Graded AlGaN Barrier Layer," *Journal of Electronic Materials*, vol. 53, no. 5, pp. 2533–2543, Mar. 2024, doi: https://doi.org/10.1007/s11664-024-10968-3




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:53:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

# HEP-DHMI: The High Efficiency and Payload Data Hiding Scheme for AMBTC Compressed Medical Images

**Chia-Chen Lin[1], En-Ting Chu[2], Iuon-Chang Lin[3], Rajeev Kumar[4]**

[1] Department of Information Computer Science and Information Engineering, National Chin-Yi University of Technology, No.57, Sec. 2, Zhongshan Rd., Taiping Dist., Taichung 411030, Taiwan (R.O.C.) (e-mail: ally.cclin@ncut.edu.tw, orcid: 0000-0003-4480-7351)

[2] Prospective Technology of Electrical Engineering and Computer Science, National Chin-Yi University of Technology, No.57, Sec. 2, Zhongshan Rd., Taiping Dist., Taichung 411030, Taiwan (R.O.C.) (e-mail: corpse0712@gmail.com, orcid: 0009-0000-4509-6024)

[3] Department of Management Information Systems, National Chung Hsing University, 145 Xingda Road, South Dist., Taichung City 402, Taiwan (R.O.C.)

[4] Department of Computer Science and Engineering, Delhi Technological University, Delhi, India (rajeevkumar@dtu.ac.in, orcid: 0000-0002-5000-7644)

Corresponding author: Iuon-Chang Lin (e-mail: iclin@nchu.edu.tw).

**ABSTRACT** Current AMBTC-based data hiding schemes generally exhibit strong performance with general images, but their effectiveness varies significantly when applied to medical images in terms of hiding capacity, stego image quality, and efficiency. Moreover, maintaining structural similarity between the original and reconstructed images remains challenging due to the lossy nature of AMBTC compression. Given the increasing adoption of telemedicine applications, especially in the wake of Covid-19, where Internet infrastructure varies across regions, AMBTC's efficiency and ease of implementation make it well-suited for such applications. To enhance hiding capacity while preserving image quality and structural similarity, we propose a high-efficiency, high-payload data hiding scheme tailored for AMBTC-based medical images, termed HEP-DHMI. This approach involves classifying AMBTC compressed blocks into four types—flat, smooth, moderate, and complex—and designing distinct data hiding strategies for each block type. Experimental results validate that HEP-DHMI achieves an average PSNR of 34.25 dB, a payload capacity of 245,788 bits, an efficiency rating of 45.22, and an M-SSIM of 0.98 when evaluated on images selected from two prominent databases: Covid-19 and Brain Tumor MRI. These results demonstrate superior performance compared to existing schemes.

**INDEX TERMS** Absolute moment block truncation coding, data hiding, improved hiding capacity, enhanced visual quality, Hamming code (7,4)

## I. INTRODUCTION

Due to the widespread adoption of the Internet and advancements in digital processing, an increasing number of individuals opt for online communication for personal or business purposes. However, this transition carries significant implications and poses potential risks. The inherent openness of the Internet renders transmitted data susceptible to interception and unauthorized tampering, leading to potential information breaches. Consequently, there has been a notable surge in academic and industrial interest towards enhancing information security through technologies like data encryption and concealment.

Data hiding technology involves concealing data within multimedia objects like images, audio, and videos. This method employs algorithms to embed data seamlessly into these media without any noticeable alterations. For instance, in the case of images, data hiding strives to make the stego image indistinguishable from the original to the human eye. This feature ensures that multimedia objects containing sensitive information do not readily attract malicious attention, thereby safeguarding the confidentiality of the hidden data.

In the domain of image hiding, data hiding techniques are categorized into two main groups: irreversible data hiding methods [1-9] and reversible data hiding methods also called RDH [10-22]. Irreversible methods are typically employed in contexts with lower security requirements as they embed secret data into cover media but cannot fully restore the cover



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



media after data extraction. These methods are commonly used in fragile watermarking to maintain content integrity. An exemplary technique is the Least Significant Bit Substitution method [1], which adjusts the least significant bits of image pixels for data embedding.

Conversely, RDH permits the lossless extraction of embedded data and full restoration of the cover image without compromising its image quality. However, the extraction phase of RDH tends to be more complex, rendering RDM suitable for high-security fields like military and medical imaging. RDH not only conceals confidential data and ensures similarity between the original and stego media but also guarantees reversibility for both hidden data and original media. Notable RDH techniques include Pixel-error Expansion (PEE) [20], which achieves high-quality data embedding by predicting error values between pixels, and Pixel Value Differencing (PVD) [3], which embeds data by adjusting the difference values between adjacent pixels, difference expansion (DE) [10], histogram shifting (HS) [22], and interpolation technology (IT). PEE [20] achieves high-quality data embedding by predicting error values between pixels. PVD [3] embeds data by adjusting the difference values between adjacent pixels. DE [10] embeds 1-bit of secret data into two neighboring pixels in a cover image by expanding the difference between two neighboring pixels. Later, the auxiliary data is recorded to solve the overflow and underflow problem and ensure reversibility. HS [22] hides secret data by choosing the combination of peak and zero-point pixels from the histogram of a cover image. Then, the secret data is hidden in peak pixels after shifting non-zero-point pixels either by adding 1 or subtracting 1.

Besides the spatial domain of images, scholars also consider the compression codes generated by image compression algorithms as the carrier to expand the candidates of the carrier in data hiding. Among various image compression algorithms, Absolute Moment Block Truncation Coding (AMBTC) [18, 25-29] is one of the popular algorithms whose compression codes have widely played the carrier in different irreversible/reversible data hiding methods over the last two decades because AMBTC's computation is simple and its performance is efficient.

In 2015, Ou et al. [24] introduced an AMBTC-based data hiding scheme, where blocks are categorized as smooth or complex based on the difference between two quantizers, compared to a predefined threshold. In their scheme, smooth blocks replace the bitmap with 16 bits of embedded secret data, while complex blocks embed 1-bit secret message by swapping quantizers for "1" or leaving them unchanged for "0". In 2019, Kumar et al. [25] proposed an improved scheme for classifying blocks into smooth, less complex, and highly complex categories, leveraging the quantizers' differences and two thresholds. In their scheme, smooth blocks embed 16-bit secret data directly, less complex blocks use Hamming distance for concealing 8-bit secrets, and highly complex blocks adjust quantizers using the PVD method to embed varying bit lengths for optimal payload and image quality. In the same year, Chang et al. [26] introduced a scheme for AMBTC-compressed images categorizing blocks as smooth, less complex, or highly complex based on quantizers and thresholds. In their scheme, smooth blocks embed 16-bit secret data, less complex blocks use Hamming codes to modify a divided bitmap for concealing 6-bit secrets, and highly complex blocks embed 2-bit secrets in the least significant bits of quantizers. In 2023, Chen et al. [27] proposed a gradient-based scheme by classifying blocks into gradient and non-gradient types using gradient consistency ($Gc$), horizontal difference ($Hd$), and vertical difference ($Dv$). Gradient blocks embed 13-17 bits in compressed encoding, while non-gradient blocks employ AMBTC compression with further categorization into low-complexity (embedding 2-bits) and high-complexity (embedding 1-bit by swapping quantization values) types. This scheme achieves an average embedding capacity of approximately 140,000 bits and image quality of around 31 dB for PSNR, with efficiency around 36. For complex images, capacity and quality decrease to 110,000 bits and 30 dB.

In their experiments, Chen et al. [27] used medical images as test images, demonstrating that the PSNR of these images ranged from 29.73 to 31.07 dB with a threshold set to 6. Increasing the threshold to 8 allowed for an improved payload of up to 136,810 bits while maintaining similar image quality. In the same year, Lai et al. introduced a hybrid data hiding method that combines AMBTC and LSB substitution [28]. They asserted that their approach achieves a high payload, averaging around 224,806 bits with a quality of 30.395 dB when the threshold is set to 8. Following this, Al-Chaab et al. investigated ways to securely exchange medical images while minimizing transmission size [30]. Their experiments showed a 30% reduction in final transmission size through the use of compressive sensing and LSB audio steganography. In 2024, Lin et al. [29] and Qu et al. [21] employed secret sharing or a mix of data hiding and encryption to bolster the security of medical images in various application scenarios. These studies indicate a growing interest in medical image security, with AMBTC-based hybrid data hiding emerging as a key strategy for safeguarding the confidentiality of transmitted data. Despite these advancements, further improvements are still feasible. Given the rising demand for medical image applications, this paper aims to introduce a high-efficiency and payload reversible data hiding method suitable for AMBTC-compressed medical images called HEP-DHMI, aiming to enhance effective payload and image quality for both general and medical images. Our approach draws inspiration from Kumar et al.'s scheme [25], but we have enhanced it further. Specifically, our scheme classifies all blocks in a cover image into four types: flat block, smooth block, moderate block, and complex block, based on pixel value differences. Each block type undergoes a specific embedding algorithm tailored to optimize payload and image quality.



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



The structure of the remainder of this paper is as follows: Section II introduces the related methods. Section III provides a detailed description of the designed embedding algorithms, along with their corresponding extraction and reconstruction algorithms, for the proposed HEP-DHMI. Section IV presents comprehensive experimental results and comparisons with other methods. Finally, Section V concludes the paper with our findings and implications.

## II. RELATED WORK

### A. ABSOLUTE MOMENT BLOCK TRUNCATION CODING

Absolute Moment Block Truncation Coding also known as AMBTC is a lossy compression technique designed for digital images, proposed by Lema et al. [23]. AMBTC divides an image into multiple non-overlapping blocks of size 4×4 pixels. Each block's pixels are then simplified into two quantizers $avgH$, $avgL$ along with a bitmap to achieve image compression.

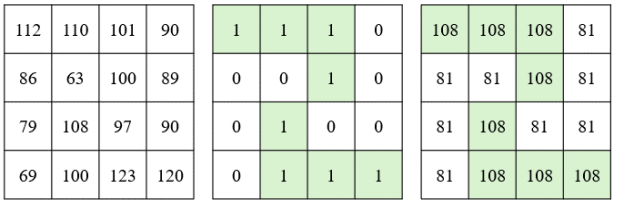



**FIGURE 1. Example of AMBTC compression.**

To clearly illustrate this process, Fig. FIGURE 1. provides an example of AMBTC compression. First, the mean value $avgB_k$ of the $k$-th block in an image is calculated. Pixels within the $k$-th block in an image are then classified based on $avgB_k$. Pixels greater than or equal to $avgB_k$ belong to the high-mean category, while those below $avgB$ belong to the low-mean category. The mean value $avgB_k$ for $k$-th block in an image is calculated using Eq. (錯誤! 找不到參照來源。)

$$avgB_k = \frac{1}{16}\sum_{i=1}^{16} p_i, \tag{1}$$

where $p_i$ represents the $i$-th pixel of the $k$-th block $B$ in an image. After classifying the pixel, each category within the $k$-th block is simplified into a high mean ($avgH_k$) and a low mean ($avgL_k$). The mean values for each category are calculated using Eqs. (錯誤! 找不到參照來源。) and (錯誤! 找不到參照來源。).

$$avgH_k = \frac{1}{16-n}\sum_{p_i \geq avgB_k} p_i, \tag{2}$$

$$avgL_k = \frac{1}{n}\sum_{p_i < avgB} p_i, \tag{3}$$

where $n$ is the number of pixels in the low-mean category. Next, the positions of pixels that belong to the $avgH_k$ and $avgL$ are recorded as "1" and "0" in the bitmap, respectively. Finally, the AMBTC compression code for $k$-th 4×4 block is represented as ($avgL_k$, $avgH_k$, $BM_k$). In other words, using AMBTC, a 4×4 image block sized as at 16×8 (=128) bits is reduced to 4×8 (=32) bits.

### B. KUMAR ET AL.'S AMBTC-BASED DATA HIDING METHOD

In 2019, Kumar et al. proposed a data hiding method based on AMBTC [25]. This method embeds data by adjusting the two quantizers and bitmap $BM$ of each AMBTC block. The method classifies AMBTC blocks into three types: smooth block, low-complexity block, and high-complexity block, based on the difference $D$ between the two quantizers and the two thresholds $thr1$ and $thr2$.

If $D < thr1$, the block is classified as a smooth block, and 16-bit of secret data are embedded by completely replacing the bitmap of the smooth block. If $thr2 \geq D \geq thr1$, the block is classified as a low-complexity block. In this case, 8-bit of secret data are embedded by replacing the odd and even rows of the bitmap $BM$ with the data. Hamming distances $h1$ and $h2$ are then calculated. If $h1 < h2$, it indicates that embedding data into the odd rows of the bitmap results in a higher similarity to the original bitmap, so the data is embedded into the odd rows. Conversely, if $h1 \geq h2$, embed the data into the even rows of the block and swap the positions of two quantizers, $avgH$ and $avgL$, in the block.

If $D > thr2$, the block is classified as a high-complexity block. In this case, the Pixel Value Differencing (PVD) [3] is used to embed 3-7 bits of secret data. Kumar et al.'s defined Table. 0 to search for the embedding capacity based on the difference $D$.

**TABLE 1. Pixels value embedding capacity based on pixel difference range**

| Difference range | 0~7 | 8~15 | 16~31 | 32~63 | 64~127 | 128~255 |
|---|---|---|---|---|---|---|
| Hiding capacity (bits) | 3 | 3 | 4 | 5 | 6 | 7 |

## III. PROPOSE HEP-DHMI SCHEME

After examining the method introduced by Kumar et al. [25], we found that adding block classification to an image can boost the payload while maintaining competitive image quality. However, the effectiveness of their approach in terms of payload and image quality varies with medical images, with image quality ranging from 27.82 dB to 31.33 dB. Consequently, this paper aims to develop various data hiding strategies tailored to different block types and then achieve more consistent performance in image quality, payload capacity, and efficiency when handling medical images.

To achieve our objective, the proposed HEP-DHMI classifies image blocks into four categories: flat, smooth, moderate, and complex. For each block category, we propose specific embedding algorithms to enhance payload capacity while preserving high image quality and efficiency, the



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



flowchart of HEP-DHMI is shown in Fig. FIGURE 1. . The proposed scheme is organized into three phases: embedding, extraction, and reconstruction, which are detailed in Sections III.A to III.C, respectively.

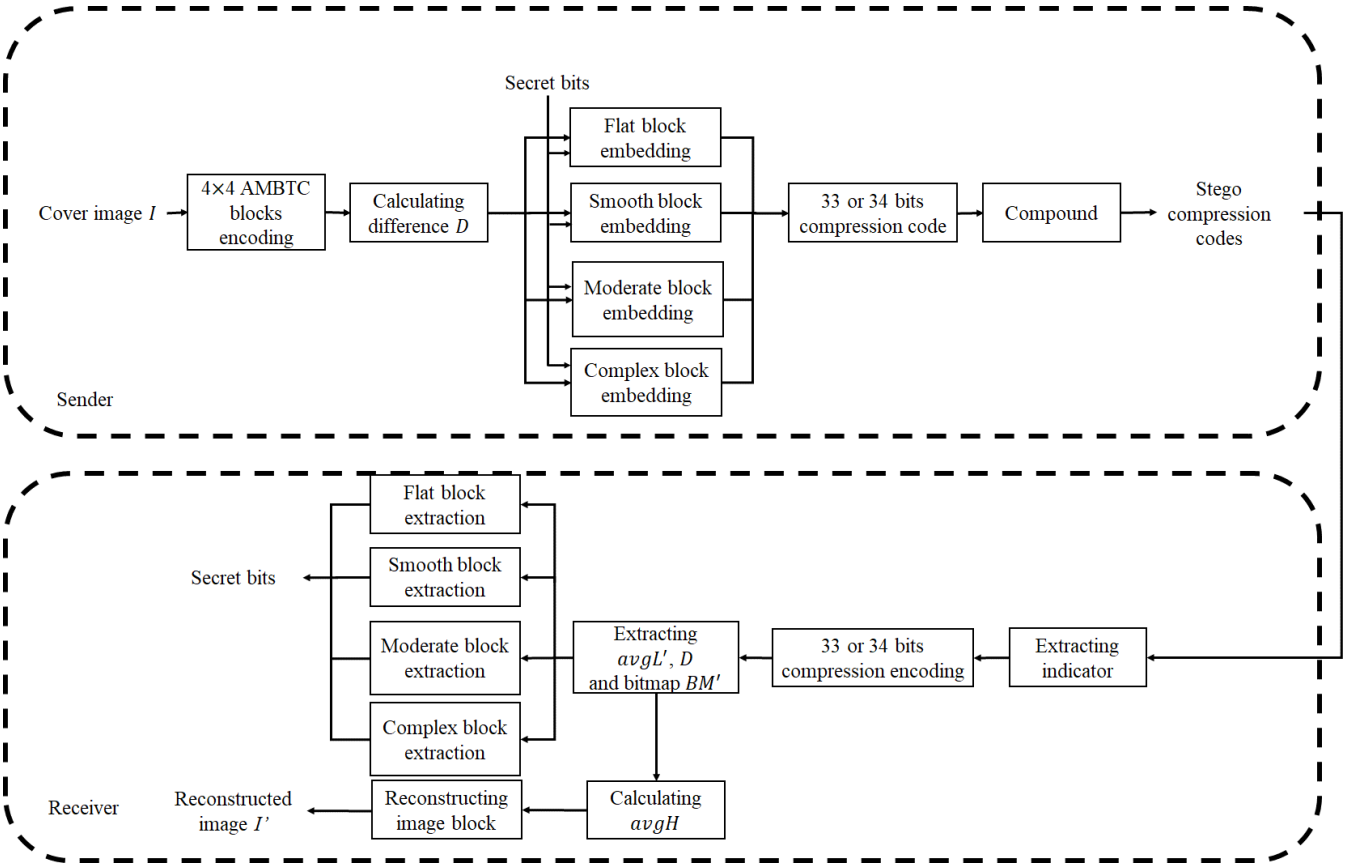



**FIGURE 2. Flowchart of the proposed HEP-DHMI.**

In Section II, it was previously noted that each 4×4 non-overlapping block within AMBTC compressed images includes two quantizers, $\boldsymbol{avgH}$ and $\boldsymbol{avgL}$, along with a bitmap $\boldsymbol{BM}$. It has been observed that for a 4×4 AMBTC compressed block, $\boldsymbol{avgH}$ can always be reconstructed from $\boldsymbol{avgL}$ and $\boldsymbol{D}$, where $\boldsymbol{D}$ represents the difference between $\boldsymbol{avgL}$ and $\boldsymbol{avgH}$. Consequently, the proposed HEP-DHMI records $\boldsymbol{avgL}$ and $\boldsymbol{D}$ instead of both $\boldsymbol{avgH}$ and $\boldsymbol{avgL}$, as conventional AMBTC does. Ultimately, using HEP-DHMI, the stego AMBTC compression code for each block is represented as [*avgL*, *D*, *BM*], with a size of either 33 or 34 bits.

### *A. EMBEDDING PHASE*

As previously mentioned at the beginning of Section III, each 4×4 block in AMBTC compressed images has been changed as $[avgL, D, BM]$ in the HEP-DHMI instead of $[avgL, avgH, BM]$, as conventional AMBTC does. Later, during the data embedding phase of the HEP-DHMI, two extra information: the *indicator* defined in Eq. (錯誤! 找不到參照來源。) that presents the corresponding block type and $data$ that indicates the hidden secret message will be added. Finally, the AMBTC compression code is stored in the format of $[\,indicator\,,\,avgL\,,\,D\,,\,data\,,\,BM\,]$ or $[indicator, avgL, D, BM]$:

$$(block, indicator) = \begin{cases} (flat, 0), & if\ D \le 4 \\ (smooth, 0), & if\ 4 < D \le 14 \\ (moderate, 10), & if\ 14 < D \le 59 \\ (complex, 11), & if\ 59 < D \end{cases} \quad (4)$$

The detailed descriptions of the embedding algorithm of HEP-DHMI are listed below:

**Algorithm: HEP-DHMI Embedding**

**Input:** An image *I*, secret bits
**Output:** Stego AMBTC compression codes

Step 1. Divide an image into 4×4 non-overlapping blocks.
Step 2. For each block, conduct AMBTC encoding and generate the AMBTC compression code $[avgL, avgH, BM]$.



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



Step 3. Calculate the difference $D$ between the two quantizers $D = |avgH - avgL|$.
Step 4. Based on the difference $D$, classify the block into one of four categories according to Eq. (錯誤! 找不到參照來源。)
Step 4.1. If the current block belongs to a flat block, its indicator is defined as "0" according to Eq. (錯誤! 找不到參照來源。). The format of compressed code is defined as $[indicator, avgL, D, data, BM]$, and conduct flat block embedding strategy which will be described in Subsection III.A.1.
Step 4.2. If the current block is determined as a smooth block, its indicator is defined as "0" according to Eq. (錯誤! 找不到參照來源。). The format of compressed code is defined as $[indicator, avgL, D, data, BM]$, and conduct smooth block embedding strategy which will be described in Subsection III.A.2.
Step 4.3. If the current block is determined as a moderate block, its indicator is defined as "10" according to Eq. (錯誤! 找不到參照來源。). The format of compressed code is defined as $[indicator, avgL, D, data, BM]$, and conduct moderate block embedding strategy which will be described in Subsection III.A.3.
Step 4.4. If the current block is a complex block, its indicator is defined as "11" according to Eq. (錯誤! 找不到參照來源。). The format of compressed code is defined as $[indicator, avgL, D, BM]$, and conduct complex block embedding strategy which will be described in Subsection III.A.4.
Step 5. Output the stego compression code derived from Step 4.
Step 6. If the current block is not the last one in the image, read the next block and repeat Steps 4-5 to perform the corresponding embedding operation according to its block type.
7. Collect and contact the stego compression codes generated at each step to obtain the final AMBTC stego compression codes.

### 1) FLAT BLOCK EMBEDDING

For a flat block, the stego compression code is represented as $SBC$ and $SBC$= [$indicator$ (1 bit), $avgL$ (8 bits), $D$ (4 bits), $data$ (4 bits), $BM$ (16 bits)]. Fig. FIGURE 3. (a) illustrates the detailed embedding operation for the flat block. First, read 20 bits from the secret bits, and record 4 secret bits in the $data$. The remaining 16 secret bits replace the original AMBTC bitmap $BM$. Finally, set the $indicator$ as 0 and output $SBC$ as shown in Fig. 3(b).

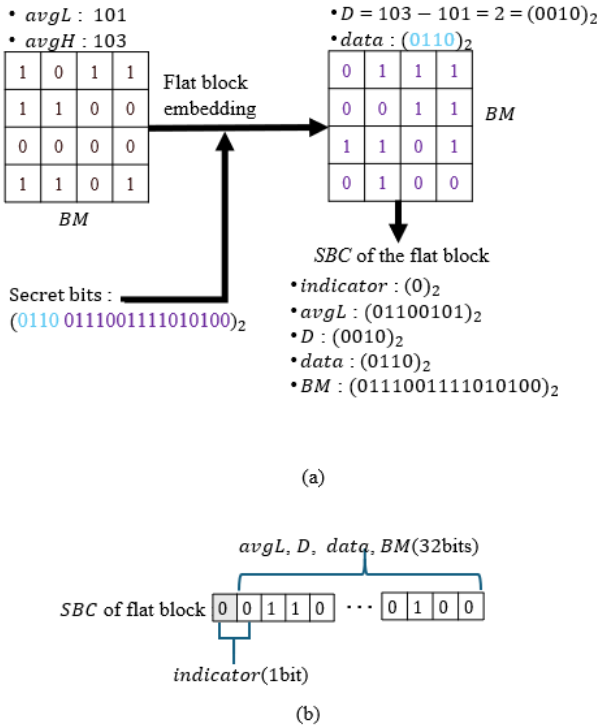



**FIGURE 3. Example of the flat block embedding: (a) embedding procedure (b) stego block compression code ($SBC$).**

### 2) SMOOTH BLOCK EMBEDDING

For the smooth blocks, the stego compression code is constructed as $\boldsymbol{SBC}$ = [$\boldsymbol{indicator}$ (1 bit), $\boldsymbol{avgL}$ (8 bits), $\boldsymbol{D}$ (4 bits), $\boldsymbol{data}$ (4 bits), $\boldsymbol{BM}$ (16 bits)]. To conduct the embedding strategy, First, the $\boldsymbol{BM}$ is divided into two parts as shown in Fig. 4, The first seven bits of $\boldsymbol{BM}$ those framed with the red line form the first part, and the last eight bits those framed with the blue line form the second part. The smooth block embedding is then illustrated in Fig. 5.

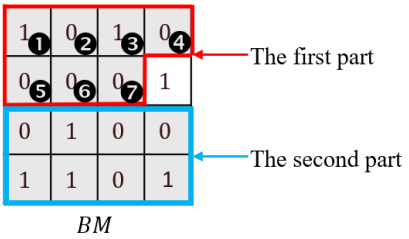



**FIGURE 4. Example of two parts of $\boldsymbol{BM}$.**



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



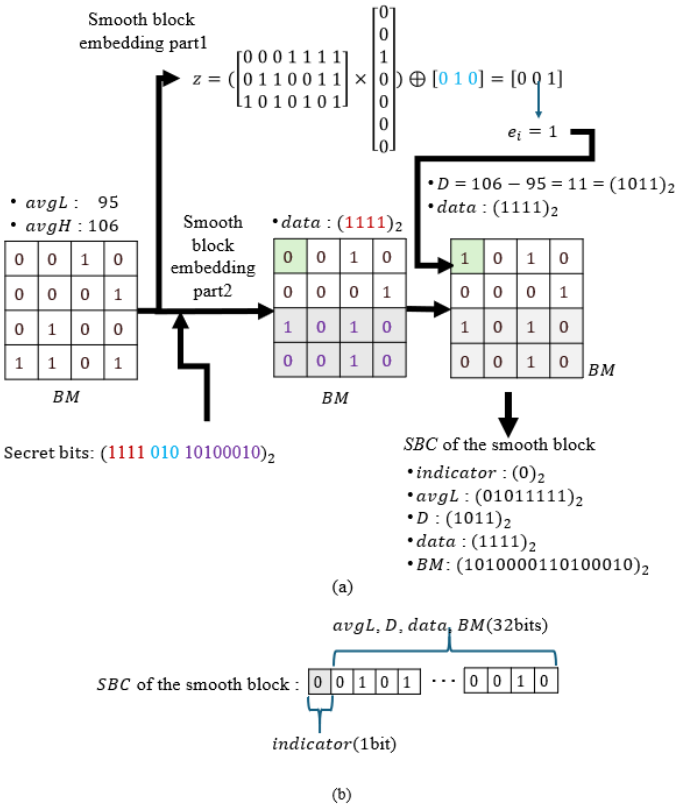



**FIGURE 5. Example of the smooth block embedding: (a) embedding procedure (b) stego block compression code ($SBC$).**

Next, read 15 bits from the secret bits and record the first 4 secret bits into the ***data*** field. Next, read the following 3 secret bits and carry these 3 secret bits with the first part of *BM* by using Hamming code (7,4) defined in Eq. (**錯誤! 找不到參照來源。**). Finally, the second part of *BM* carries 8 secret bits by direct replacement. The corresponding ***indicator*** is set as "0" and output ***SBC***.

$$z = (M \times B^T)^T \oplus s,$$
$$M = \begin{bmatrix}0&0&0&1&1&1&1\\0&1&1&0&0&1&1\\1&0&1&0&1&0&1\end{bmatrix}. \quad (4)$$

where $B$ is the first 7-bit element located at the first part of $BM$, $s$ is the 3-bit secret data to be embedded via Hamming code (7,4), $M$ is the parity-check matrix and z is the syndrome vector. z is used to indicate whether an error has occurred, and if so, for which bit position in the first part of *BM*. Later, map z to Table. **錯誤! 找不到參照來源。** to find the corresponding $e_i$ which indicates the position should be flipped in the first part of *BM* to carry the 3-bit. It is important to note that $e_i$ listed in Table 2 corresponds to the number indicated in the first part of the *BM* shown in Fig. 5, which means the value of $e_i$ ranges from 1 to 7. Once the specific position $e_i$ is identified, Eq. (**錯誤! 找不到參照來源。**) is used to flip the corresponding bit.

**TABLE 2. Hamming table indicating the position for adjustment**

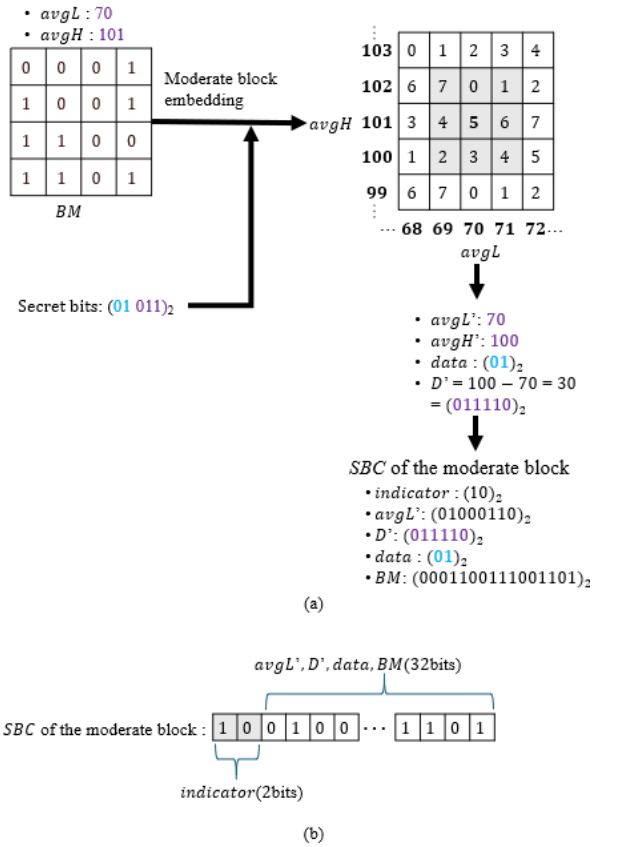
| z | $(0\,0\,1)_2$ | $(0\,1\,0)_2$ | $(0\,1\,1)_2$ | $(1\,0\,0)_2$ | $(1\,0\,1)_2$ | $(1\,1\,0)_2$ | $(1\,1\,1)_2$ |
|---|---|---|---|---|---|---|---|
| $e_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

$$BM' = BM[e_i] \oplus 1. \quad (5)$$

For example, consider Figs. 4 and 5: let $B$=$(0010000)_2$, s=$(010)_2$ and z is obtained as $(001)_2$ with $e_i$=1. This indicates that the first bit in the first part of *BM* needs to be changed to "1." Once the smooth block embedding is complete, the indicator is set to 0, and the *SBC* is generated and ready to be outputted.

### *3) MODERATE BLOCK EMBEDDING*

In the context of the moderate block, the stego compression code is denoted as ***SBC***= [***indicator*** (2 bits), ***avgL*** (8 bits), ***D*** (6 bits), ***data*** (2 bits), ***BM*** (16 bits)]. Fig. FIGURE 6. provides an illustration of the moderate block embedding process. Initially, 5 bits are read from the secret bits. Among these, 2 secret bits are directly recorded into the *data* field. The remaining 3 secret bits are converted into a decimal number, ***DS***. This ***DS*** value is subsequently embedded into the coordinates ***avgH*** and ***avgL*** by referring to the embedding matrix ***EM***, depicted in Fig. FIGURE 7. . Here, ***avgH*** and ***avgL*** respectively correspond to the ***Y*** and ***X*** axes of the embedding matrix ***EM***.






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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



**FIGURE 6. Example of the moderate block embedding: (a) embedding procedure (b) stego block compression code ($SBC$).**

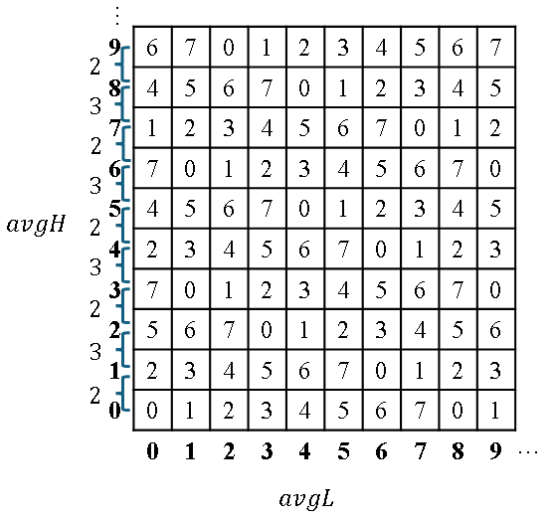



**FIGURE 7. Embedding matrix $EM$.**

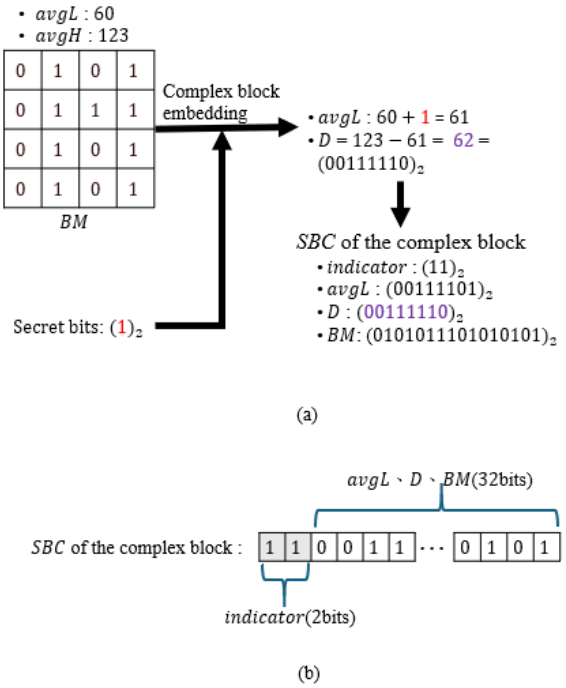



**FIGURE 8. Example of complex block embedding: (a) embedding procedure (b) stego block compression code ($SBC$).**

Start by initializing the search with the pair $(\boldsymbol{avgH}, \boldsymbol{avgL})$ **from** $\boldsymbol{EM}$. Identify a quantizer pair and find a new quantizer pair of ($\boldsymbol{avgH'}$, $\boldsymbol{avgL'}$) that minimizes the Euclidean distance to the original quantizer pair $(\boldsymbol{avgH}, \boldsymbol{avgL})$ while satisfying the condition that the value located at $(\boldsymbol{avgH'}, \boldsymbol{avgL'})$ in $\boldsymbol{EM}$ matches the $\boldsymbol{DS}$ value. Next, construct the $\boldsymbol{SBC}$ = [$\boldsymbol{avgL'}$ , $\boldsymbol{D'}$ = |$\boldsymbol{avgH'}$ - $\boldsymbol{avgL'}$|, $\boldsymbol{data}$, $\boldsymbol{BM}$]. Finally, set the $\boldsymbol{indicator}$ to "10" and output $\boldsymbol{SBC}$.

#### *4)COMPLEX BLOCK EMBEDDING*

In the context of complex block, the compression code for the block is constructed as $\boldsymbol{SBC}$ = [$\boldsymbol{indicator}$ (2 bits), $\boldsymbol{avgL}$ (8 bits), $\boldsymbol{D}$ (4 bits), $\boldsymbol{data}$ (4 bits) and $\boldsymbol{BM}$ (16 bits)]. Fig. FIGURE 8. illustrates the complex block embedding strategy.

For a complex block, only 1 secret bit is carried. LSB is adopted as shown in Eq. (錯誤! 找不到參照來源。) to embed 1 secret bit to $\boldsymbol{avgL}$ , and then the $\boldsymbol{D'}$ is recalculated for the block with the stego $\boldsymbol{avgL'}$ and the original $\boldsymbol{avgH}$.

$$avgL' = avgL - (avgL\%2) + s, \tag{6}$$

where $\boldsymbol{s}$ is the 1-bit data to be embedded. Finally, set the $\boldsymbol{indicator}$ to "11" and output $\boldsymbol{SBC}$.

### *B. EXTRACTION PHASE*

After the recipient receives the stego AMBTC compression codes generated by the sender, the HEP-DHMI extraction algorithm is then conducted. Each block's stego compression code $SBC$ may have a different length ranging between 33 and 34 bits. Eq. (錯誤! 找不到參照來源。) is then used to determine the size of $SBC$ and $D$ according to the received *indicator*. The $avgH$ of each block can be restored by computing $avgH = avgL + D$, and then the corresponding extraction strategies are employed according to the block type. The detailed description of the HEP-DHMI extraction algorithm is listed below:

$$\text{size of } (SBC, D) = \begin{cases} (33,4), & \text{if } indicator = 0 \\ (33,4), & \text{if } indicator = 0 \\ (34,6), & \text{if } indicator = 10 \\ (34,8), & \text{if } indicator = 11 \end{cases}. \tag{7}$$

**Algorithm 2: HEP-DHMI Extraction**

**Input:** Stego AMBTC compression codes
**Output:** Hidden secret bits
Step 1. Read the first bit from the received stego AMBTC compression codes.
Step 1.1. If the read bit is "0" then read the following 32 bits; otherwise, read the following 33 bits.
Step 2. Read $D$ field to classify the current block into one of four categories: flat, smooth, moderate or complex block, based on the difference $D$ values according to Eq. (錯誤! 找不到參照來源。).



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



Step 2.1. If *indicator* = "0" and $D < 5$, the current block is determined as a flat block, and 20 secret bits are carried. The extraction strategy described in Subsection III.B.1 is conducted.
Step 2.2. If *indicator* = "0" and $5 \leq D < 15$, the current block is determined as a smooth block, and 15 secret bits are carried. The extraction strategy described in Subsection III.B.2 is conducted
Step 2.3. If the $indicator$ = "10" and the current block is determined as a moderate block, then it contains 5 secret bits. The extraction strategy described in Subsection III.B.3 is conducted.
Step 2.4. If the $indicator$ = "11" and the current block is determined as a complex block, then it contains 1 secret bits. The extraction strategy described in Subsection III.B.4 is conducted.
Step 3. Output the extracted secret bits.
Step 4. If the current compression code is not the last one in the received final stego compression code, read the next compression code and repeat Steps 2-3 to perform the extraction operation.
Step 5. Collect and contact all the extracted secret bits to obtain the final secret message.

### *1) FLAT BLOCK EXTRACTION*

In the flat blocks, the *data* field contains 4 secret bits, while the stego *BM* hides an additional 16 secret bits. This means there are a total of 20 hidden bits that can be extracted from the stego compression code. Fig. FIGURE 9. illustrates the extraction process for the flat block.

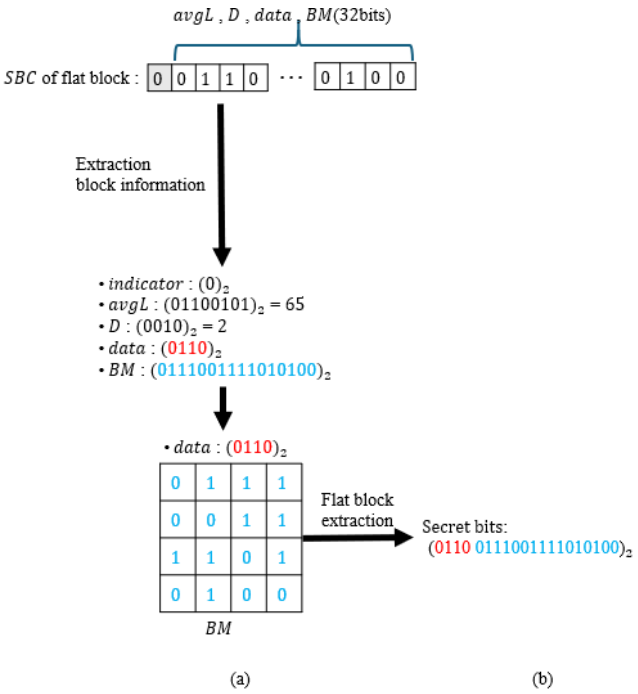



**FIGURE 9. Example of flat block data extraction (a) extraction procedure (b) extracted 20 secret bits.**

### *2) SMOOTH BLOCK EXTRACTION*

In the smooth blocks, the *data* field contains 4 secret bits, while the second part of the stego ***BM*** hides an additional 8 secret bits. The extra 3 secret bits are carried by the first part of the stego *BM* with Hamming code (7,4) by changing a single bit. This means there are a total of 15 hidden bits that can be extracted from the stego compression code. Fig. FIGURE 10. depicts the extraction process of the smooth block. Note that 3 secret bits can be extracted from the first part of the stego $BM$ by using the Hamming code (7,4) defined in Eq. (**錯誤! 找不到參照來源。**).

$$s = (M \times B^T)^T, \text{wherSe } M = \begin{bmatrix} 0\,0\,0\,1\,1\,1\,1 \\ 0\,1\,1\,0\,0\,1\,1 \\ 1\,0\,1\,0\,1\,0\,1 \end{bmatrix}, \tag{8}$$

where $B$ represents the first 7-bit element located at the first part of the stego $BM$, $s$ is the 3-bit hidden secret can be extracted via Hamming code (7,4) and $M$ is the parity-check matrix. Convert $s$ into a 3-bits binary representation to reveal the hidden secret bits. Finally, the remaining 8 hidden secret bits are directly retrieved from the second part of the stego $BM$.

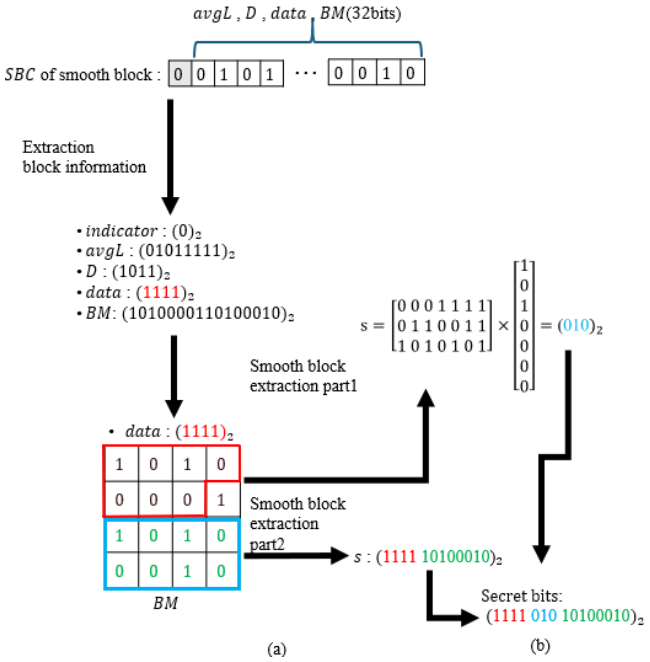



**FIGURE 10. Example of smooth block's data extraction (a) the extraction procedure (b) extracted 15 secret bits.**

### *3) MODERATE BLOCK EXTRACTION*

In moderate blocks, there are 5 hidden secret bits, with two of them being recorded in the *data* field. The remaining 3 secret bits are carried by $avgH'$ and $avgL'$ using embedding matrix $EM$ in the Subsection III.A.3. Fig. FIGURE 11. depicts the extraction process of the moderate block. Note that $avgH'$ and $avgL'$ corresponds to the $Y$ and $X$ axes of embedding matrix $EM$, respectively. To obtain the



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



hidden bit, find $s = EM(avgL', avgH')$ and then convert $s$ into 3-bits binary representation to get the hidden 3-bits secret.

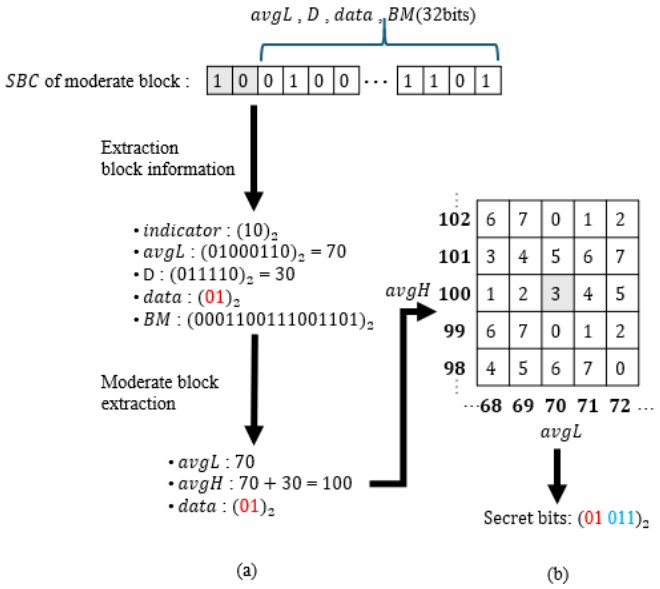



**FIGURE 11. Example of moderate block's data extraction (a) the extraction procedure (b) extracted 3 secret bits.**

#### *4) COMPLEX BLOCK EXTRACTION*

Only 1 secret bit is embedded within the stego compression code for the complex block. Fig. FIGURE 12. illustrates the process of extracting data from a complex block. To retrieve the hidden 1 secret bit, the received $avgL$' is converted to an 8-bits binary representation and then extracts the least significant bit.

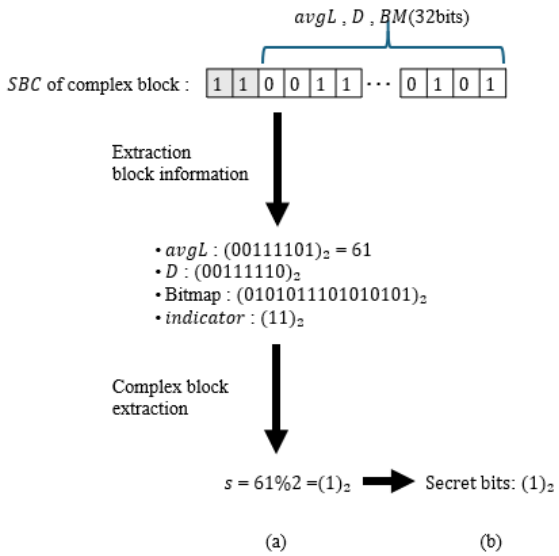



**FIGURE 12. Example of complex block's data extraction (a) the extraction procedure (b) extracted 1 secret bit.**

### *C. IMAGE RECONSTRUCTION*

After retrieving the hidden secret bits, the recipient computes $avgH$ as $avgH = avgL + D$. Using the restored $avgH$, received $avgL$ and $BM$, the recipient is then able tp reconstruct the image following the conventional AMBTC decoding process. In other words, if a bit in the received $BM$ is 1, the pixel value is restored by replacing the corresponding bit value in the received $BM$ with the recalculated $avgH$. Otherwise, the pixel value is restored by replacing the corresponding bit value in the received $BM$ with the received $avgL$. This replacement procedure continues iteratively for all blocks until the entire image has been reconstructed. Once all blocks have been processed according to the received $BM$, the recipient obtains the fully reconstructed image. While the restored image differs from the original, subsequent experiments will demonstrate that the average structural similarity (average SSIM) for medical images reaches as high as 0.9233. Detailed presentations and discussions will be provided in Section IV.

## IV. EXPERIMENTAL RESULTS

In this section, we first demonstrate the performance of the proposed HEP-DHMI scheme in terms of image quality, hiding capacity, and efficiency. Subsequently, comparisons are made between HEP-DHMI and four existing schemes. Subsection IV.A demonstrates the overall performance of HEP-DHMI with 8 general images (shown in Fig. FIGURE 13. ) and 6 medical images (shown in Fig. FIGURE 14. ) selected from the Osirix database [31] as test images. Additionally, to enrich the experimental results on medical images, we selected two different types from the Covid-19 image dataset [32] and Brain Tumor MRI dataset [33], totaling 200 images for testing as shown in Fig.FIGURE 15. . The performance of HEP-HMI with two sets of medical images are reported in Subsection 4.B. Finally, Subsection 4.C reports the comparisons between our scheme and four existing schemes. These images were used as test sets to assess the performance of our scheme in embedding random bit streams generated from a random number generator. The experimental environment involved a Windows 11 computer equipped with an Intel i5-11400H CPU with RTX3060 laptop GPU, 24GB of memory, and Python 3.9.7 was used for experimentation.

Eqs. (錯誤! 找不到參照來源。), (錯誤! 找不到參照來源。), (錯誤! 找不到參照來源。), (錯誤! 找不到參照來源。) and (錯誤! 找不到參照來源。) outline the metrics employed to evaluate the performance of our scheme and four existing methods. PSNR is a standard measure for assessing the image quality of the stego image post data embedding, defined by Eq. (錯誤! 找不到參照來源。). A PSNR exceeding 30 dB typically indicates that the difference between the cover image and the stego image is imperceptible to the human visual system:



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



$$\text{PSNR} = 10\log_{10}\left[\frac{(255)^2}{MSE}\right], \tag{9}$$

where MSE is the Mean Squared Error defined as:

$$\text{MSE} = \frac{1}{M * N}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(p_{ij} - p'_{ij}\right)^2, \tag{10}$$

where $p_{ij}$ and $p'_{ij}$ representing pixel values of the $i\text{-}th$ row and $j\text{-}th$ column in the cover and stego images, respectively, and $M$ and $N$ denoting the height and width of the image. SSIM defined in Eq. (**錯誤! 找不到參照來源。**) measures the similarity between the cover image and the stego image on a scale from 0 to 1, where higher values indicate greater similarity:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{11}$$

where $x$ and $y$ are the two input images, $\mu_x$ and $\mu_y$ are average luminance, $\sigma_x$and $\sigma_y$ are the standard deviations of luminance, $\sigma_{xy}$ is the covariance between $x$ and $y$, and both $c_1$ and $c_2$ are used for stability. The distinction lies in the fact that SSIM assesses image similarity at a single scale, whereas M-SSIM provides a more comprehensive evaluation by decomposing the image into multiple scales. The M-SSIM value ranges from 0 to 1, with a higher value indicating greater similarity between the two images, as defined in Eq. (**錯誤! 找不到參照來源。**):

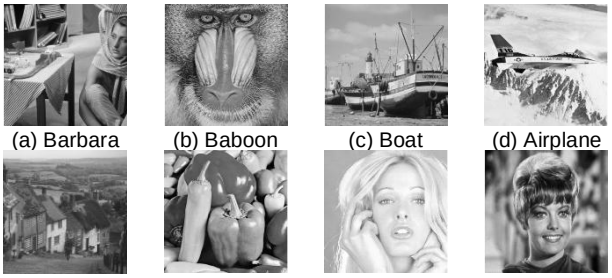
$$\text{M-SSIM}(x, y) = [l_M(x, y)]^{\alpha M}\prod_{j=1}^{M}[C_j(x, y)]^{\beta_j}[S_j(x, y)]^{y_j}, \tag{12}$$

where $x$ and $y$ are the two input images, $M$ represents multiple scales, $l$ represents the calculation of luminance similarity, $C$ represents the calculation of contrast similarity, $S$ represents the calculation of structural similarity, and $\alpha M$, $\beta_j$ and $y_j$are the weights.

Efficiency is the utilization rate of the stego image during embedding data; higher embedding rates indicate better efficiency, defined as follows:

$$Efficiency = \frac{payload}{C_{total}}, \tag{13}$$

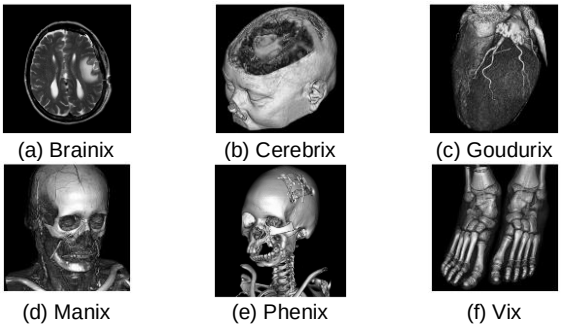
where $C_{total}$ is the bitstream length of the stego image after data embedding and payload the size of the hidden secret bits.



(a) Barbara (b) Baboon (c) Boat (d) Airplane

(e) Goldhill (f) Peppers (g) Tiffany (h) Zelda

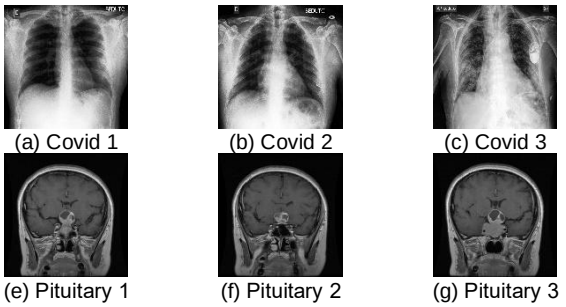
**FIGURE 13. Eight grayscale general images with a size of 512×512 pixels.**



(a) Brainix (b) Cerebrix (c) Goudurix

(d) Manix (e) Phenix (f) Vix

**FIGURE 14. Six medical images with a size of 512×512 pixels.**



(a) Covid 1 (b) Covid 2 (c) Covid 3

(e) Pituitary 1 (f) Pituitary 2 (g) Pituitary 3

**FIGURE 15. Two types of medical images with a size 512×512 pixels.**

### *A. GENERAL EMBEDDING PERFORMANCE ANALYSI1S*

Based on the proposed HEP-DHMI scheme, image blocks within an image are categorized into four distinct types: flat, smooth, moderate, and complex. Tables 0 and 0 present the corresponding distribution of block types within these test images and the test results for general grayscale images, respectively. Table 0 validates that our classification strategy effectively categorizes images, where images with predominantly smooth areas contribute more blocks classified as flat and smooth, and conversely.

**TABLE 3. Number of blocks for grayscale images**

| Image | Flat block | Smooth block | Moderate block | Complex block |
|---|---|---|---|---|
| Barbara | 3374 | 4338 | 7223 | 1449 |
| Baboon | 257 | 4928 | 9806 | 1393 |
| Boat | 7217 | 3716 | 4649 | 802 |
| Airplane | 8203 | 4110 | 3236 | 835 |
| Goldhill | 1767 | 8377 | 5995 | 245 |
| Peppers | 5438 | 7745 | 2762 | 439 |
| Tiffany | 7053 | 6468 | 2647 | 216 |
| Zelda | 3297 | 10440 | 2597 | 50 |

In test images predominantly composed of flat and smooth blocks, such as Lena and Pepper, the proposed scheme achieves an embedding capacity of 230,000 bits. Even in test images with predominantly moderate and



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



complex blocks, such as Barbara and Baboon, which are more intricate, the scheme still achieves a notable embedding capacity of 130,000 bits. This highlights the high payload capacity of the proposed HEP-DHMI, effectively concealing data within images while maintaining robust hiding performance. These findings underscore the stability and efficacy of the proposed approach, demonstrating its applicability across diverse image types and its ability to sustain substantial embedding capacity and superior hiding performance across varying levels of image complexity.

**TABLE 4. Performance experimental results on payload, PSNR, efficiency, SSIM and M-SSIM**

| Images | Payload (bits) | PSNR (dB) | Efficiency | SSIM |
|---|---|---|---|---|
| Barbara | 170,114 | 28.71 | 30.97 | 0.88 |
| Baboon | 129,483 | 27.68 | 23.46 | 0.84 |
| Boat | 224,127 | 31.11 | 41.04 | 0.87 |
| Airplane | 242,725 | 32.20 | 44.56 | 0.86 |
| Goldhill | 191,215 | 31.43 | 34.96 | 0.83 |
| Peppers | 239,184 | 32.40 | 43.98 | 0.86 |
| Tiffany | 251,531 | 34.21 | 46.28 | 0.85 |
| Zelda | 235,575 | 33.64 | 43.36 | 0.82 |

Regarding image quality, Table 0 also shows that the PSNR remains stable at approximately 30 dB, ensuring that the hidden image maintains good quality with changes imperceptible to the human eye. The SSIM score for the hidden image post-embedding stays around 0.85, indicating high similarity in luminance, contrast, and structure compared to the cover image, without significant visual differences. This demonstrates that embedding data does not noticeably alter the overall image structure. Lastly, the M-SSIM score for the hidden image post-embedding reaches approximately 0.98, indicating very high structural similarity across multiple scales between the hidden image and the cover image. This ensures the overall quality and structural integrity of the image are maintained effectively. The high structural similarity indicated by M-SSIM is due to our proposed HEP-DHME scheme, which classifies relatively complex blocks into two categories: moderate blocks and complex blocks. For these categories, the original BM, which could impact structural similarity, has been carefully preserved during the data embedding process.

Tables 0 and 0 further illustrate that across various types of medical images, the majority of blocks are categorized as flat and smooth, maintaining an embedding capacity of approximately 234,000 bits. This highlights the scheme's capability to sustain high embedding capacity across different medical image types. Regarding image quality, the results indicate that PSNR remains stable at around 32.0 dB, indicating that medical images maintain good quality after data embedding with high visual fidelity. Additionally, both SSIM and M-SSIM scores are maintained around 0.92 and 0.99, respectively. This confirms very high structural similarity between the stego and original medical images, ensuring the preservation of structural integrity. In essence, this guarantees the readability of the stego medical images.

**TABLE 5. Number of blocks for different medical images**

| Images | Flat block | Smooth block | Moderate block | Complex block |
|---|---|---|---|---|
| Brainix | 10439 | 2874 | 2490 | 581 |
| Cerebrix | 5870 | 5386 | 4783 | 345 |
| Goudurix | 6643 | 4325 | 4876 | 540 |
| Manix | 6476 | 4960 | 4610 | 338 |
| Phenix | 9295 | 2368 | 3579 | 1142 |
| Vix | 7509 | 3584 | 5061 | 230 |

**TABLE 6. Performance experimental results on payload, PSNR, efficiency, SSIM, and M-SSIM**

| Images | Payload (bits) | PSNR (dB) | Efficiency | SSIM |
|---|---|---|---|---|
| Brainix | 264,921 | 32.50 | 48.72 | 0.95 |
| Cerebrix | 222,450 | 31.91 | 40.76 | 0.90 |
| Goudurix | 222,655 | 31.17 | 40.77 | 0.91 |
| Manix | 227,308 | 32.16 | 41.66 | 0.91 |
| Phenix | 240,457 | 30.01 | 44.09 | 0.94 |
| Vix | 229,475 | 32.44 | 42.03 | 0.93 |

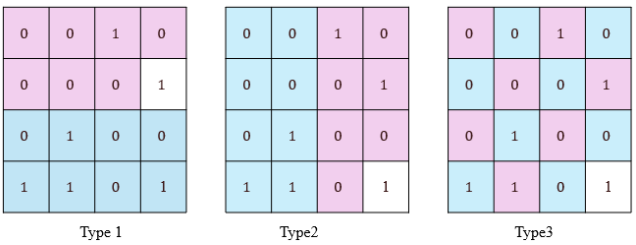



**FIGURE 16. Three types of *BM* pattern.**

**TABLE 7. Performance experimental results on PSNR and SSIM, with three patterns of *BM***

| Image | Type1 | | Type2 | | Type3 | |
|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Brainix | 32.50 | 0.95 | 32.51 | 0.95 | 32.51 | 0.95 |
| Cerebrix | 31.91 | 0.90 | 31.91 | 0.90 | 31.91 | 0.90 |
| Goudurix | 31.17 | 0.91 | 31.17 | 0.91 | 31.17 | 0.91 |
| Manix | 32.16 | 0.91 | 32.16 | 0.91 | 32.17 | 0.91 |
| Phenix | 30.01 | 0.94 | 30.01 | 0.94 | 30.01 | 0.94 |
| Vix | 32.44 | 0.93 | 32.45 | 0.93 | 32.46 | 0.93 |

To evaluate the impact of different BM patterns on the performance of our HEP-DHMI regarding image quality and SSIM, we employed three BM patterns, as shown in Fig. 16, to assess the six medical images displayed in Fig. 15. The results are summarized in Table 7. The analysis revealed that all BM patterns produced the same SSIM value, indicating no significant differences in image quality. However, we will proceed with the Type 3 BM pattern for future experiments based on these findings.



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



### *B.* PERFORMANCE OF DI*FFERENT MEDICAL IMAGE DATABASES*

To evaluate the overall performance and practicality of HEP-DHMI, we conducted tests using a selection of 100 Covid lung images and 100 pituitary gland images sourced from the Covid-19 image dataset [32] and Brain Tumor MRI dataset [33], totaling 200 images as depicted in Fig. FIGURE 15. . Figs. FIGURE 17. and FIGURE 18. illustrate that the embedding capacity consistently ranges between 200,000 bits and 250,000 bits while maintaining a PSNR range of 30 dB to 35 dB. These results demonstrate the robustness of our proposed HEP-DHMI and its ability to consistently deliver high performance across diverse image types.

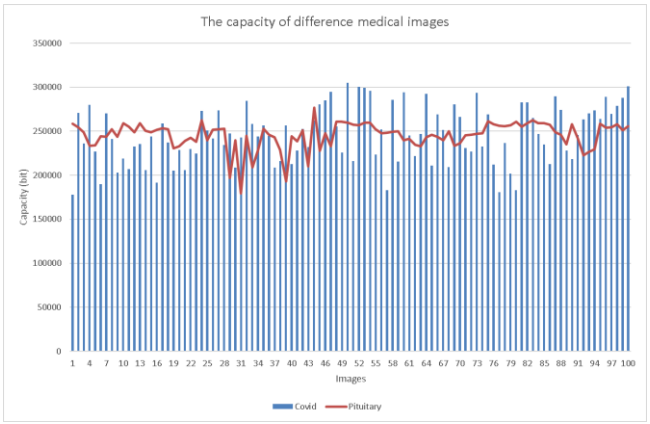



**FIGURE 17. Capacity performance of different medical images with 200 images.**

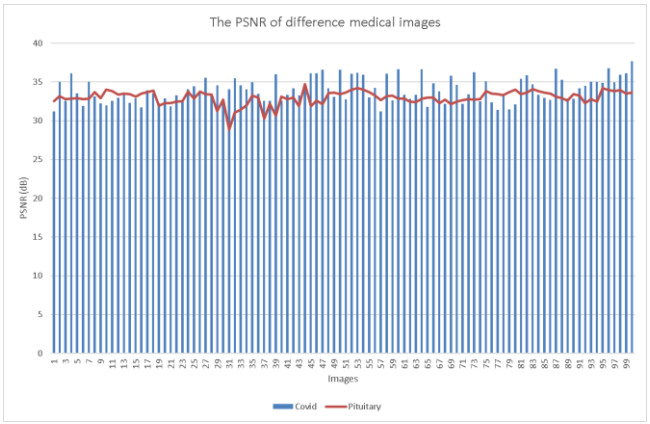



**FIGURE 18. PSNR performance of different medical images with 200 images.**

The stability of our proposed HEP-DHMI scheme in maintaining image quality and hiding capacity across two sets of medical images—Covid-19 [32] and Brain Tumor [33]—can be attributed to the high capacity provided by flat and smooth blocks. In these datasets, the primary background is relatively dark, smooth, and consistent. This characteristic enables our HEP-DHMI to extract more smooth and flat blocks. Additionally, for the textured portions of the medical images, our HEP-DHMI differentiates between two categories: moderate and complex blocks. This categorization ensures high structural similarity and contributes to superior image quality. The average performance of two different image databases in terms of payload, PSNR, efficiency, SSIM, and M-SSIM using HEP-DHMI is detailed in Table 8. Despite the average SSIM being below 0.85 in the Covid lung database, its M-SSIM reaches approximately 0.98. This indicates that when considering multiple scales, the structural similarity between the original and stego images provided by HEP-DHMI remains stable, hovering around 0.98.

**TABLE 8. Performance experimental results on payload, PSNR, efficiency, SSIM, and M-SSIM**

| Image | Payload (bits) | PSNR (dB) | Efficiency | SSIM | M-SSIM |
|---|---|---|---|---|---|
| Covid Lung | 245988.20 | 33.98 | 45.31 | 0.85 | 0.98 |
| Pituitary Brain | 245589.01 | 34.52 | 45.13 | 0.90 | 0.98 |

### *C. COMPARISON OF OUR HEP-DHMI WITH FOUR EXISTING SCHEMES*

To demonstrate the superior performance of the proposed HEP-DHMI compared to the previous schemes, four existing schemes—Ou et al. [24], Kumar et al. [25], Chang et al. [26], and Chen et al. [27]—are included in the comparisons presented in Table 9. Table 9 illustrates that when HEP-DHMI is applied to the Airplane image, its PSNR is comparable to that of other schemes, showing an improvement of approximately 1 dB, along with an additional effective payload of about 30,000 bits. Testing on medical images also shows HEP-DHMI achieving a PSNR about 1 dB higher than competing schemes, with a payload advantage of 120,000 bits. While the PSNR improvement by HEP-DHMI may not always be significant in every scenario, it consistently demonstrates significant performance advantages across various types of images in both hiding capacity and efficiency.

In addition to the detailed comparison of our HEP-DHMI method with four representative approaches in Table 9, we also analyze the performance of HEP-DHMI presented in Figs. 17 and 18 alongside two methods published in 2023: Lai et al.'s method [28] and Al-Chaab et al.'s method [30]. The former focuses primarily on general images, while the latter emphasizes protecting medical images while effectively reducing transmission file size. It was found that when comparing our HEP-DHMI with Lai et al.'s average performance—showing a hiding capacity of 224,807 bits and an image quality of 30.395 dB at a threshold of 8 [28]—our method demonstrates relatively higher image quality. Our HEP-DHMI achieves an average PSNR of 32.5 dB at a compression rate of 25%, surpassing the PSNR of 31 dB obtained by Al-Chaab et al. at a compression rate of 28%



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



[30]. Furthermore, while Al-Chaab et al.'s method produces encrypted medical images that appear noise-like, the stego media generated by our HEP-DHMI resembles AMBTC compression codes. This characteristic makes the stego media produced by our method less likely to attract the attention of malicious attackers, as discussed in [28].

## V. CONCLUSION

This paper introduces HEP-DMHI, a block-based data hiding scheme tailored for AMBTC-based medical and general images. Given the critical need for structural similarity in medical images and the potential requirement to embed additional confidential data such as Electronic Health Records, HEP-DMHI categorizes AMBTC compressed blocks into four types—flat, smooth, moderate, and complex—based on differences in block quantization. This classification informs the development of distinct data embedding and extraction strategies.

Experimental results illustrate that block distributions correspond closely to the texture of the images. Specifically, smooth images predominantly contain flat and smooth blocks, whereas complex images are characterized by moderate and complex blocks. According to our block classification rules, experiments confirm that HEP-DHMI maintains an average image quality of 34.25 dB, a hiding capacity of 245,788 bits, an efficiency of 45.22, and an M-SSIM of 0.98 when evaluated with Covid-19 and Brain Tumor MRI datasets.

Moreover, HEP-DHMI demonstrates superior performance over four established AMBTC-based data hiding schemes in terms of image quality, hiding capacity, and efficiency for both stego general and medical images. Specifically, the average hiding capacity and PSNR offered by HEP-DHMI are respectively 1.9 times and 2.99 dB higher than those of Chang et al. [25], which maintains the highest payload among the four representative schemes. With increased hiding capacity, not only can Patient Health Records be transmitted synchronously with medical images, but authentication codes can also be embedded to verify the integrity of the medical images later. Moving forward, our research will focus on exploring methods to embed both confidential data and authentication codes simultaneously, aiming to further enhance the security and integrity of transmitted medical images.

**TABLE 9. Performance comparison on payload, PSNR, and efficiency**

| Method | Metric | Test image | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Airplane | Brainix | Cerebrix | Goudurix | Manix | Phenix | Vix |
| Ou et al. [24] | PSNR | 31.71 | 30.95 | 30.83 | 29.75 | 30.7 | 28.01 | 31.17 |
| | Payload | 174,754 | 117,580 | 74,950 | 75,400 | 82,000 | 101,635 | 88,390 |
| | Efficiency | 33.33 | 36.74 | 23.42 | 23.56 | 25.63 | 31.79 | 27.62 |
| Kumar et al. [25] | PSNR | 31.33 | 30.68 | 29.98 | 29.02 | 30.1 | 27.82 | 30.48 |
| | Payload | 200,912 | 130,250 | 101,608 | 100,808 | 106,328 | 117,118 | 109,234 |
| | Efficiency | 38.32 | 40.70 | 31.85 | 31.50 | 33.23 | 36.60 | 34.14 |
| Chang et al. [26] | PSNR | 29.94 | 29.59 | 28.73 | 27.84 | 29.02 | 26.92 | 29.40 |
| | Payload | 210,142 | 134,698 | 110,668 | 110,122 | 114,513 | 123,444 | 116,316 |
| | Efficiency | 40.08 | 42.0 | 34.58 | 34.41 | 35.79 | 38.58 | 36.35 |
| Yung et al. [27] | PSNR | 31.85 | 31.05 | 30.96 | 29.73 | 31.06 | 28.04 | 31.42 |
| | Payload | 209,086 | 136,810 | 107,323 | 99,463 | 109,795 | 118,348 | 115,369 |
| | Efficiency | 38.67 | 41.46 | 32.52 | 30.14 | 33.27 | 35.86 | 34.93 |
| Ours | PSNR | 32.20 | 32.5 | 31.91 | 31.17 | 32.16 | 30.01 | 32.44 |
| | Payload | 242,725 | 264,921 | 222,450 | 222,655 | 227,308 | 240,457 | 229,475 |
| | Efficiency | 44.56 | 48.72 | 40.76 | 40.77 | 41.66 | 44.09 | 42.03 |

## REFERENCES

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

[2] W.-J. Wang, C.-T. Huang, and S.-J. Wang, "VQ applications in steganographic data hiding upon multimedia images," *IEEE Systems Journal*, vol. 5, no. 4, pp.528–537, December 2011.

[3] A. Shukla, A. Singh, B. Singh, and A. Kumar, "A secure and high-capacity data-hiding method using compression, encryption and optimized pixel value differencing," *IEEE Access*, vol. 99, no. 6, pp. 51130–51139, September 2018.

[4] S. Li, and X. Zhang, "Toward construction-based data hiding: from secrets to fingerprint images," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1482–1497, October 2018.

[5] D. Xiao, F. Li, M. Wang, and H. Zheng, "A novel high-capacity data hiding in encrypted images based on compressive sensing progressive recovery," *IEEE Signal Process Lett*., vol. 27, pp. 296–300, January 2020.

[6] S. Chen, C.-C. Chang, and Q. Liao, "Fidelity preserved



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



data hiding in encrypted images based on homomorphism and matrix embedding," *IEEE Access*, vol. 8, pp. 2169–3536, January 2020.

[7] Z. Yu, C.-C. Lin, and C.-C. Chang, "ABMC-DH: an adaptive bit-plane data hiding method based on matrix coding," *IEEE Access*, vol. 8, pp. 27634–27648, January 2020.

[8] W. Chen, C.-C. Chang, S. Weng, and B. Ou, "Multi-layer mini-Sudoku based high-capacity data hiding method," *IEEE Access*, vol. 8, pp. 69256–69267, April 2020.

[9] J.-H. Horng, S. Xu, C.-C. Chang, and C.-C. Chang, "An efficient fata-hiding scheme based on multidimensional mini-SuDoKu," *Sensors*, vol. 20, no. 9, pp. 2739, May 2020.

[10] J. Tian, "Reversible data embedding using a difference expansion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 890–896, August 2003.

[11] 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, March 2006.

[12] W. Hong, T.-S. Chen, and C.-W. Shiu, "Reversible data hiding for high quality images using modification of prediction errors," *Journal of Systems and Software*, vol. 82, no. 11, pp. 1833–1842, November 2009.

[13] J. He, J. Chen, W. Luo, S. Tang, and J. Huang, "A novel high-capacity reversible data hiding scheme for encrypted JPEG bitstreams," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3501–3515, November 2018.

[14] Y.-C. Chen, T.-H. Hung, S.-H Hsieh, and C.-W. Shiu, "A new reversible data hiding in encrypted image based on multi-secret sharing and lightweight cryptographic algorithms," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp.3332–3343, May 2019.

[15] J. He, J. Chen, and S. Tang, "Reversible data hiding in JPEG images based on negative influence models," *IEEE Transactions on Information Forensics and Security*., vol. 15, pp. 2121–2133, December 2019.

[16] Y. Qiu, Q. Ying, X. Lin, Y. Zhang, and Z. Qian, "Reversible data hiding in encrypted images with dual data embedding," *IEEE Access*, vol. 8, pp. 23209–23220, January 2020.

[17] Y. Yang, X. Xiao, X. Cai, and W. Zhang, "A secure and privacy-preserving technique based on contrast-enhancement reversible data hiding and plaintext encryption for medical images," *IEEE Signal Processing Letters*, vol. 27, pp.256–260, January 2020.

[18] C.-C. Lin, T.-S. Nguyen, and C.-C. Chang, "Efficient reversible data hiding scheme for AMBTC-compressed images," *Applied Sciences.*, vol. 11, no. 15, pp. 6741, July 2021.

[19] J.-Y. Lin, J.-H. Horng, and C.-C. Chang, "A reversible and authenticable secret sharing scheme using dual images," *Multimedia Tools and Applications*, vol. 81, pp. 17527–17545, March 2022.

[20] M. Fan, S. Zhong, and X. Xiong, "Reversible data hiding method for interpolated images based on modulo operation and prediction-error expansion," *IEEE Access*, vol. 11, pp. 27290–27302, March 2023.

[21] L. Qu, M. Li, and P. Chen, "Reversible data hiding in encrypted image with secure multi-party for telemedicine applications," *Biomedical Signal Processing and Control*, vol. 93, July 2024.

[22] X. Liang, and S. Xiang, "General distortion metric based histogram shifting for reversible data hiding," *IEEE Signal Processing Letters*, vol. 31, pp. 2420–2424, September 2024.

[23] M. Lema, and O. Mitchell, "Absolute moment block truncation coding and its application to color images," *IEEE Transactions on Communications*, vol. 32, no. 10, pp. 1148–1157, October 1984.

[24] D. Ou, and W. Sun, "High payload image steganography with minimum distortion based on absolute moment block truncation coding," *Multimedia Tools and Applications*, vol. 74, pp. 9117–9139, May 2014.

[25] R. Kumar, D. Kima, and K. Jung, "Enhanced AMBTC based data hiding method using hamming distance and pixel value differencing," *Information Security and Applications*, vol. 47, pp. 94–103, August 2019.

[26] C.-C. Chang, X. Chang, and J.-H. Horng, "Hybrid data hiding method for strict AMBTC format images with high-fidelity," *Symmetry*, vol. 11, no. 10, pp.1314, October 2019.

[27] Y.-Y. Chen, Y.-C. Hu, H.-Y. Kao, and Y.-H. Lin, "Security for eHealth system: data hiding in AMBTC compressed images via gradient-based coding," *Complex & Intelligent Systems*, vol. 9, pp. 2699–2711, May 2021.

[28] P.-C. Lai, J.-J. Shen, and Y.-C. Chou "High embedding capacity data hiding technique based on hybrid AMBTC and LSB substitutions," *International Journal of Network Security*, vol. 25, no. 2, pp. 221–234, March 2023.

[29] Y. Lin, C.-C. Lin, C.-C. Chang, and C.-C. Chang, "An IoT-based electronic health protection mechanism with AMBTC compressed images," *IEEE Internet of Things Journal*, vol. 99, 2024. Available: DOI: 10.1109/jiot.2024.3467152

[30] W. Al-Chaab, Z.-A. Abduljabbar, E.-W. Abood, V.-O. Nyangaresi, H.-M. Mohammed, and J. Ma, "Secure and low-complexity medical image exchange based on compressive sensing and LSB audio steganography," *Informatica*, vol. 47, no. 6, pp. 65–74, 2023.

[31] Osirix Database [Online]. Available: https://www.osirix-viewer.com/support/ knowledge-



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/

This article has been accepted for publication in IEEE Access. 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/ACCESS.2025.3568601



base/, Accessed on: October. 26, 2024.
[32] Covid-19 Image Dataset [Online]. Available: https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset/data, Accessed on: October. 26, 2024.
[33] Brain Tumor MRI Dataset [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset, Accessed on: October. 26, 2024.



**Chia-Chen Lin** (also known as MIN-HUI LIN) received her Ph.D degree in information management in 1998 from the National Chiao Tung University. Dr. Lin is currently a professor of the Department of Computer Science and Information Engineering, National Chin-Yi University of Technology. Since 2018, she is the Fellow of IET and the School counselor for Providence University. In additions, from 2009 to 2012, she served as the Vice Chairman of Tainan Chapter IEEE Signal Processing Society. She also serves Associate Editor and Editor for several representative EI, SCIE journals. Her research interests include image and signal processing, information hiding, mobile agent, and electronic commerce.



**En-Ting Chu** received his M.S. degree from National Chin-Yi University of Technology in 2023, where he is currently pursuing the Ph.D. degree in prospective technology of electrical engineering and computer science. His research interests include image processing, data hiding, and deep learning.



**Iuon-Chang Lin** received the Ph.D. in Computer Science and Information Engineering in March 2004 from National Chung Cheng University, Chiayi, Taiwan. He is currently a professor of the Department of Management Information Systems, National Chung Hsing University, Taichung, Taiwan. His current research interests include electronic commerce, information security, blockchain security, and cloud computing.



**Rajeev Kumar** received Ph.D. in Computer Engineering from University of Delhi, India, in 2017. He completed 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 was awarded a fellowship from National Research Foundation (NRF), Korea, for the period 2018-2019 and 2021-2024 to work on projects in the domain of multimedia security. He has been awarded with Commendable research excellence awards from Delhi Technological University for his research in 2021 and 2022. Currently, he is serving as an Associate Editor with Journal of Information Security and Applications (Elsevier), Journal of Electronic Imaging (JEI), SPIE, SN Computer Science (Springer) and review editor with journal 'Frontiers in Signal Processing'. He is also serving as lead guest editor with "Journal of Information Security and Applications", Elsevier and guest editor with Journal of Internet Technology (JIT). His research interests include steganography, reversible data hiding, multimedia forensics, Image Processing, compression, and wireless sensor networks.



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/

Check for updates

*International Journal of Circuit Theory and Applications*

WILEY

ORIGINAL ARTICLE

# High-Gain Fault-Tolerant Multiport Bidirectional DC–DC Converter With Time Multiplexing Control

Abhishek Kumar Gupta | Mayank Kumar | Madhusudan Singh

Centre of Excellence for Electric Vehicle and Related Technologies, Department of Electrical Engineering, Delhi Technological University, Delhi, India

**Correspondence:** Mayank Kumar (mayankkumar@dtu.ac.in)

**Received:** 19 December 2024 | **Revised:** 8 March 2025 | **Accepted:** 28 April 2025

**Keywords:** fault tolerant | high-gain dc-dc converter | multiport converter | open-circuit switch faults (OCSFs) | time multiplexing control

## ABSTRACT

In low-power multiple-voltage applications, single-input, multiple-output nonisolated converters are becoming popular as alternative dc-dc topologies to increase efficiency and reduce the size and cost of the converter. This article presents a novel high-gain fault-tolerant multiport (HGFTM) dc-dc converter for low- and medium-voltage applications. The proposed converter is a combination of conventional boost and buck converter topologies with time multiplexing control. The switched boost action divides power into two ports using time multiplexing control. The noncascading structure of the converter provides high gain with improved efficiency. The proposed converter provides open circuit switch fault (OCSF) tolerant capability without any additional device requirement. This provides controlled output voltages for a wide range of applications. The main objective of this research is to attain robust and fault-tolerant operations in multiport converter systems, ensuring continued power delivery even under component failure. A closed-loop control technique is implemented to control the output voltages and assure the system's stability and accuracy under different load conditions. The physical implementation of the converter is tested in various scenarios to ensure its real-time application. The experimental findings closely match the simulation results, showing the effectiveness of the proposed fault-tolerant multiport dc-dc converter.

## 1 | Introduction

Reliability of power electronic converters is important concern for the critical applications such as aeronautical devices, medical equipment, electric vehicles, and distributed generating units [1, 2]. The demand for dc-dc conversion, including both step-up and step-down operations, has been increasing due to its utilization in photovoltaic generating systems, battery systems, and dc power loads [3]. Therefore, this article focuses on implementing a fault-tolerant design to enhance the reliability of dc-dc converters.

The fault-tolerant method consists of two main processes: fault detection and corrective action. Fault detection is mainly responsible for identifying and isolating the semiconductor switch faults, whereas corrective action involves reconfiguring the converter to ensure uninterrupted service. The ideal fault-tolerant capability of the system can be categorized as having zero additional power losses and being cost-effective [4]. Previous research has shown that power semiconductor failure is the primary cause of converter issues. A sudden malfunction in the power switch can lead to a catastrophic and irreversible system failure [5, 6]. In power semiconductor devices, failures can be categorized into short-circuit switch faults (SCSFs) and open-circuit switch faults (OCSFs) [7]. The SCSFs can occur due to overvoltage, static or dynamic latch-up, and so on. whereas the OCSFs can occur due to the factors such as wire bonding disruptions, gate driver circuitry, and energy shocks. Since most traditional power converters lack redundancy in their design, any fault in the power devices can lead to system disruptions. Consequently, there is a growing research interest in power converters that possess fault-tolerant capabilities and offer

© 2025 John Wiley & Sons Ltd.

enhanced system reliability, particularly in applications that prioritize uninterrupted power supply.

The concept of a high-gain fault-tolerant multiport bidirectional dc-dc converter with time multiplexing control is an advanced approach in power electronics, particularly for applications in dc microgrids and renewable energy systems. This type of converter is designed to handle multiple input and output ports, providing high conversion efficiency and fault tolerance with reduced number of semiconductor devices, which are critical for maintaining system reliability and performance [8]. The integration of time multiplexing control further enhances the converter's ability to manage power flow dynamically across different ports, optimizing the use of available energy resources. The proposed multiport high-gain bidirectional dc-dc converter offers significant advantages such as high step-up/step-down conversion gain, reduced switch voltage stress, and a lower component count, making it suitable for dc applications [9].

In [10], a novel multiphase and multilevel converter architecture with enhanced fault tolerance under OCSF and SCSF circumstances is proposed. All of the above-discussed methods, taken together, offer efficient fault detection and subsequent converter protection. But once the problem arises, their sluggish reaction time renders them useless in stopping the system from collapsing. To determine the OCSF and SCSF in phase-shifted full-bridge converters, a switching mode-based diagnosis technique based on frequency data from electromagnetic measurement is employed in [11]. With the use of current error signals [12] and modular error voltages [13], the OCSF diagnosis in matrix converters has been proven. To maintain system continuity, a failing full-bridge converter is converted into a half-bridge topology using the fault-tolerant technique [14]. In [15], the suggested fault-tolerant multilevel H-bridge dc-dc converter can maintain output voltage even after OCSF and SCSF.

Multiple-input and multiple-output (MIMO) ports, single-input multiple-output (SIMO), or multiple-input single-output (MISO) converters are features that can be found in multiport dc-dc converters (MPDCs) [16–18]. In order to create MPDCs, multiple topologies are combined, either by combining converters with comparable input side architectures [17] or by using semiconductor switches to facilitate a range of activities. The MPDC topology offers decreased dimensions, weight, and expense. Time multiplexing control action is used to transfer the power at two ports with regulated output voltage [16].

The integration of advanced control strategies and innovative topologies will play a pivotal role in enhancing the performance and adaptability of these converters, ultimately supporting the transition towards greener energy solutions [19]. In [1], a three-level boost converter with OCSF diagnostics and fault tolerance behavior of the converter has been presented, where the maximum power point tracking (MPPT) utilizes output capacitor voltage and control parameters as variables. In [20], a dc-dc converter was presented that combines many resonant switched-capacitor cells to form a voltage multiplier circuit. This structure enables the fault-tolerant functionality and the variable output voltage ratio. If an active cell fails, the converter will keep the reference voltage stable by skipping the damaged switch. Several interleaved converters with fault tolerance characteristics were introduced in [21, 22]. An adjustable multiphase PWM control based on an OCSF detection technique is used to build a fault-tolerant topology. Thus, the large number of circuit elements contributes to the guarantee of uninterrupted service even in the event of a switch-leg failure [23–25]. This comprehensive exploration highlights the evolving landscape of DC-DC converters, particularly focusing on their fault tolerance capabilities and multiport functionalities that cater to various applications in renewable energy systems and electric vehicles. As the demand for efficient energy conversion rises, addressing these challenges becomes crucial in developing next-generation power management solutions that ensure reliability and sustainability.

The ability to seamlessly manage power distribution across multiple sources and outputs not only enhances reliability but also optimizes overall system efficiency, ensuring that critical loads receive uninterrupted power even in the event of component failures. This capability is particularly important in applications such as renewable energy systems, electric vehicles, and data centers, where maintaining a consistent power supply is essential for performance and safety.

This article presents a novel high-gain fault-tolerant multiport (HGFTM) dc-dc converter. The converter provides the step-down (at output Port-2) and a high-gain step-up (output Port-2), bidirectional and output port, respectively, using time multiplexing control. The major contributions of this research work are as follows: *(i)* the proposed converter ensuring high-voltage gain at output Port-1, *(ii)* fault-tolerant capability of the converter for Port-1 output and partial fault-tolerant capability at Port-2, *(iii)* multiport feature using time multiplexing control for power transfer at both the ports, *(iv)* it supports the battery charging and discharging characteristics, that is, bidirectional behavior, and *(v)* an experimental prototype is built and validated the features of the proposed converter in real-time.

This article is arranged as follows: Section 1 explains the background of the various multiport converter topologies and fault-tolerant methodologies for dc-dc converters; the brief introduction and operation principle of the proposed converter are presented in Section 2. Section 3 presents the state space averaged (SSA) model and steady-state behavior of the proposed converter; the behavior of the proposed converter in different operating regions under OCSF conditions is presented in Section 4. An experimental setup and the results are demonstrated to verify the behavior of the proposed converter in Section 5, and the conclusions are presented in Section 6.

## 2 | Proposed HGFTM DC-DC Converter and Operation Principle

The proposed HGFTM dc-dc converter is developed for managing power flow between a source, typically a constant dc or photovoltaic system, and multiple output ports. The block diagram of the proposed converter is shown in Figure 1. The converter design aims to ensure high gain to the output Port-1 and efficient power distribution while offering fault tolerance and



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

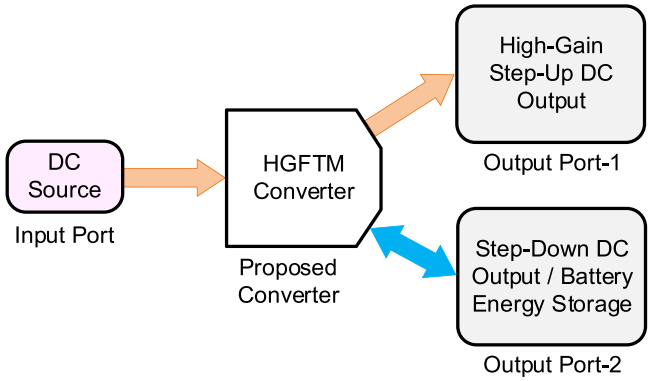



**FIGURE 1** | Block diagram of the proposed converter.

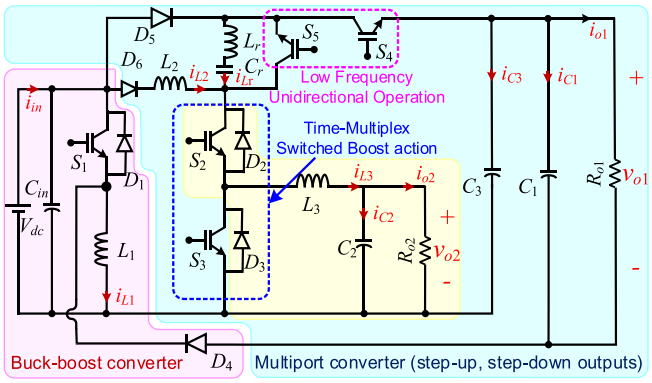



**FIGURE 2** | Schematic diagram of the proposed HGFTM dc-dc converter.

bidirectional energy flow capabilities. The HGFTM converter consists of two distinct output ports: a high-gain step-up port and a step-down port. These ports serve different purposes and allow for versatile power management.

As shown in Figure 2, the proposed HGFTM dc-dc converter combines the characteristics of a switched boost converter with time multiplexing control and the traditional buck-boost topology with a storage element. An important aspect during power flow from input to output is as follows: A portion of the power is supplied directly to the output Port-1, whereas the other portion is supplied through the storage element ($C_3$). An important behavior of this converter is its noncascading structure with time multiplexing control, which provides the capability of multiport output with improved efficiency and high step-up gains to output Port-1. The important features of the output Port-1 and output Port-2 of the proposed converter are as follows: *High Gain Fault-Tolerant Step-up Output Port*-1: The nonisolated output Port-1 is designed to amplify the input voltage to a higher level, making it suitable for applications requiring increased voltage levels compared to the input source. The high-gain nature of this port enables efficient voltage rise, making it ideal for high-voltage applications. The switch $S_4$ is used to properly shut down the output Port-1; this switch can also be replaced with a diode if a unity gain output is required during the off state of the converter.

*Bidirectional Step-Down Output Port*-2: The second output port facilitates voltage reduction, making it suitable for applications where the voltage needs to be lowered before being supplied to connected devices or systems. Additionally, this port has bidirectional features that enable the battery charging/discharging characteristics. The integration of a battery into the second output port adds another layer of functionality to the HGFTM converter. The switch $S_5$ is used to enable the discharging of the battery to the output Port-1. The bidirectional capability of the output Port-2 enhances system reliability and resilience by enabling seamless power flow management between the source, battery, and connected load. It ensures uninterrupted power supply to critical components with the use of an energy storage system connected at output Port-2, thereby optimizing overall system efficiency.

The circuit diagram of the proposed HGFTM dc-dc converter is shown in Figure 2. Where $V_{dc}$ is the dc input voltage, $v_{o1}$ is the step-up voltage across the output Port-1, and the $v_{o2}$ is the step-down voltage across the output Port-2. The loads connected at Port-1 and Port-2 are $R_{o1}$ and $R_{o2}$, respectively. The proposed converter consists of three energy transfer inductors ($L_1$, $L_2$, and $L_3$), a resonant inductor $L_r$, three capacitors ($C_1$, $C_2$, and $C_3$), a resonant capacitor $C_r$, three bidirectional high frequency switches ($S_1$ $S_2$, and $S_3$), two unidirectional low-frequency switches ($S_4$, $S_5$), and two high-frequency diodes ($D_4$ and $D_5$). The filter capacitors $C_1$ and $C_2$ are used at output Port-1 and Port-2, respectively. Based on the direction of the flow of power, there are two different states in healthy conditions. In *State*-I, a dc source is connected at the input port which can supply the power to both output ports as shown in Figure 3a, and in *State*-II as shown in Figure 3b, the input supply is absent; in this case, the battery energy storage is connected at the output Port-2 started supply to the output Port-1. Hence, the converter ensures the uninterrupted supply to the load.

The principle of operation of the proposed converter is presented for healthy conditions and OCSF conditions. Further, the operation of the converter is classified as *State*-I, that is, during power flow from the input source to output Port-1 and Port-2 or charging at output Port-2 (with BESS), and *State*-II, that is, discharging of BESS to the output Port-1. Each state is further classified into modes of operation, which are discussed in the following subsections.

## 2.1 | Operating Modes for *State*-I Under Healthy Conditions

The switch-on time period of the switch $S_3$ (i.e., $G_3$) is considered as $DT_s$, where $D$ is the duty ratio and $T_s$ is the switching time period. The switch-on time period of the switches $S_2$ (i.e., $G_2$) and $S_1$ (i.e., $G_1$) is defined as $\lambda DT_s$ and $\beta DT_s$, respectively, where $\lambda$ and $\beta$ are the duty factors of $S_2$ and $S_1$, respectively. To enable the output Port-2, the value of $\lambda$ should be greater than 1, that is, $\lambda > 1$ where the value of $\lambda D \leq 1$. The value of $\beta$ depends on the contribution of power cascaded from the buck-boost converter to the output Port-1. Figure 4a–c shows the switching logic of the proposed converter under *State*-I at $\beta > \lambda$, $1 < \beta \leq \lambda$, and $\beta \leq 1$ with respect to the contribution of power cascaded from the buck-boost converter to the output Port-1, where the value of $\beta D \leq 1$.



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

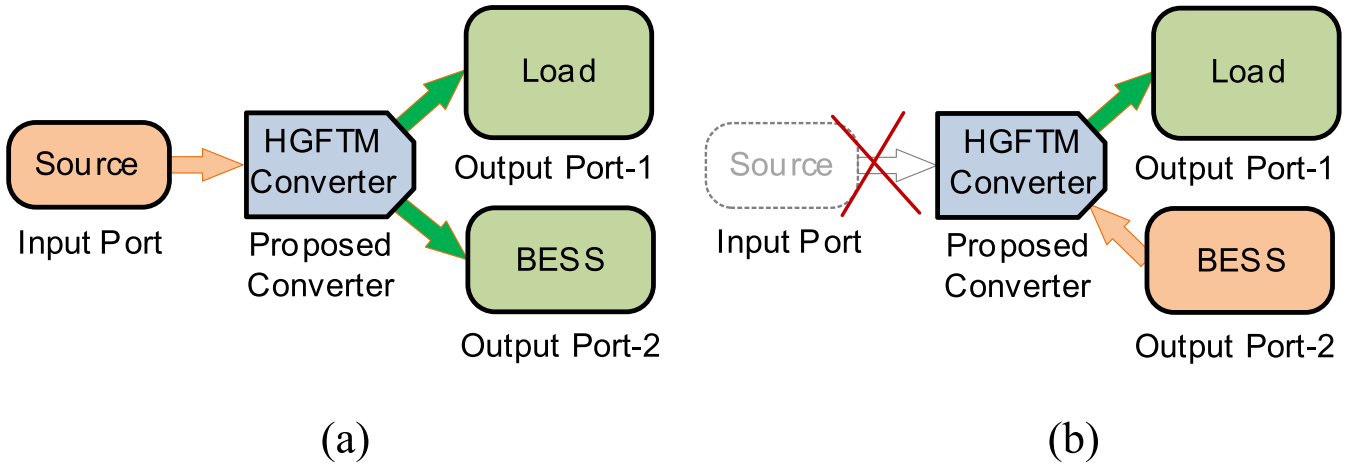



**FIGURE 3** | Different states of operation of proposed converter. (a) *State*-I: charging mode. (b) *State*-II: discharging mode.

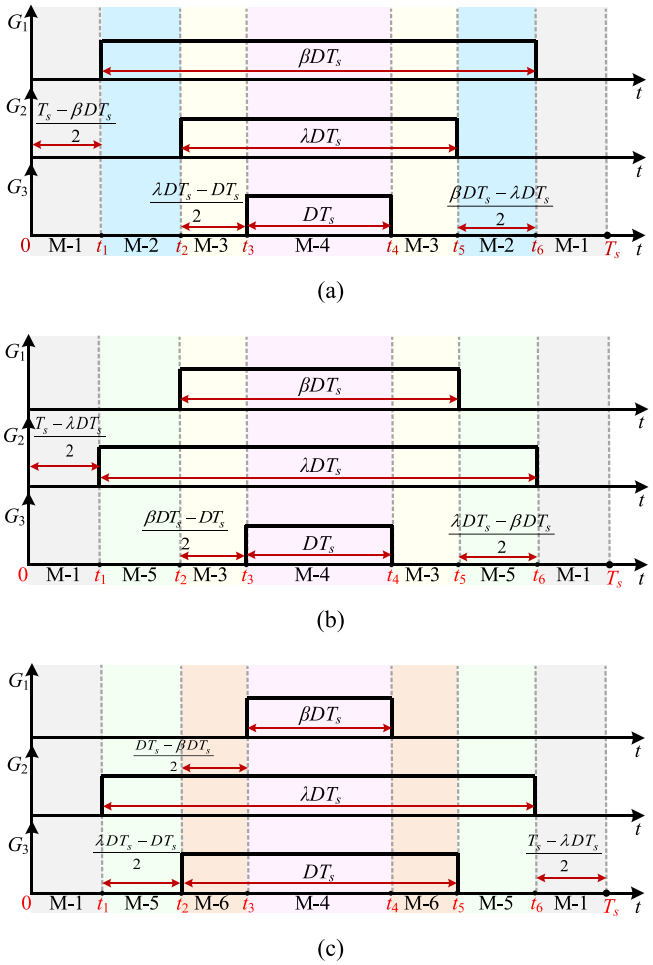



**FIGURE 4** | Switching logic of proposed HGFTM converter for *State*-I under healthy conditions for $1 \le \lambda \le 1/D$ at (a) $\lambda \le \beta \le 1/D$, (b) $1 \le \beta \le \lambda$, and (c) $\beta \le 1$.

Six different modes of operations are possible during *State*-I for all the possible values of duty factor of $S_1$ (i.e., $\beta$). The first four modes for $\beta > \lambda$ are shown in Figure 5. The switching signals for $S_1$, $S_2$, and $S_3$ are shown in Figure 4a. All four modes, that is, Mode-1 (M-1), Mode-2 (M-2), Mode-3 (M-3), and Mode-4 (M-4), are represented with respect to their time periods. The remaining modes for $1 < \beta \le \lambda$ and $\beta \le 1$ are presented in Figure 4b,c, respectively. Two unidirectional low-frequency switches $S_4$ and $S_5$ are used. The switch $S_4$ is used to properly shut down the output Port-1; this switch can also be replaced with a diode if a unity voltage gain output is required during the off-state of the converter. The switch $S_5$ is used to enable *State*-II, that is, discharging of the battery to the output Port-1 during the unavailability of the input source.

A brief discussion of all the operating modes for *State*-I at $\beta > \lambda$ of the proposed converter is as follows: M-1: This mode occurs during the interval $t_0$ - $t_1$ and $t_6$ - $T_s$. Within a switching period ($T_s$), the total duration of Mode-1 is ($T_s$ - $\beta DT_s$). In this mode, switch $S_4$ is *on*, whereas switches $S_1$, $S_2$, and $S_3$ are *off*. $D_3$ and $D_4$ freewheel the energy stored in $L_3$ and $L_1$, respectively. The instantaneous values of inductor currents, $i_{L1}$, $i_{L2}$, and $i_{L3}$, are continuously decreasing during this mode. The circuit for Mode-1 is shown in Figure 5a. M-2: This mode occurs during the interval $t_1$ - $t_2$ and $t_5$ - $t_6$. The duration of the interval for Mode-2 within a switching period is ($\beta DT_s$ - $\lambda DT_s$). In Mode-2 as shown in Figure 5b, switches $S_1$ and $S_4$ are in the *on* state, whereas switches $S_2$ and $S_3$ are in the *off* state. Similar to Mode-1, $D_3$ continuously freewheels the energy stored in $L_3$. Since $S_1$ is in *on* state, due to which the $V_{L1}$ is positive $L_1$ starts charging and $i_{L1}$ is increasing while same as Mode-1, $V_{L2}$ and $V_{L3}$ are negative so $i_{L2}$ and $i_{L3}$ are continuously decreasing.

M-3: This mode occurs during the interval $t_2$ - $t_3$ and $t_4$ - $t_5$. The duration of the interval for Mode-3 in a switching period is ($\lambda DT_s$ - $DT_s$). The circuit for Mode-3 is shown in Figure 5c. In this mode, except $S_3$, all the switches are in *on*-state, and all the diodes are reverse biased. Similar to Mode-2, $S_1$ remains in *on*-state in Mode-3, so $i_{L1}$ continuously increases while $S_2$ turned *on* and $S_3$ remains in *off* state so the $V_{L3}$ is positive and $i_{L3}$ is increasing. Since the $V_{L2}$ is negative so $i_{L2}$ is decreasing.

M-4: This mode occurs during the interval $t_3$ - $t_4$. The duration of the interval for Mode-4 in a switching period is $DT_s$. The circuit



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

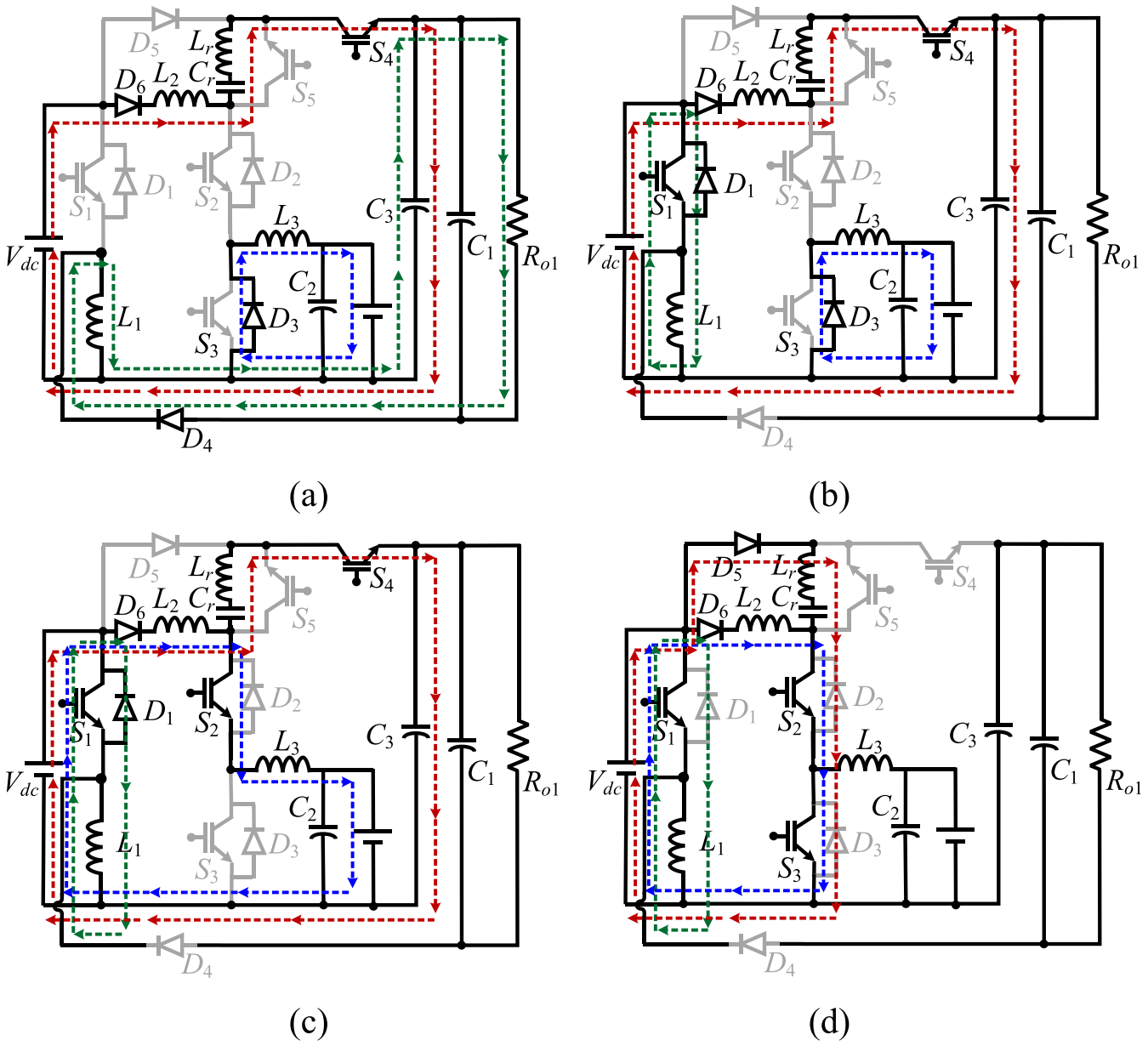



**FIGURE 5** | Operation of proposed HGFTM converter for *State*-I under healthy condition during (a) Mode-1, (b) Mode-2, (c) Mode-3, and (d) Mode-4.

for Mode-4 is shown in Figure 5d. In this mode, except $S_4$, all the switches are in *on*-state, and $D_5$ is forward biased. Similar to Mode-2 and Mode-3, $S_1$ remains *on*-state in Mode-4, so $L_1$ is continuously charging. Switches $S_2$ and $S_3$ both are in *on*-state so $L_2$ is charging but $L_3$ is discharging.

The characteristics of all the currents and switching states for the operation in *State*-I for $\beta > \lambda$ are summarized in Table 1.

### 2.2 | Operating Modes for *State*-II Under Healthy Condition

Two modes of operations are possible in *State*-II under healthy condition, as shown in Figure 6. The duration of both the modes, that is, Mode-7 (M-7) and Mode-8 (M-8), is shown in Figure 7. In this mode, the input dc supply ($V_{dc}$) is absent which is sensed by a low-cost sensor (explained in the next section), and thereafter, the requirement of power at output Port-1 is achieved by the battery connected at the output Port-2.

In *State*-II, switches $S_1$ and $S_2$ are *off*, whereas the switches $S_4$ and $S_5$ are turned *on*. A brief discussion of both the operating modes for *State*-II of the proposed converter is as follows: M-7: This mode occurs during the interval $t_0$ - $t_1$ and $t_2$ - $T_s$ shown in Figure 7. In a switching period ($T_s$), the duration of the interval for Mode-7 is ($T_s$ - $DT_s$). The switch $S_3$ is *off*, and the diode $D_2$ is forward bias and freewheels the energy stored in $L_3$ at the same time $L_1$ is charging through diode $D_4$ (Table 2).



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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** | Different modes of operation during *State*-I for $\beta > \lambda$: Charging.

| Modes of operation | Time interval | Inductor current characteristics | Active switches |
|---|---|---|---|
| Mode-1 (M-1) | $t_0 - t_1 = t_6 - T_s = \frac{T_s - \lambda DT_s}{2}$ | $i_{L_1}\downarrow, i_{L_2}\downarrow, i_{L_3}\downarrow$ | $D_3, D_4, S_4, D_6$ |
| Mode-2 (M-2) | $t_1 - t_2 = t_5 - t_6 = \frac{\beta DT_s - \lambda DT_s}{2}$ | $i_{L_1}\uparrow, i_{L_2}\downarrow, i_{L_3}\downarrow$ | $S_1, D_3, S_4, D_6$ |
| Mode-3 (M-3) | $t_2 - t_3 = t_4 - t_5 = \frac{\lambda DT_s - DT_s}{2}$ | $i_{L_1}\uparrow, i_{L_2}\downarrow, i_{L_3}\uparrow$ | $S_1, S_2, S_4, D_6$ |
| Mode-4 (M-4) | $t_3 - t_4 = DT_s$ | $i_{L_1}\uparrow, i_{L_2}\uparrow, i_{L_3}\downarrow$ | $S_1, S_2, S_3, D_5, D_6$ |

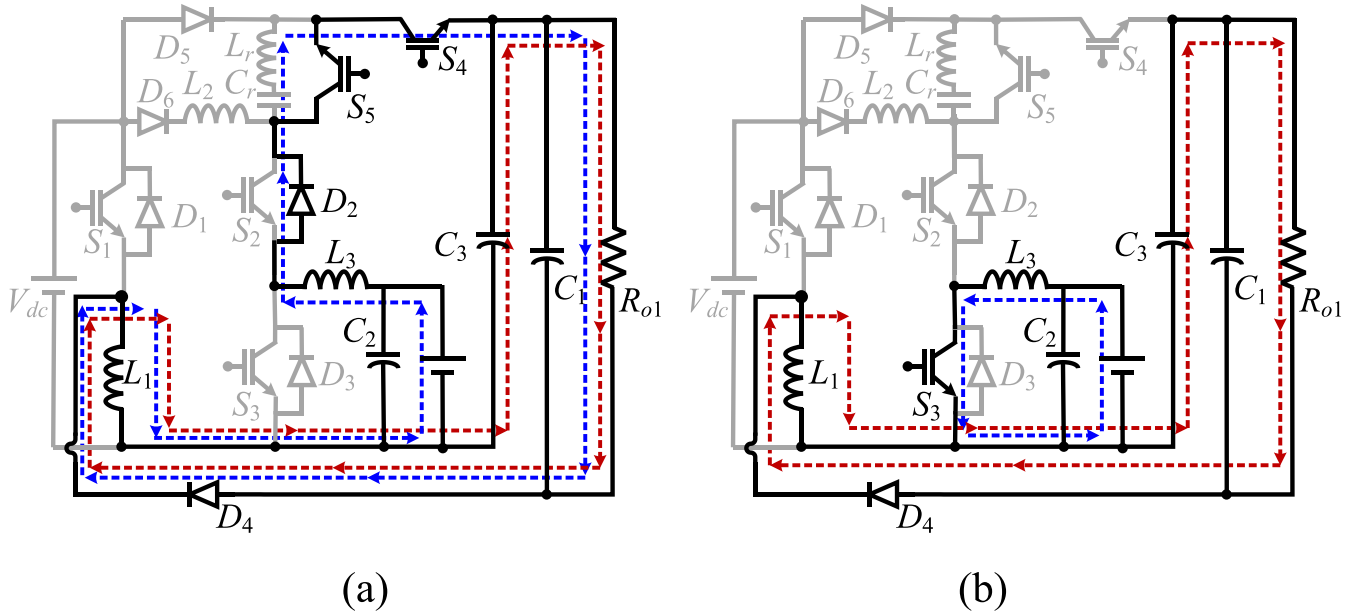



**FIGURE 6** | Operation of proposed HGFTM converter for *State*-II under healthy conditions during (a) Mode-7, and (b) Mode-8.




**FIGURE 7** | Switching logic of proposed HGFTM converter for *State*-II under healthy conditions.

M-8: This mode occurs in the interval $t_1 - t_2$. In a switching period ($T_s$), the duration of the interval for Mode-8 is $DT_s$. In this mode, the switch $S_3$ is *on*. Since $S_3$ is *on*, that is, the battery connected at output Port-2 is shorted with $L_3$ and so $L_3$ charges. The energy stored in $L_1$ discharges in Mode-8 through $D_4$.

Figure 8 shows the behavior of current characteristics with respect to modes of operation under *State*-I and *State*-II in healthy conditions. The nature of the input current is continuous. During Mode-4 under *State*-I, the dc input source is connected with $S_2$, $S_3$, $D_5$, $C_r$, and $L_r$ to reverse the polarity or reduce the voltage of the capacitor $C_r$. Since Mode-4 operates for $DT_s$ time period, therefore, the value of $L_r$ can be written as follows:

$$L_r \leq \frac{D^2}{f_S{}^2 \pi^2 C_r} \tag{1}$$

## 3 | State Space Averaged Modeling and Steady-State Gain Analysis of HGFTM Converter

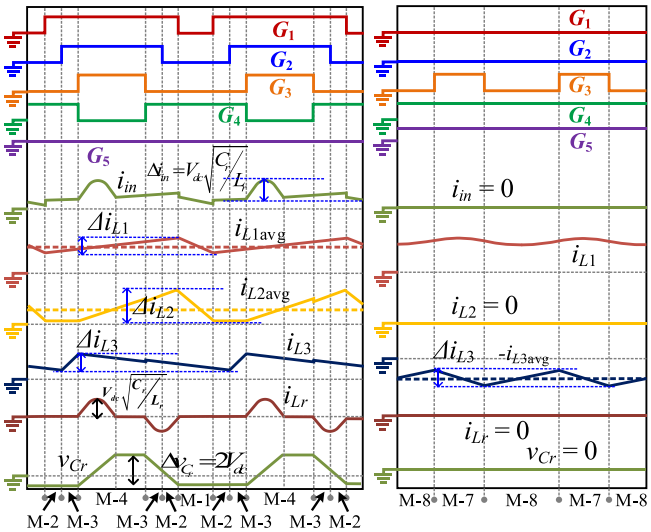
A detailed mathematical analysis of the proposed HGFTM dc-dc converter is presented with the duty factors $d_1$ (= $D$), $d_2$ (= $\lambda D$), and $d_3$ (= $\beta D$). For the *State*-I discussed in the previous section, the operating modes are characterized in three conditions. These conditions can be mathematically explained as follows:



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**TABLE 2** | Different modes of operation during *State*-II: Discharging.

| Modes of operation | Time interval | Inductor current characteristics | Active switches |
|---|---|---|---|
| Mode-7 (M-7) | $t_0 - t_1 = t_2 - T_s = \frac{T_s - DT_s}{2}$ | $i_{L_1}\uparrow, i_{L_2} = 0, i_{L_3}\downarrow$ | $D_2, D_4, S_4, S_5$ |
| Mode-8 (M-8) | $t_1 - t_2 = DT_s$ | $i_{L_1}\downarrow, i_{L_2} = 0, i_{L_3}\uparrow$ | $S_3, D_4$ |




(a) (b)

**FIGURE 8** | Behavior of current characteristics with respect to modes of operation under healthy condition for (a) *State*-I and (b) *State*-II.

$$\begin{aligned} &Condition{:}\,1 \;\rightarrow\; 1/D \geq \lambda \geq 1 \;\text{ and }\; 1/D \geq \beta \geq \lambda \\ &Condition{:}\,2 \;\rightarrow\; 1/D \geq \lambda \geq 1 \;\text{ and }\; \lambda \geq \beta \geq 1 \\ &Condition{:}\,3 \;\rightarrow\; 1/D \geq \lambda \geq 1 \;\text{ and }\; \beta \leq 1 \end{aligned} \tag{2}$$

A linear network can be expressed by expressing the first derivative of the state variables as a linear combination of the independent inputs of the system and the state variables themselves. There are eight state variables in the proposed converter. The state variables $x(t)$ are defined as $[i_{L1}, i_{L2}, i_{L3}, i_{Lr}, v_{C1}, v_{C2}, v_{C3}, v_{Cr}]^T$. The state space model can be written as follows:

$$\frac{d}{dt}x(t) = \dot{x} = Ax(t) + Bu(t) \tag{3a}$$

$$y(t) = Cx(t) \tag{3b}$$

where $A$, $B$, and $C$ stand for state, input, and output matrices. The voltage $v_{in}$ is input voltage, and the voltages across $C_1$ and $C_2$ are actually the load voltages $v_{o1}$ and $v_{o2}$, respectively. In the steady-state condition, we have the following:

$$\dot{x}(t) = 0$$

So, the SSA model given in (3a) and (3b) can be rewritten as follows:

$$X = -A^{-1}BU \tag{4a}$$

$$Y = \left(-CA^{-1}B + D\right)U \tag{4b}$$

The operating duty factors for different all the three conditions are given in Table 3. For the proposed converter, the state matrix $A$ and input matrix $B$ are averaged within the conditions $C_j$, for $j = 1, 2, 3$. The SSA model is developed with the respective values of duty factor given in Table 3 within any of the three conditions given in (2) as follows:

$$A|_{C_j} = \sum_{i=1}^{n=6} A_i M_i C_j \tag{5a}$$

$$B|_{C_j} = \sum_{i=1}^{n=6} B_i M_i C_j \tag{5b}$$

where $A_i$ and $B_i$ are the representation of the respective state and input matrices corresponding to the six operation modes, while $M_i C_j$ denotes the modulation factors associated with condition.

**TABLE 3** | Operating duty factor under different conditions.

| Modes of operation | *Condition*-1 | *Condition*-2 | *Condition*-3 |
|---|---|---|---|
| Mode-1: (M-1) | $1 - d_3$ | $1 - d_2$ | $1 - d_2$ |
| Mode-2: (M-2) | $d_3 - d_2$ | 0 | 0 |
| Mode-3: (M-3) | $d_2 - d_1$ | $d_3 - d_1$ | 0 |
| Mode-4: (M-4) | $d_1$ | $d_1$ | $d_3$ |
| Mode-5: (M-5) | 0 | $d_2 - d_3$ | $d_2 - d_1$ |
| Mode-6: (M-6) | 0 | 0 | $d_1 - d_3$ |



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

$$A_{C_1}=\begin{bmatrix} 0 & 0 & 0 & 0 & \frac{d_3-1}{L_1} & 0 & \frac{1-d_3}{L_1} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{L_r k(d_1-d_2)}{L_2} & \frac{d_2-1}{L_2^{'}}+\frac{(d_1-d_2)kL_3}{L_2} & \frac{d_2-1}{L_2^{'}}+\frac{(d_1-d_2)kL_3}{L_2} \\ 0 & 0 & 0 & 0 & 0 & \left(\frac{d_2-d_1-1}{L_3}+\frac{(d_1-d_2)k(L_2+L_r)}{L_2}\right) & k(d_2-d_1) & k(d_2-d_1) \\ 0 & 0 & 0 & 0 & 0 & k(d_2-d_1) & \left(\frac{d_2-1}{L_2^{'}}+\frac{(d_1-d_2)k(L_2+L_3)}{L_2}\right) & \left(\frac{d_2-1}{L_2^{'}}-\frac{d_1}{L_r}+\frac{(d_1-d_2)k(L_2+L_3)}{L_2}\right) \\ \frac{1-d_3}{C_1} & 0 & 0 & 0 & -\frac{1}{R_1C_1} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{C_2} & 0 & 0 & -\frac{1}{R_2C_2} & 0 & 0 \\ \frac{d_3-1}{C_3} & \frac{1-d_1}{C_3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{C_r} & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{6a}$$

$$B_{C_1}=\begin{bmatrix}\left(\frac{d_3}{L_1}\right) & \left(\frac{1-d_2}{L_2^{'}}+\frac{d_1}{L_2}+\frac{(d_2-d_1)k(L_3+L_r)}{L_2}\right) & \left(\frac{(d_2-d_1)kL_r}{L_2}\right) & \left(\frac{1-d_2}{L_2^{'}}-\frac{d_1}{L_r}+\frac{(d_2-d_1)kL_3}{L_2}\right) & 0 & 0 & 0 & 0\end{bmatrix}^T$$

$$A_{C_2}=\begin{bmatrix} 0 & 0 & 0 & 0 & \frac{d_3-1}{L_1} & 0 & \frac{1-d_3}{L_1} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{L_r k(d_1-d_2)}{L_2} & \frac{d_2-1}{L_2^{'}}+\frac{(d_1-d_2)kL_3}{L_2} & \frac{d_2-1}{L_2^{'}}+\frac{(d_1-d_2)kL_3}{L_2} \\ 0 & 0 & 0 & 0 & 0 & \left(\frac{d_2-d_1-1}{L_3}+\frac{(d_1-d_2)k(L_2+L_r)}{L_2}\right) & k(d_2-d_1) & k(d_2-d_1) \\ 0 & 0 & 0 & 0 & 0 & k(d_2-d_1) & \left(\frac{d_2-1}{L_2^{'}}+\frac{(d_1-d_2)k(L_2+L_3)}{L_2}\right) & \left(\frac{d_2-1}{L_2^{'}}-\frac{d_1}{L_r}+\frac{(d_1-d_2)k(L_2+L_3)}{L_2}\right) \\ \frac{1-d_3}{C_1} & 0 & 0 & 0 & -\frac{1}{R_1C_1} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{C_2} & 0 & 0 & -\frac{1}{R_2C_2} & 0 & 0 \\ \frac{d_2-1}{C_3} & \frac{1-d_2}{C_3} & 0 & \frac{d_2-d_1}{C_3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{C_r} & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{6b}$$

$$B_{C_2}=\begin{bmatrix}\left(\frac{d_3}{L_1}\right) & \left(\frac{1-d_2}{L_2^{'}}+\frac{d_1}{L_2}+\frac{(d_2-d_1)k(L_3+L_r)}{L_2}\right) & \left(\frac{(d_2-d_1)kL_r}{L_2}\right) & \left(\frac{1-d_2}{L_2^{'}}-\frac{d_1}{L_r}+\frac{(d_2-d_1)kL_3}{L_2}\right) & 0 & 0 & 0 & 0\end{bmatrix}^T$$



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

$$A_{C_3}=\begin{bmatrix} 0 & 0 & 0 & 0 & \frac{d_3-1}{L_1} & 0 & \frac{1-d_3}{L_1} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{L_r k(d_1-d_2)}{L_2} & \frac{d_2-1}{L_2^{'}}+\frac{(d_1-d_2)kL_3}{L_2} & \frac{d_2-1}{L_2^{'}}+\frac{(d_1-d_2)kL_3}{L_2} \\ 0 & 0 & 0 & 0 & 0 & \left(\frac{d_2-d_1-1}{L_3}+\frac{(d_1-d_2)k(L_2+L_r)}{L_2}\right) & k(d_2-d_1) & k(d_2-d_1) \\ 0 & 0 & 0 & 0 & 0 & k(d_2-d_1) & \left(\frac{d_2-1}{L_2^{'}}+\frac{(d_1-d_2)k(L_2+L_3)}{L_2}\right) & \left(\frac{d_2-1}{L_2^{'}}-\frac{d_1}{L_r}+\frac{(d_1-d_2)k(L_2+L_3)}{L_2}\right) \\ \frac{1-d_3}{C_1} & 0 & 0 & 0 & -\frac{1}{R_1C_1} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{C_2} & 0 & 0 & -\frac{1}{R_2C_2} & 0 & 0 \\ \frac{d_3+d_2-d_1-1}{C_3} & \frac{1-d_2}{C_3} & 0 & \frac{d_2-d_1}{C_3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{C_r} & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$B_{C_3}=\begin{bmatrix} \left(\frac{d_3}{L_1}\right) & \left(\frac{1-d_2}{L_2^{'}}+\frac{d_1}{L_2}+\frac{(d_2-d_1)k(L_3+L_r)}{L_2}\right) & \left(\frac{(d_2-d_1)kL_r}{L_2}\right) & \left(\frac{1-d_2}{L_2^{'}}-\frac{d_1}{L_r}+\frac{(d_2-d_1)kL_3}{L_2}\right) & 0 & 0 & 0 & 0 \end{bmatrix}^T \tag{6c}$$

where $A_{C1}$, $A_{C2}$, and $A_{C3}$ are averaged state matrices for *Condition*-1, *Condition*-2, and *Condition*-3, respectively; $B_{C1}$, $B_{C2}$, and $B_{C3}$ are averaged input matrices for *Condition*-1, *Condition*-2, and *Condition*-3, respectively; and $k$ is defined as follows:

$$k=\frac{L_2}{L_2L_r+L_3L_r+L_2L_3} \tag{7}$$

Using (4) and (6), the steady-state voltage gains at the output Port-1 (i.e., $A_{V\text{boost}}=v_{o1}/V_{dc}$), and Port-2 (i.e., $A_{V\text{buck}}=v_{o2}/V_{dc}$) can be derived using initial conditions of resonant component and neglecting the smaller values of $L_r$. It can be written as follows:

$$A_{V\text{boost}}=\frac{v_{o1}}{V_{dc}}=\left(\frac{2-D}{1-D}+\frac{\beta D}{1-\beta D}\right) \tag{8}$$

$$A_{V\text{buck}}=\frac{v_{o2}}{V_{dc}}=\frac{(\lambda-1)D}{1-D} \tag{9}$$

Further, the steady-state values of state variables such as $i_{L1}$, $i_{L2}$, $i_{L3}$, and $v_{C3}$ can also be derived using (4) and (6).

## 4 | Characterization of HGFTM Converter Under OCSF Conditions

The proposed HGFTM converter is capable of operating under single OCSF conditions. The operation principle of the converter is discussed with respect to the behavior of currents. The OCSFs under each semiconductor switches $S_1$, $S_2$, and $S_3$ are presented in the following subsections: *State-I under OCSF condition in $S_1$:* During open circuit switch fault in $S_1$ (i.e., $SF_{1OC}$), modes M-1, M-5, and M-6 occur as shown in Figures 5a and 9a,b, respectively, where modes M-5 and M-6 also occurred in healthy conditions, if $\beta \le 1$ as shown in the switching logic diagram presented in Figure 4c. The updated switching modes during $SF_{1OC}$ are shown in Figure 10. In this condition, the power flow from the buck-boost converter is zero; only the multiport converter is used to transfer the power to the output Port-2 and $C_3$, and further, the energy stored in $C_3$ is transferred to the output Port-1 through $L_1$.

Further, the gain at output Port-2 during $SF_{1OC}$ can be derived as (10).

$$A_{v\text{buck}}\big|_{SF_{1OC}}=\frac{v_{o2}}{V_{dc}}\bigg|_{SF_{1OC}}=\frac{D(\lambda-1)}{1-D} \tag{10}$$

Since there is $SF_{1OC}$ condition, that is, the $S_1$ is turned off for all time, which is equivalent to the condition when $\beta D=0$ under healthy conditions. Hence, the gain at output Port-1 during $SF_{1OC}$ can be derived by putting $\beta D=0$ in (8) that is given in (11).

$$A_{V\text{boost}}\big|_{SF_{1OC}}=\frac{v_{o1}}{V_{dc}}\bigg|_{SF_{1OC}}=\left(\frac{2-D}{1-D}\right) \tag{11}$$

*State-I under OCSF condition in $S_2$:* The updated switching modes during open circuit switch fault in $S_2$ (i.e., $SF_{2OC}$) are shown in Figure 11. In a complete switching period, modes M-1a, M-2a, and M-9 occur under $SF_{2OC}$, where modes M-1a and



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

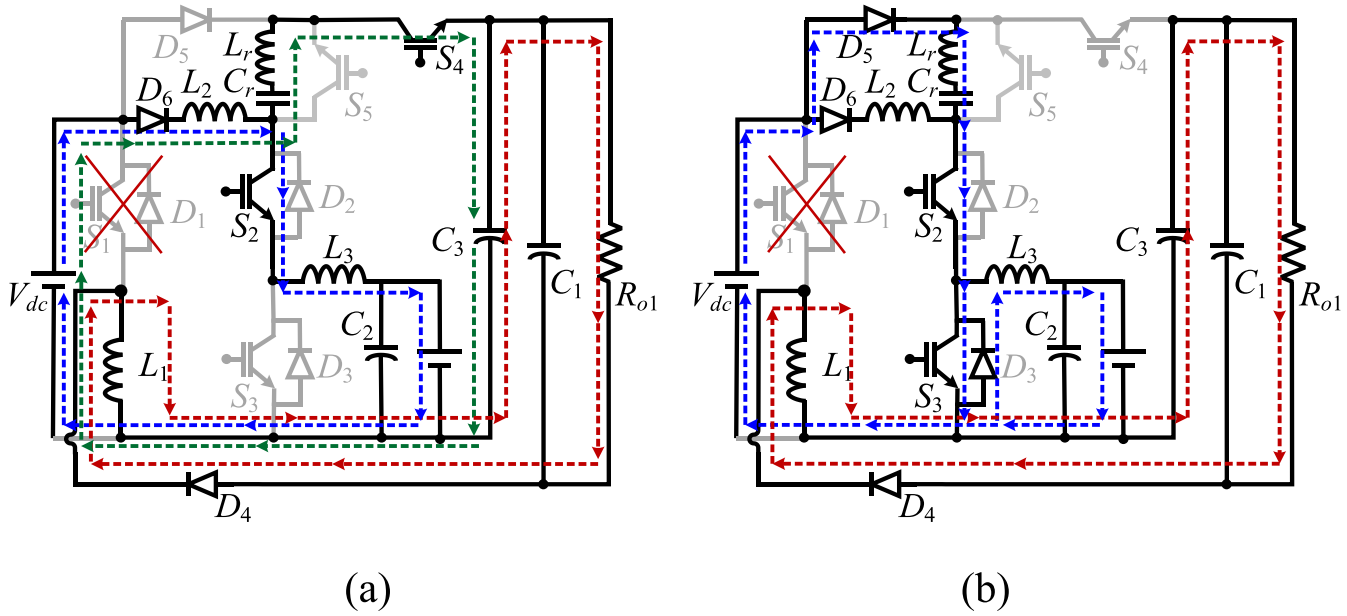



**FIGURE 9** | Operation of proposed HGFTM converter under *State*-I during $SF_{1OC}$ condition for (a) Mode-5 and (b) Mode-6.

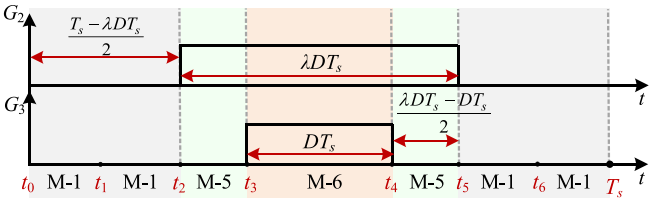


**FIGURE 10** | Switching logic of proposed HGFTM converter for *State*-I under $SF_{1OC}$ condition.

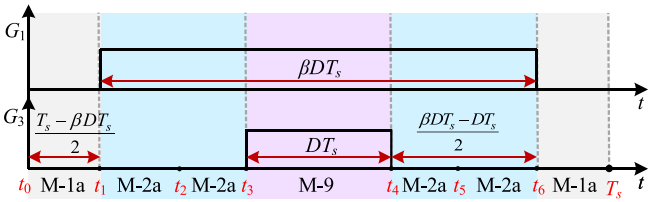


**FIGURE 11** | Switching logic of proposed HGFTM converter for *State*-I under $SF_{2OC}$ condition.

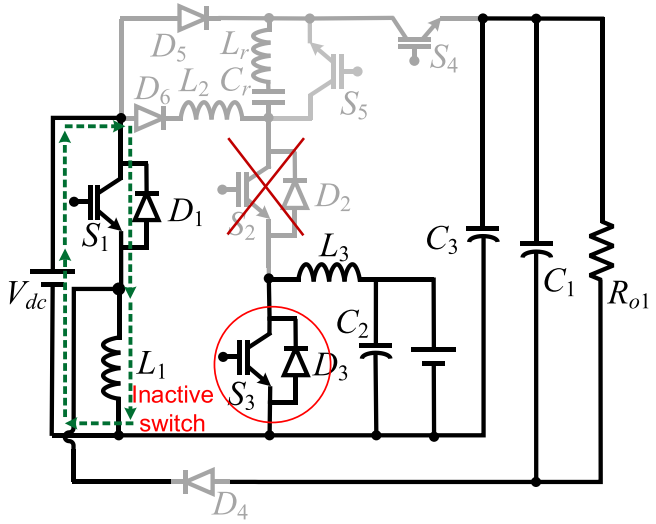



**FIGURE 12** | Operation of proposed HGFTM converter under *State*-I during $SF_{2OC}$ condition for Mode-9.

M-2a are the slight modification of modes M-1 (i.e., Figure 5a) and M-2 (i.e., Figure 5b), respectively. The current paths for M-1a and M-2a under steady-state operation are only dotted paths shown in green in Figure 5a,b, respectively. Figure 12 shows the operation of the converter under Mode-9 during $SF_{2OC}$. The duty cycle $S_3$ is $D$, but the switching of $S_3$ will not affect the gain of the output Port-1 during $SF_{2OC}$. Hence, the gain of the output Port-1 during $SF_{2OC}$ is derived and given in (12).

$$A_{V\text{boost}}\big|_{SF_{2OC}} = \frac{v_{o1}}{V_{dc}}\bigg|_{SF_{2OC}} = \left(\frac{\beta D}{1-\beta D}\right) \tag{12}$$

Since there is an $SF_{2OC}$ condition, that is, the $S_2$ is off for all time and the output Port-2 is isolated from the main supply as shown in Figure 12. Hence, the gain at output Port-2 during $SF_{2OC}$ is zero.

$$A_{v\text{buck}}\big|_{SF_{2OC}} = \frac{v_{o2}}{V_{dc}}\bigg|_{SF_{2OC}} = 0 \tag{13}$$

*State-I under OCSF condition in $S_3$:* The updated switching modes during open circuit switch fault in $S_3$ ($SF_{3OC}$) are shown in Figure 13. Modes M-1, M-2, and M-3 occur during $SF_{3OC}$.



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

All three modes appear under healthy conditions as shown in Figures 5 a–c, respectively.

The voltage gain at output Port-1 and output Port-2 during $SF_{3OC}$ is derived and given in (14) and (15) respectively.

$$A_{V\text{boost}}\big|_{SF_{3OC}} = \frac{v_{o1}}{V_{dc}}\bigg|_{SF_{3OC}} = \left(\frac{1}{1-\beta D}\right) \tag{14}$$

$$A_{v\text{buck}}\big|_{SF_{3OC}} = \frac{v_{o2}}{V_{dc}}\bigg|_{SF_{3OC}} = \lambda D \tag{15}$$

The current and voltage characteristics of the proposed HGFTM converter under OCSF condition in $S_1$, $S_2$, and $S_3$ are shown in Figure 14a–c, respectively. The output voltage gains of output Port-1 ($A_{v\text{boost}}$) and output Port-2 ($A_{v\text{buck}}$) under healthy conditions and OCSF conditions are summarized in Table 4. Figure 15 shows the gains of the converter for the range of duty ratio and different values of duty factor (β) under different operating conditions and without the consideration of circuits parasitic. It can be seen from Figure 15a that for $\beta > 1$ and $\beta D \leq 1$, the gain of the converter is greater than six for a duty ratio greater than 0.8 under all the operating conditions. The gain of the proposed converter can be further increased with the increase in the value of duty factor $\beta$, which can be seen in Figure 15c. It can also be seen from Figure 15c that the voltage gain of the converter under $SF_{1OC}$ is independent of the value of duty factor $\beta$. Figure 15d shows the voltage gain of the output Port-2. The gain profile of the converter at output Port-2 is similar under healthy operation and $SF_{1OC}$ fault conditions, whereas it varies linearly under $SF_{3OC}$. The voltage stress across each switch in the proposed converter is calculated under *Condition*-3 and presented in Table 5 to optimize the circuit design for reliable, efficient and safe operation of the system.

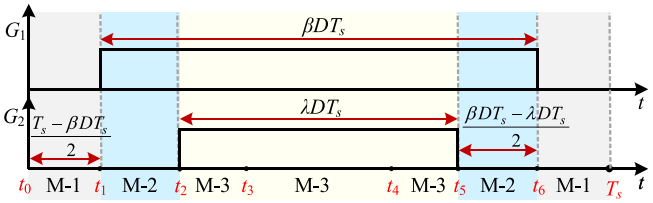



**FIGURE 13** | Switching logic of proposed HGFTM converter for *State*-I under $SF_{3OC}$ condition.

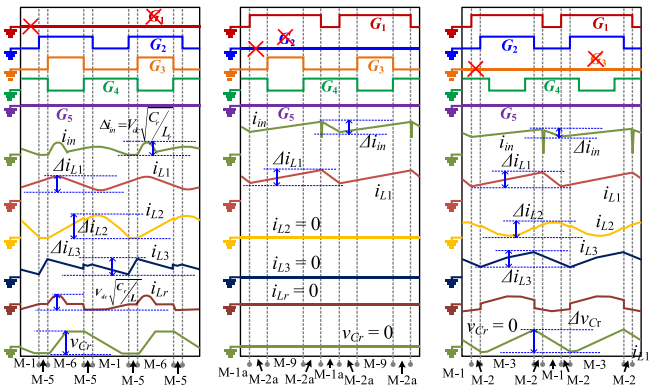



**FIGURE 14** | Behavior of current characteristics with respect to modes of operation under OCSF condition in (a) $SF_{1OC}$, (b) $SF_{2OC}$, and (c) $SF_{3OC}$.

**TABLE 4** | Output voltage gains of HGFTM DC-DC converter under different operating conditions.

| Voltage gains | Healthy condition | $SF_{1OC}$ | $SF_{2OC}$ | $SF_{3OC}$ |
|---|---|---|---|---|
| Output Port-1 | $\left(\frac{2-D}{1-D}+\frac{\beta D}{1-\beta D}\right)$ | $\left(\frac{2-D}{1-D}\right)$ | $\left(\frac{\beta D}{1-\beta D}\right)$ | $\left(\frac{1}{1-\beta D}\right)$ |
| Output Port-2 | $\frac{D(\lambda-1)}{1-D}$ | $\frac{D(\lambda-1)}{1-D}$ | 0 | $\lambda D$ |

## 5 | Results and Discussion

A prototype of the proposed HGFTM dc-dc converter is developed to validate the performance of the proposed converter. The laboratory type prototype model of an experimental setup is shown in Figure 16. An isolated driver circuit is developed using TLP350 optocouplers, and isolated dc-dc converters (MORNSUNQA153-1509R3) are used to isolate the ground of noncommon emitter terminal of the IGBTs. The SGT40N60FD2 IGBT power switches and RHRP30120 hyper-fast freewheeling diodes are used for proper switching in the developed converter. The GW-Instek ASR-2100 programmable AC/DC power source is used as an input dc supply for the proposed converter. The converter is tested in the laboratory for the input voltage range of 12–35 V dc voltage, and the output level at the output Port-1 (step-up side) is 36–96 V dc and at the output Port-2 (step-down side) is 6 V–14 V dc. The system is tested for 100 W output power. The system parameters used to develop the laboratory prototype are given in Table 6.

A unit-magnitude voltage feedback low-cost sensor circuit is designed using an optocoupler TLP350 to sense the availability of input supply. The lower sensor output enables the discharging property of the proposed converter. The sensor circuit gives *logic* '0' when the source voltage is below 12 V. The cutoff range can be set with the help of the input voltage divider circuit of the low-cost sensor. The schematic diagram of the unit-magnitude sensor circuit is shown in Figure 17. The closed-loop constant voltage operation is achieved using LV25 voltage sensor circuits at both output ports.

The experimental results of the proposed HGFTM dc-dc converter are presented in Figures 18–23. The digital probe connected at channel 1 has 8 pins, out of which 5 pins are used for gate signals for switches $S_1$ to $S_5$ (i.e., $G_1$, $G_2$, $G_3$, $G_4$, and $G_5$). Four current probes are connected at Channels 2–5 for the current measurement of $i_{in}$, $i_{L1}$, $i_{L2}$, and $i_{L3}$, respectively. Three voltage differential probes are connected at Channels 6–8 for the



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

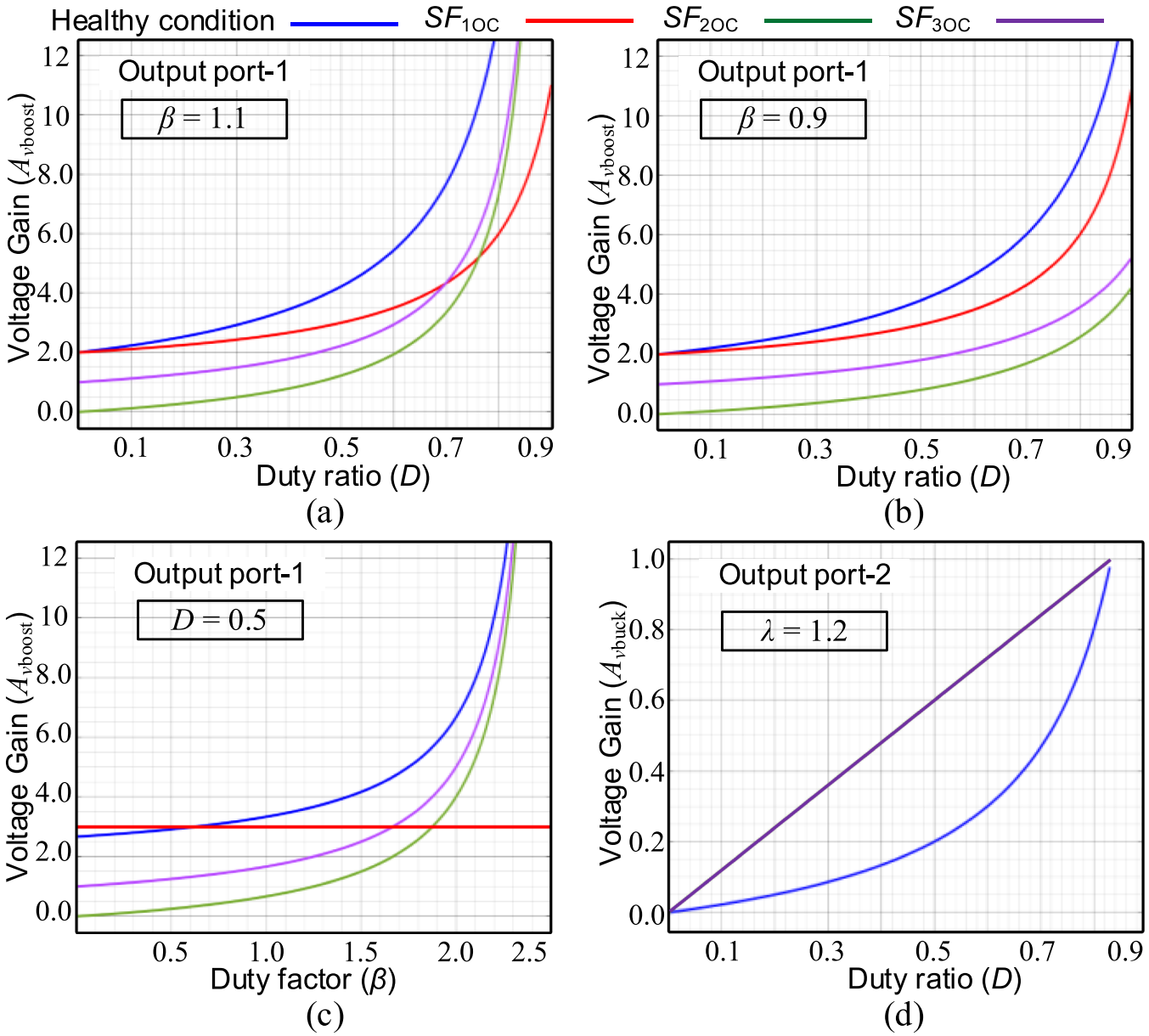



**FIGURE 15** | Voltage gains of the proposed converter under different operating conditions and ideal circuit parameters at output Port-1 ($A_{vboost}$) with (a, b) $\beta = 1.1$, 0.9 at varying duty ratio, (c) $D = 0.5$ at different values of duty factor, and at output Port-2 ($A_{vbuck}$) with (d) $\lambda = 1.2$ at varying duty ratio.

**TABLE 5** | Maximum voltage stress in the switches.

| Voltage stress at $S_1$: $V_{S1}$ | Voltage stress at $S_2$: $V_{S2}$ | Voltage stress at $S_3$: $V_{S3}$ |
|---|---|---|
| $\frac{V_{dc}}{1-\beta D}$ | $\frac{V_{dc}}{1-\lambda D}$ | $\frac{V_{dc}}{1-\lambda D}$ |

measurement of $V_{dc}$, $v_{o1}$, and $v_{o2}$, respectively. Figure 18 shows mainly the high gain characteristics of the proposed converter, where the input voltage $V_{dc}$ is 18 V and the output Port-1 ($v_{o1}$) is 96 V which is about six times the voltage gain of the input supply voltage. Other output parameters $i_{in}$, $i_{L1}$, $i_{L2}$, and $i_{L3}$ are the input current and inductor currents in $L_1$, $L_2$, and $L_3$, respectively. At the same time, the voltage at output Port-2 ($v_{o2}$) is set to 13.2 V for charging the battery under constant voltage mode, and $G_5$ is set to low (i.e., $S_5$ is *off*). It can be verified using the inductor current $i_{L3}$, which is positive (current entering the positive terminal of the battery connected at output Port-2).

The fault tolerant characteristics of the converter under different OCSFs like $SF_{1OC}$, $SF_{2OC}$, and $SF_{3OC}$ at $V_{dc} = 24$ V are validated and shown in Figures 19–21, respectively. Figure 19 shows that the $SF_{1OC}$, triggered at −0.2 s. After the fault is triggered, the $i_{L1}$ becomes zero, but there is no change in the output characteristics of the proposed converter and $i_{L3}$ is positive that is, the proposed converter tolerates the $SF_{1OC}$ and continuously supplies $v_{o1} = 48$ V and $v_{o2} = 13.5$ V before and after the fault occurs as shown in the zoomed section of Figure 19. The $SF_{2OC}$ triggered



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

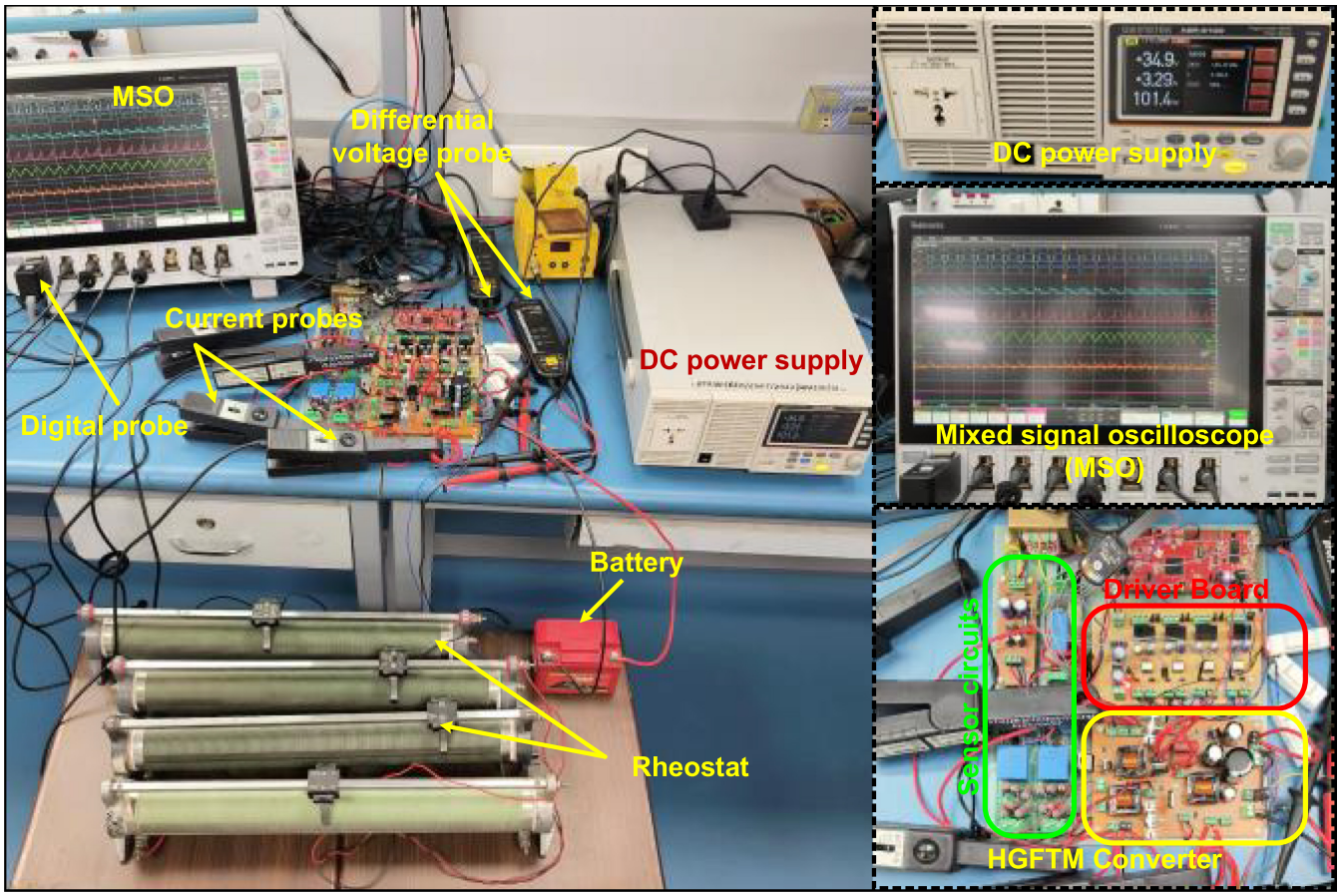



**FIGURE 16** | A 100 W prototype laboratory setup of proposed HGFTM dc-dc converter.

**TABLE 6** | System parameters.

| Name of component | Quantity | Specifications |
|---|---|---|
| Power semiconductor switch (with diode) | 03 | 40 N60FD2 |
| Power semiconductor switch (without diode) | 02 | 40 N60 |
| High-frequency diodes | 02 | RHR30120 |
| Inductors (μH) | 03 | $L_1 = 165$, $L_2 = 590$, $L_3 = 206$ |
| Capacitors (μF) | 03 | $C_1 = 680$, $C_2 = 470$, $C_3 = 200$ |
| Resonant elements ($L_r$, $C_r$) | 01 | $L_r = 10\,\mu$H, $C_r = 4.4\,\mu$F/250 V (polyester film capacitor) |
| Load | 01 | Output Port-01: $75W_{peak}$ Output Port-02: $25W_{peak}$ |

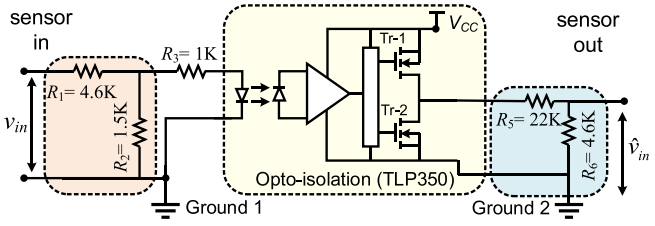



**FIGURE 17** | Schematic diagram for unit-magnitude voltage sensor circuit.

at 0.1 s as shown in Figure 20, the $i_{L3}$ becomes zero, that is, no energy is supplied from the source to the battery, and the battery is neither charging nor discharging. Since the energy supplied to the battery becomes zero, the power delivered from the input source decreases ($i_{in}$ decreases). As shown in the zoomed section of Figure 20, before $SF_{2OC}$ occurs, $v_{o1} = 48$ V and $v_{o2} = 13.5$ V, and after $SF_{2OC}$ occurs, $v_{o1} = 48$ V and $v_{o2} = 13.3$ V, where $v_{o2} = 13.3$ V is the open-circuit battery voltage, and hence, the proposed converter tolerates the $SF_{2OC}$ condition and continuously supply



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

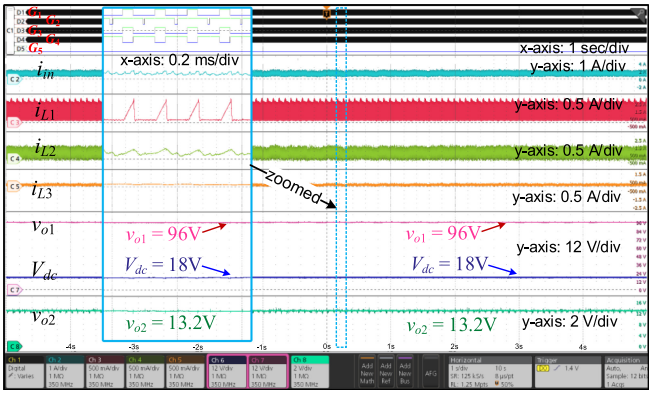



**FIGURE 18** | Generation of high-voltage gain (i.e., $v_{o1} \approx 6 \times V_{dc}$) at boost side port with the proposed HGFTM converter.

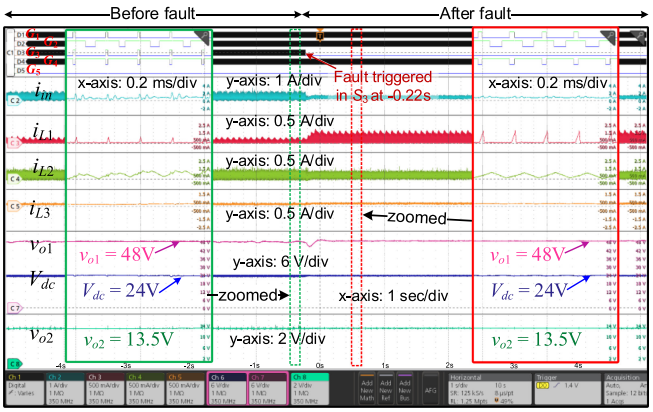



**FIGURE 21** | Fault tolerant characteristics of proposed HGFTM converter under open circuit switch fault in $S_3$ ($SF_{3OC}$) at $V_{dc} = 24$ V.

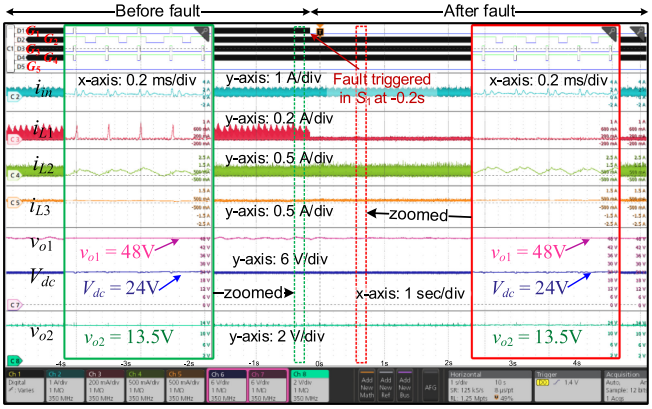



**FIGURE 19** | Fault tolerant characteristics of proposed HGFTM converter under open circuit switch fault in $S_1$ ($SF_{1OC}$) at $V_{dc} = 24$ V.

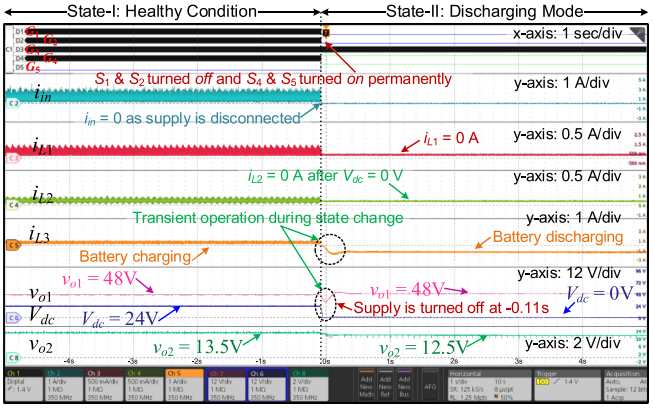



**FIGURE 22** | Bidirectional converter operation: *State*-I at $V_{dc} = 24$ V and State-II at $V_{dc} = 0$ V with continuous load voltage $V_{o1} = 48$ V.




**FIGURE 20** | Fault tolerant characteristics of proposed HGFTM converter under open circuit switch fault in $S_2$ ($SF_{2OC}$) at $V_{dc} = 24$ V.

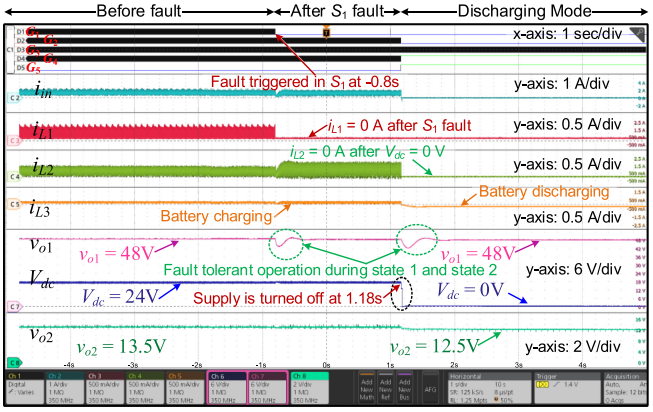



**FIGURE 23** | Fault tolerant characteristics of proposed HGFTM converter under OCSF in $S_1$ ($SF_{1OC}$) at $V_{dc} = 24$ V and bidirectional operation of the converter.

the desired voltage at the output Port-1. Figure 21 shows the fault tolerant behavior of the converter, the $SF_{3OC}$ is triggered at −0.22 s, $i_{L3}$ has no change, and it is positive that is, the battery connected at output Port-2 is continuously charging. The output Port-1 voltage ($v_{o1}$) is maintained at 48 V even after the $SF_{3OC}$.

The proposed converter also has the bidirectional operation capability which can be validated by the result shown in Figures 22 and 23. Before $t = -0.11$ s, the converter was operating in *State*-I and that too in healthy condition as shown in Figure 22. During *State*-I with healthy operation, switches



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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 7** | Comparative analysis of the proposed HGFTM DC-DC converter with similar converters.

| Reference | No. of power switches | No. of inductors | No. of capacitors | No. of diodes | Voltage gain | No. of output ports |
|---|---|---|---|---|---|---|
| [19] | 1 | 3 | 5 | 4 | High | 1 |
| [26] | 3 | 3 | 4 | 1 | Medium | 2 |
| [27] | 4 | 2 | 3 | 3 | Low | 2 |
| [28] | 4 | 1 | 1 | 2 | High | 1 |
| [29] | 3 | 3 | 3 | 1 | Medium | 2 |
| Proposed | 3, 2 | 3, 1 | 3, 1 | 2, 1 | Medium | 2 |

$S_1$ to $S_4$ are operating. The $v_{o1}$ at this case is 48 V, and battery connected at Port-2 is charging. At $t=-0.11$ s, the $V_{dc}$ is disconnected, and the converter is now working in bidirectional mode (*State*-II). In this case, $S_1$ and $S_2$ are turned *off* and $S_4$ and $S_5$ are turned *on* permanently. In this case, only $S_3$ is switching to get the desired output voltage at Port-1. The battery is discharging after $t=-0.11$ s, and the voltage at Port-1 is maintained at 48 V. In Figure 23, before the fault occurs ($t<-0.8$ s), the switches $S_1$ to $S_4$ operate, and the converter operates in healthy condition with $V_{dc}=24$ V, $v_{o1}=48$ V, and $v_{o2}=13.5$ V. Since $i_{L3}$ is positive, the battery is charging. The $SF_{1OC}$ triggered at $t=-0.8$ s switches $S_2$ to $S_4$ operates, and the $i_{L3}$ is positive; that is, the battery is charging and the output voltages $v_{o1}$ and $v_{o2}$ are maintained at 48 and 13.5 V. During $SF_{1OC}$, the input supply to the converter ($V_{dc}$) is turned off at $t=1.18$ s. In this case, $S_1$ and $S_2$ are turned *off*, $S_4$ and $S_5$ are turned *on*, and $S_3$ operates to maintain the $v_{o1}$ at the desired level using the battery as a source connected at output Port-2. The $i_{L3}$ is negative; that is, the battery is discharging and the battery voltage ($v_{o2}$) suddenly dropped by 1 V, that is, 13.5 to 12.5 V. Table 7 presents the superiorities of the proposed converter and compared with the recent research. The proposed converter has multiple output ports for different applications, and it has the fault-tolerant capability for multiple switches with high-voltage gain characteristics.

## 6 | Conclusion

In this article, a novel HGFTM bidirectional dc-dc converter is presented and validated using a 100-W laboratory prototype. The converter uses a time multiplexing control technique to transfer power at output Port-1 (at step-up voltage) and output Port-2 (at step-down voltage). It is a combination of conventional boost and buck converter topologies with a noncascading structure to provide high gain with improved efficiency. Without any additional device requirement, the proposed converter provides OCSF tolerant capability. This provides controlled output voltages for a wide range of applications. The voltage gain characteristics of the proposed converter are presented for healthy conditions and OCSF conditions for both the output ports. Finally, this topology is compared and presented in Table 6. The proposed converter is a high-gain converter and have the fault tolerant capability for multiple switches.

### Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

### References


1. E. Ribeiro, A. J. M. Cardoso, and C. Boccaletti, "Fault-Tolerant Strategy for a Photovoltaic dc–dc Converter," *IEEE Transactions on Power Electronics* 28, no. 6 (2013): 3008–3018.

2. E. Jamshidpour, P. Poure, and S. Saadate, "Photovoltaic Systems Reliability Improvement by Real-Time FPGA-Based Switch Failure Diagnosis and Fault Tolerant dc–dc Converter," *IEEE Transactions on Industrial Electronics* 62, no. 11 (2015): 7247–7255.

3. O. Khan, W. Xiao, and M. S. E. Moursi, "A New PV System Configuration Based on Submodule Integrated Converters," *IEEE Transactions on Power Electronics* 32, no. 5 (2017): 3278–3284.

4. Y. Song and B. Wang, "Survey on Reliability of Power Electronic Systems," *IEEE Transactions on Power Electronics* 28, no. 1 (2013): 591–604.

5. S. Nie, X. Pei, Y. Chen, and Y. Kang, "Fault Diagnosis of PWM dc–dc Converters Based on Magnetic Component Voltages Equation," *IEEE Transactions on Power Electronics* 29, no. 9 (2014): 4978–4988.

6. M. Kumar, "Multiple Open Switch Fast Fault Detection and Localization Algorithm for Tolerant CHB-MLI," *IEEE Transactions on Transportation Electrification* 10, no. 3 (2024): 6789–6800.

7. M. Kumar, "Time-Domain Characterization and Detection of Open-Circuit Faults for the H-Bridge Power Cell," *IEEE Transactions on Power Electronics* 37, no. 2 (2022): 2152–2164.

8. K. Varesi, N. Hassanpour, and S. Saeidabadi, "Novel High Step-Up DC–DC Converter With Increased Voltage Gain per Devices and Continuous Input Current Suitable for DC Microgrid Applications," *International Journal of Circuit Theory and Applications* 48, no. 10 (2020): 1820–1837.

9. J. Ahmad, M. D. Siddique, A. Sarwar, C. H. Lin, and A. Iqbal, "A High Gain Noninverting DC–DC Converter With Low Voltage Stress for Industrial Applications," *International Journal of Circuit Theory and Applications* 49, no. 12 (2021): 4212–4230.

10. M. Gleissner, S. Member, M. M. Bakran, and S. Member, "Design and Control of Fault-Tolerant Non-Isolated Multiphase Multilevel DC-DC Converters for Automotive Power Systems," *IEEE Transactions on Industry Applications* 52, no. 2 (2016): 1785–1795.

11. Y. Chen, X. Pei, S. Nie, and Y. Kang, "Monitoring and Diagnosis for the DC–DC Converter Using the Magnetic Near Field Waveform," *IEEE Transactions on Industrial Electronics* 58, no. 5 (2011): 1634–1647.




1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

12. S. Cruz, M. Ferreira, A. Mendes, and A. J. M. Cardoso, "Analysis and Diagnosis of Open-Circuit Faults in Matrix Converters," *IEEE Transactions on Industrial Electronics* 58, no. 5 (2011): 1648–1661.

13. S. Khwan-On, L. De Lillo, L. Empringham, and P. Wheeler, "Fault Tolerant Matrix Converter Motor Drives With Fault Detection of Open Switch Faults," *IEEE Transactions on Industrial Electronics* 59, no. 1 (2012): 257–268.

14. L. Costa, G. Buticchi, and M. Liserre, "A Fault-Tolerant Series-Resonant DC-DC Converter," *IEEE Transactions on Power Electronics* 32, no. 2 (2017): 900–905.

15. K. Ambusaidi, V. Pickert, and B. Zahawi, "New Circuit Topology for Fault Tolerant H-Bridge DC-DC Converter," *IEEE Transactions on Power Electronics* 25, no. 6 (2010): 1509–1516.

16. P. K. Pedapati and A. Hemachander, "Fault-Tolerant Reconfigurable Multiport Converter for Hybrid Power Applications," *IEEE Transactions on Circuits and Systems II: Express Briefs* 71, no. 3 (2024): 1496–1500.

17. W. Xiong, Z. Yan, D. Tang, W. Zhou, and R. Mai, "A Hybrid Topology IPT System With Partial Power Processing for CC-CV Charging," *IEEE Transactions on Power Electronics* 39, no. 1 (2024): 1701–1712.

18. E. Chiodo and D. Lauria, "Some Basic Properties of the Failure Rate of Redundant Reliability Systems in Industrial Electronics Applications," *IEEE Transactions on Industrial Electronics* 62, no. 8 (2015): 5055–5062.

19. F. Mohammadi, H. Rastegar, and M. Pichan, "Circuit Design and Implementation of Fault-Tolerant Step-Up Converter With Variable Voltage Gain," *International Journal of Circuit Theory and Applications* 50, no. 12 (2022): 4395–4418.

20. R. Stala, Z. Waradzyn, A. Penczek, A. Mondzik, and A. Skała, "A Switched-Capacitor DC–DC Converter With Variable Number of Voltage Gains and Fault-Tolerant Operation," *IEEE Transactions on Industrial Electronics* 66, no. 5 (2019): 3435–3445.

21. D. Guilbert, M. Guarisco, A. Gaillard, A. N'Diaye, and A. Djerdir, "FPGA Based Fault-Tolerant Control on an Interleaved DC/DC Boost Converter for Fuel Cell Electric Vehicle Applications," *International Journal of Hydrogen Energy* 40, no. 45 (2015): 15815–15822.

22. M. Kabalo, D. Paire, B. Blunier, D. Bouquain, M. G. Simões, and A. Miraoui, "Experimental Evaluation of Four-Phase Floating Interleaved Boost Converter Design and Control for Fuel Cell Applications," *IET Power Electronics* 6, no. 2 (2013): 215–226.

23. M. Kumar, "Detection and Localization of Open Switch Faults for Level-Shifted PWM Cascaded H-Bridge Inverter," *IEEE Transactions on Circuits and Systems II: Express Briefs* 71, no. 4 (2024): 2409–2413.

24. A. K. Gupta and M. Kumar, "Characterization and Localization of Open Circuit Faults for *n*-Phase Interleaved Buck Converter," *IEEE Transactions on Industry Applications* 60, no. 2 (2024): 3273–3283.

25. M. Kumar, "Open Circuit Fault Detection and Switch Identification for LS-PWM H-Bridge Inverter," *IEEE Transactions on Circuits and Systems II: Express Briefs* 68, no. 4 (2021): 1363–1367, https://doi.org/10.1109/TCSII.2020.3035241.

26. H. Zhu, D. Zhang, B. Zhang, and Z. Zhou, "A Nonisolated Three-Port DC–DC Converter and Three-Domain Control Method for PV-Battery Power Systems," *IEEE Transactions on Industrial Electronics* 62, no. 8 (2015): 4937–4947, https://doi.org/10.1109/TIE.2015.2393831.

27. Z. Zhou, H. Wu, X. Ma, and Y. Xing, "A Non-Isolated Three-Port Converter for Stand-Alone Renewable Power System," in IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society (Montreal, QC, Canada, 2012), 3352–3357, https://doi.org/10.1109/IECON.2012.6389360.

28. L. Kumar and S. Jain, "Multiple-Input DC/DC Converter Topology for Hybrid Energy System," *IET Power Electronics* 6, no. 8 (2013): 14831501.

29. M. R. Al-Soeidat, H. Aljarajreh, H. A. Khawaldeh, D.D.-C. Lu, and J. Zhu, "A Reconfigurable Three-Port DC–DC Converter for Integrated PV-Battery System," *IEEE Journal of Emerging and Selected Topics in Power Electronics* 8, no. 4 (2020): 3423–3433, https://doi.org/10.1109/JESTPE.2019.2941595.



1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4595 by Delhi Technological University, Wiley Online Library on [19/05/2025]. 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

# Intelligent Nanorobots for Precision Cancer Diagnosis and Treatment using Deep Reinforcement Learning

Kapil Sharma
*Dept. Information Technology*
*Delhi Technological university*
Delhi, India
kapil@ieee.org

Ravi Ranjan*
*Dept. of Information Technology*
*Delhi Technological University*
Delhi, India
raviranjan_23phdit502@dtu.ac.in

2025 11th International Conference on Mechatronics and Robotics Engineering (ICMRE) | 979-8-3315-0929-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICMRE64970.2025.10976292

***Abstract*—Early detection of cancer is essential as it significantly improves treatment outcomes and survival rates by addressing the disease before it spreads to other organs. Many cancers, such as breast, cervical, and colorectal, are highly treatable in their initial stages. With the development of Artificial intelligence, nano technology, the engineers, doctors, and biologist are working on nanorobots for diagnosis and treatment of cancer. It can help detect cancer at an early stage. Early diagnosis often reduces the complexity and cost of treatment, alleviating the physical and financial burdens on patients. Nanorobots are at the forefront of advanced biomedical applications, offering potential breakthroughs in targeted drug delivery and precision diagnostics, particularly for cancer treatment. By integrating artificial intelligence and Deep Reinforcement Learning, these devices can autonomously navigate complex biological environments to detect cancer cells via biomarker gradients. This work explores the deployment of Deep Q-Learning to improve nanorobot decision-making, enabling efficient path optimization and obstacle navigation.**

***Keywords—Nanorobots, Cancer Diagnosis, Drug Delivery, Deep-Q learning, Deep Reinforcement learning***

## I. INTRODUCTION

Cancer remains one of the leading causes of morbidity and mortality worldwide, with nearly 10 million deaths annually, as reported by the World Health Organization (WHO)[1]. Cancer is a dangerous disease, but in most of the cancers, the prominent reason for mortality is the late detection. Apart from that, the treatments of cancer like chemotherapy also impact the body negatively, as it moves in the body through blood, it impacts other body cells as well, whichever are growing rapidly in the body like white blood cells. Early detection of cancer can significantly increase the chances of survival. As the cancer progress it metastasis into other body parts, and then it becomes very challenging for doctors to cure the cancer. Advancements in artificial intelligence, sensors and robotics have brought emergent changes in the field of the medicine as well. With the development of nanotechnology, a lot of research is being conducted to use it for detection and treatment of cancer [2]. Nanorobots are explored for drug delivery for the treatment of complex medical diseases like Neurological disorders, as it can breach the blood-brain barrier (BBB) [3]. BBB filters the blood before sending it to the brain to protect it from harmful substances. Although BBB is necessity for the human bodies, as it protects us from a lot of potential diseases, but at the same time it sometimes do not let the medicine drug to pass this barrier. But since nanorobot can surpass the BB Barrier, it could be used to provide the drug to the specific location.

Treatments such as Cancer, have their own side-effects, the medicine also impacts the health of the patient severely, as rather than acting just on the cancer cell, the medicine travels through whole body and could impact any organ. Nanorobots could be used to deliver the drug to the specific parts of the body that need to be treated rather than the whole body as shown in figure 1.

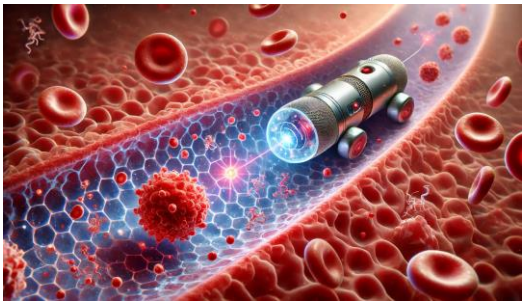


Fig. 1. Nanorobot transforming medicine

These miniature devices have biologically inspired designs that could offer precise disease detection and treatment. Biologist, Engineers and Doctors are working on integration of artificial intelligence, microfabrication and nano technology to find a way to track the inside body actions without interfering any bodily function [4]. These nanorobots can be used to navigate, track, targeting and delivery through BBB without causing any harm to the body. The advancements in this field are also explored by the researchers for treatment of more complicated diseases like Alzheimer's and Parkinson's disease, as the making the drug reach in the brain to target the issues in brain is a very difficult task due to the Blood-brain-barrier [5]. Drug delivery using these nanorobots could be the future of the medicine, but a lot of research still need to be done to have an effective system, before it could even be tested in clinical trials.

In this paper, we present a simulation for a nanorobot, that detects the cancer cells in the body and deliver a drug to them. The process is conducted by creating a simulating environment of the body by creating the cancer cells in a 3-D environment. As a human body is a very complicated mixture of different types of cells, tissues, nerves, organs etc.. As the blood travels through the whole body, it will pass besides cancer cells as well, so to detect cancer and delivering the drug, the nanorobots need to travel along with the blood [6]. These nanorobots are very


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:49:22 UTC from IEEE Xplore. Restrictions apply.

small like of size 0.1 to 10 micrometers. In the simulation we consider the nanorobot will travel to cancer cell eventually too, but it should know which cell is cancer and which are not. To understand which cell is a cancer cell, the biologist and doctors too depend on the biomarkers, some of these biomarkers can be traced through blood cells, some using the imaging technology such as CT scan and PET scan. The biomarkers that are identifiable through blood tests are CA125 for Ovarian cancer, Prostate-specific antigen for prostate cancer, CA19.9 for Pancreatic cancer. These biomarkers are present in blood, because cancer cells release these into the blood, so the concentration of these biomarkers are always more near the cancer cells. The simulation was created by displaying cancer cells, biomarkers, nanorobots on the 3-Dimension space. For rest of the blood cells, or organ are being shown with the help of obstacles in the simulation. In this simulation the nanorobot avoids the obstacles and move towards the cancer cell by identifying where the concentration of the biomarkers is the most. Deep Q-Learning, a reinforcement learning technique enhanced by neural networks, enables nanorobots to adapt dynamically to environmental complexities [7]. Unlike traditional Q-Learning, which uses tabular methods, Deep Q-Learning approximates Q-values through deep neural networks, making it ideal for high-dimensional decision spaces encountered in biological systems.

Researchers are exploring the use of reinforcement learning for optimizing the nanorobot navigation through the body. They aim to minimize the time taken to reach the cancer cell, by reducing the distance covered. Kavousinejad used Q-learning to refine nanorobot movement strategies using biomarker concentration data [8,11]. Cao et al designed an Improved bacterial foraging optimization algorithm along with cooperative learning for multi-nanorobots cancer detection [9]. Liu et al. developed a computational nano biosensing framework using reinforcement learning for tumor detection and treatment [10]. In this work, we have used deep reinforcement learning technique Deep-Q to refine the movement strategies of nanorobots and to optimize the time taken to identify the cancer cell.

## II. Methodology

In this work we explore the deployment of Deep Q-Learning to improve nanorobot decision-making, enabling efficient path optimization and obstacle navigation. We use Deep Q learning for nanorobot movement towards higher biomarker concentration. To find the path through to reach highest biomarker concentration we use gradient analysis.

### A. Biomarker Detection and Gradient Analysis

Cancer biomarkers, such as circulating tumor DNA and proteins, diffuse from tumor cells and create gradients in the surrounding tissue. These gradients serve as navigational cues for nanorobots [12]. Using a Gaussian distribution model, the biomarker concentration at any point $(x, y, z)$ is calculated as:

$$\rho(x,y,z) = \rho_0 e^{\frac{-((x-x_0)^2+(y-y_0)^2+(z-z_0)^2)}{2\sigma^2}}$$

Where, $\rho_0$: Peak concentration at the cancer cell.

$(x_0, y_0, z_0)$ : Cancer cell coordinates. $\sigma$ : Spread of the biomarker concentration. Nanorobots measure biomarker density gradients to move toward cancer cells.

### B. Deep Q-Learning Framework

To enable autonomous navigation and precise targeting of cancer cells, we employed a Deep Q-Network (DQN) architecture. DQN is a reinforcement learning algorithm that learns optimal policies through interaction with an environment [13,14]. In our case, the environment is a simulated blood vessel network, replete with obstacles and varying biomarker concentrations.

*State Representation:* The state of the nanorobot at any given time is defined by a vector:

$$S = [B_1, B_2, ..., B_n, D_1, D_2, ..., D_m, X, Y, Z, \theta]$$

Where $B_i$ is the concentration of biomarker i, $D_i$ is the distance to obstacle i. X, Y, Z represents the position of the nanorobot, and θ is the orientation of the nanorobot

*Action Space:* The action space, A, consists of discrete actions that the nanorobot can take:

A = {forward, backward, left, right, rotate_left, rotate_right}

*Reward Function:* The reward function, R(S, A), is designed to incentivize desired behaviours:

R(S, A) =
+5, if the nanorobot reaches a target cell
-1, if the nanorobot collides with an obstacle
-2, if the nanorobot exceeds a time limit
+1,if the nanorobot moves closer to a target cell

*DQN Training:* The DQN learns to approximate the optimal Q-value function, Q(S,A), which represents the expected cumulative reward for taking action A in state S. The Q-network is trained to minimize the temporal difference error:

$$TD_error = R(S, A) + \gamma \times \max_{a'} Q(S', a') - Q(S, A)$$

Where γ is the discount factor, S' is the next state, a' is the action taken in the next state.

To stabilize training and improve convergence, we employed experience replay. A replay buffer stores a history of (state, action, reward, next state) tuples. During training, a random sample of experiences is drawn from the buffer to update the Q-network.

*Nanorobot Control:* Once the DQN is trained, it can be used to control the nanorobot in real-time. At each timestep, the current state S is fed into the DQN, which outputs the Q-values for all possible actions. The action with the highest Q-value is selected and executed by the nanorobot's actuators.

Simulation is conducted by taking 100 obstacles 40 biomarkers that were placed near 2 of the cancer cells, and one nanorobot. The whole objective is for Nanorobots to demonstrate effective navigation toward cancer cells while avoiding obstacles. This objective is achieved through


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:49:22 UTC from IEEE Xplore. Restrictions apply.

reinforcement learning, by maximizing the cumulative reward award. Whenever the nanorobot move towards the cancer cell, it gets a reward. Whether the nanorobot is moving towards the cancer cell is determined by the concentration of the biomarkers. If nanorobot start to move away or struck into an obstacle, the framework gets a penalty. By moving towards higher concentration of biomarkers by following the gradient, nanorobot tries to reach the cancer cell. The simulation environment was created on google colab by taking these different components position randomly on a 3-D graph. We tested the simulation multiplied times to validate that deep Q learning algorithm works very well to find the cancer cells in the body. Although the model should be validated very effectively before it could be used for medical purposes. Working in simulation environment do not guarantee that it would be effective in real-time environment as well.

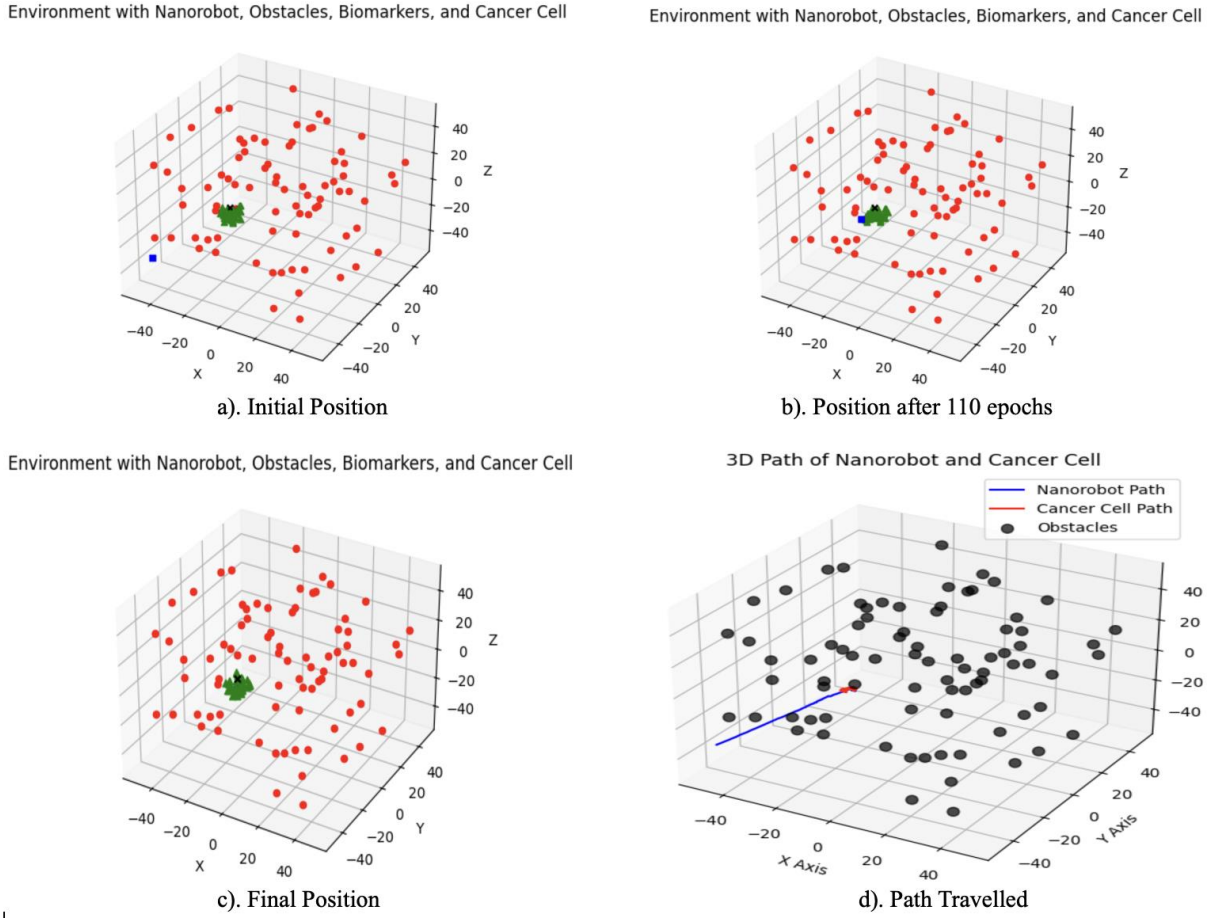



Fig. 2. Nanorobot movement in the Simulation Environment

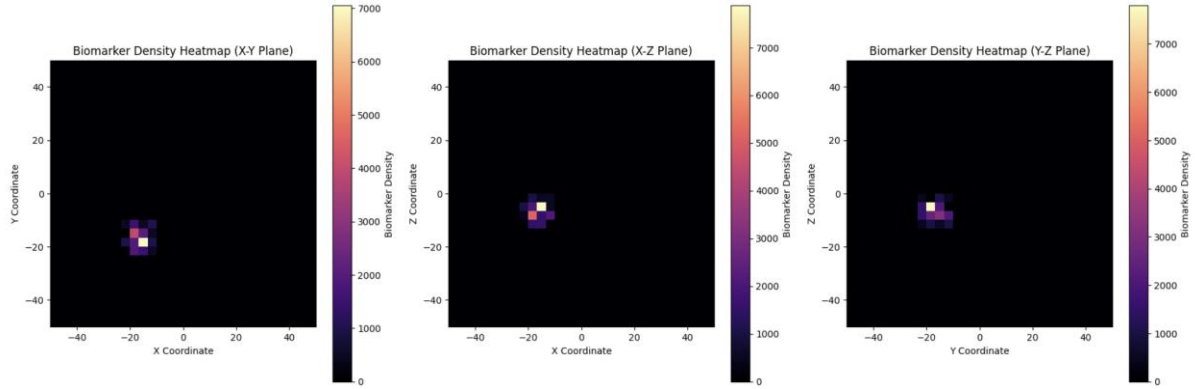



Fig. 3. Heatmap of Biomarker Concentration


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:49:22 UTC from IEEE Xplore. Restrictions apply.

## III. Result & Discussion

To evaluate the model, the efficiency and speed of the model is analysed, how accurately the nanorobot detect the cancer cell and in how much time, the movement of the nanorobots to cancer cell is shown in figure 2. We have calculated the speed in terms of number of epochs it took to reach the cancer cell, and the accuracy is being measured by the distance hat remains between the nanorobot and the cancer cell. The use of biomarker gradients significantly reduced travel distance and time. Figure 3 shows the biomarker of concentration on 3-D plan using heatmap representation. We have set the target threshold value of distance at 0.05. The maximum epochs set was at 350, with discount factor of 5%, and learning rate of 2%. Nanorobot reached the desired distance at 339 epoch, but as can be seen from figure 4, the nanorobot reached near the cancer cell at near 300 epochs. The result of the simulated environment is:

*Total Steps: 339*
*Final Distance: 0.04*
*Average Distance: 12.22*
*Average Biomarker Concentration: 0.85*
*Simulation Duration: 7.22 seconds*

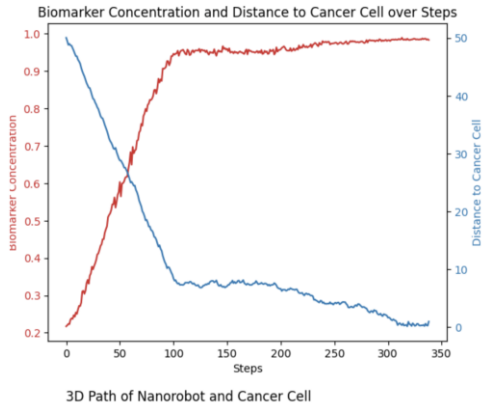



Fig. 4. Converging of distance between nanorobot and cancer cell

As could be observed from the results nanorobots learned to avoid obstacles dynamically and prioritized paths with higher biomarker densities. The simulation was performed multiple times, to ensure the efficiency of the model. Although the convergence time varied in each simulation, except two occurrences the model converged before the threshold of 350 epochs. Deep Q-Learning enabled faster convergence compared to traditional Q-Learning. While DQN offers a promising approach for nanorobot control, several challenges remain like, Uncertainty in the environment: Blood flow dynamics, obstacle distributions, and biomarker concentrations can vary significantly. Real-world sensors may be subject to noise and inaccuracies. Also, Real-time decision-making requires efficient inference and control algorithms. Researchers could address these challenges developing algorithms that can handle uncertainty and noise. It could also be resolved by decomposing the complex task of navigation and targeting into simpler subtasks. Transfer learning could also be used to leverage the knowledge gained from simulations to accelerate real-world learning. By overcoming these challenges this could possibly be used in medical field to precisely diagnose and treat the complex diseases, especially cancer.

## IV. Conclusion

Healthcare has gone through multiple generations as well, but even with all these advancements in the medical filed there are still the scenarios when either it is difficult to detect the disease until it's too late, or after diagnosis, the treatment is difficult as well. Some of these problems can be resolved by integrating the engineering knowledge with the medicine. Nanotechnology has helped to develop such small devices that can be traversed through the smallest places like human veins. The nanorobots are being designed to target specific cells of the body such as cancer to diagnose it and to deliver the drug to those particular cells. It is an ongoing research field, having a huge scope for real-time applications with an effective system. For designing an effective model we have used the deep reinforcement learning technique Deep-Q for training the nanorobots to learn the behavior of cancer cells, i.e. by tracking the concentration of biomarkers in the blood cells and near specific organs. Our model has worked very well for the simulated environment, but it was only tested for one cancer cells. Multi nanorobots should be able to detect the different cancer cells at the same time. Swarm techniques can be used by the researchers for designing a more effective multi-robot simulation. Deep-Q approach could also be incorporated with other models to design the nanorobots more efficiently.

## References


[1]. Kim, Y. B., Lee, I. J., Byun, H. K., Choi, Y. Y., Hong, B., & Lee, J. (2024). Symptom network and quality of life of breast cancer patients receiving multimodal cancer treatment: Cross-sectional study. *European Journal of Oncology Nursing*, *71*, 102661.

[2]. LoPiccolo, J., Gusev, A., Christiani, D. C., & Jänne, P. A. (2024). Lung cancer in patients who have never smoked—An emerging disease. *Nature Reviews Clinical Oncology*, *21*(2), 121-146.

[3]. Cummings, J., Osse, A. M. L., Cammann, D., Powell, J., & Chen, J. (2024). Anti-amyloid monoclonal antibodies for the treatment of Alzheimer's disease. *BioDrugs*, *38*(1), 5-22.

[4]. Wang, J., Dong, Y., Ma, P., Wang, Y., Zhang, F., Cai, B., ... & Liu, B. F. (2022). Intelligent Micro-/Nanorobots for Cancer Theragnostic. *Advanced Materials*, *34*(52), 2201051.

[5]. Dolev, Shlomi, Ram Prasadh Narayanan, and Michael Rosenblit. "Design of nanorobots for exposing cancer cells." *Nanotechnology* 30, no. 31 (2019): 315501.

[6]. Shinde, A., Prasad, S. B., Srinivasarao, D. A., Shah, S., Famta, P., Khairnar, P., ... & Srivastava, S. (2024). Nano voyagers: Pioneering a new frontier in cancer treatment with nanorobots as drug transporters. *Applied Materials Today*, *38*, 102162.

[7]. Van Hasselt, H., Guez, A., & Silver, D. (2016, March). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 30, No. 1).

[8]. Kavousinejad, S. (2024). Simulation of Nanorobots with Artificial Intelligence and Reinforcement Learning for Advanced Cancer Cell Detection and Tracking. *arXiv preprint arXiv:2411.02345*.

[9]. Cao J, Li M, Wang H, Huang L, Zhao Y, editors. An improved bacterial foraging algorithm with cooperative learning for eradicating cancer cells using nanorobots. 2016 IEEE International Conference on Robotics and Biomimetics (ROBIO); 2016: IEEE.

[10]. Liu L, Sun Y, Shi S, Chen Y, editors. Smart tumour targeting by reinforcement learning. 2021 IEEE 15th International Conference on Nano/Molecular Medicine & Engineering (NANOMED); 2021: IEEE.

[11]. Kong, X., Gao, P., Wang, J., Fang, Y., & Hwang, K. C. (2023). Advances of medical nanorobots for future cancer treatments. *Journal of Hematology & Oncology*, *16*(1), 74.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:49:22 UTC from IEEE Xplore. Restrictions apply.

[12]. Hochreiter, S., Younger, A. S., & Conwell, P. R. (2001). Learning to learn using gradient descent. In *Artificial Neural Networks—ICANN 2001: International Conference Vienna, Austria, August 21–25, 2001 Proceedings 11* (pp. 87-94). Springer Berlin Heidelberg.

[13]. Mateusz Święcki, Józef Szymelewicz, Jerzy Matusiewicz, and Rafał Grądzki, "Tests of Useful Field of View of Laser Sensors Used in Autonomous Nano Sumo Robots," International Journal of Mechanical Engineering and Robotics Research, Vol. 10, No. 8, pp.414-421, August 2021. DOI: 10.18178/ijmerr.10.8.414-421

[14]. Tan, F., Yan, P., & Guan, X. (2017). Deep reinforcement learning: From Q-learning to deep Q-learning. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24* (pp. 475-483). Springer International Publishing.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:49:22 UTC from IEEE Xplore. Restrictions apply.

Iranian Journal of Science and Technology, Transactions of Civil Engineering
https://doi.org/10.1007/s40996-025-01881-8

RESEARCH PAPER

# Investigation of Pollution by Toxic Metals in the Bed Sediments of Ganga River Along Kanpur to Prayagraj City of Uttar Pradesh, India

**Mohit Aggarwal[1,2] · Srinivasan Anbukumar[1] · Thangamani Vijaya Kumar[1]**

Received: 19 April 2023 / Accepted: 23 April 2025
© The Author(s), under exclusive licence to Shiraz University 2025

**Abstract**
We examined the fluctuations in the concentrations of Cd, Cr, Cu, Fe, Mn, Ni, Pb and Zn in the bed sediment of the Ganga River at ten sites, along 210 km span between Kanpur and Prayagraj. We utilized an Atomic Absorption Spectrometer (AAS) for metal analysis. The Ganga, India's largest river, poses a serious threat to both people and the environment. Sediment samples were collected in the summer of 2021 after the relaxation in Covid-19 lockdown. Metal concentration averages are noticed in the following order: Fe > Mn > Cr > Zn > Ni > Pb > Cu > Cd. High heavy metal concentrations are measured at mid locations i.e., Dalmau Bridge, Unchahar, Nawabganj and Manikpur as compared to other locations. Pollution indicators used in this study indicate moderate contamination by Pb and Cd metal. Relative mobility of metals calculated in the following decreasing order: Mn > Cr > Pb > Zn > Ni > Cu > Cd. A comparative study of Ganga River sediment with the previous data of 2019 indicates the reduction in concentrations of all metals. This can be attributed due to the complete shutdown in the country during 2020 and 2021 because of COVID-19. This also indicates the self-healing capacity of the Ganga River. Further, to prevent the river from further degradation industrial waste containing metals should be treated before disposal to protect the river from further degradation.

**Keywords** Heavy metals · Ganga river · Sediments · Enrichment factor · Pollution load index · Relative mobility of metals

## 1 Introduction

The Ganga, India's national river is one of the most heavily inhabited large alluvial tracts. It is an active foreland basin formed by the colliding of the Asian and Indian plates. The Himalayas, which were created by the cut and stacked up sedimentary cover north of the basin, are a major source of sediment for the basin's waterways. Around $5.21 \times 10^8$ metric tons per year of sediment are transported by the river Ganga (Milliman and Meade 1983). Along its 2600 km length, the river is home to approximately one-third of India's total population in 700 cities and towns (Dasgupta 1984). Ganga River drainage basin has been used for the disposal of agricultural runoff, domestic waste, and industrial effluents (Buaisha et al. 2020) that has harmed the quality of sediments, water, and soil within its path. In addition to the existence of organic waste and pathogens, the sediments and water of the Ganga River have a significant amount of heavy metal contamination which is causing toxicity, especially in the grasslands (Ansari et al. 1998; Haritash et al. 2014). According to a few investigations (Kansal et al. 2013), the Ganga River tributaries also contain toxic elements resulting from industrial effluent and flyash (Subramanian et al. 1987a, b; Chander et al. 1994) in the populated river stretches of Delhi, Varanasi, Agra, and Patna. River pollution can be determined by evaluating its water quality or its bed sediment quality. Though the quality of water varies within the different seasons, the quality of sediments is least influenced by changes in seasons and reflects the river's quality for a longer period of duration (Jain et al. 2005). While various studies in the last few years have resulted in useful information on the quality of water, the quality of sediment is still overlooked. (Jain et al. 2005). Unlike water, which immediately introduces aquatic

✉ Mohit Aggarwal
mohitjvg@gmail.com

[1] Department of Civil Engineering, Delhi Technological University, Delhi, India

[2] School of Engineering & Technology, Noida International University, Gautam Budh Nagar, Uttar Pradesh, India

Published online: 03 May 2025

organisms to heavy metal toxicity, sediments may control the quality of water over a longer period of duration, despite contamination, because sediments act as sinks for metals, and are the careers to transfer metals to faraway locations. In addition to this, sediments serve as both a sink and a source, governing the availability, toxicity, and mobility of metals in a river's environment. As a result, river sediments are regarded to be a genuine representation of the stage of river pollution. There are times when the sediments are polluted, but the river water is not, which presents a risk, particularly when toxic elements are mobilized into the sediments to the stream of water. Therefore, research on the quality of river water would be insufficient without the investigation of sediments. This polluted river water is utilized for agricultural purposes and for drinking in the downstream areas, resulting in the toxification of the food chain through these metal-containing clay and fine silt. Because of the high energy requirement for treating sewage (Mirra et al. 2020) and the limitation of resources and finances (Khan et al. 2022), wastewater treatment in developing countries is frequently inefficient, leading to the deterioration of the quality of water in lakes and rivers of ecological significance (Chitrakshi and Haritash 2022). Aquatic systems can be polluted by heavy metals in wastewater from industries, wastes from hospitals and chemical factories, runoff from highways and farming, sewage, and leachates from mines which discover their passage into the river systems (Karaca et al. 2017). In addition, the formation of rocks due to natural weathering within the river basin contributes to the introduction of trace quantities of heavy metals. Bio-accumulation of heavy metals happens naturally, which indicates that once they enter the biotic system, they often remain in the body, and over time their concentration rises (Satarug et al. 2010). Concerning heavy metals (As, Cd, Cr, Cu, Hg, Mn, Ni, Pb, Cu, and Zn) show a propensity to become adsorbed on the sediment surface, particularly during the low flow. These metals are frequently moved from the sediments into the water, based on their pH value, digenesis, flooding, turbulence, and redox potential (Osakwe et al. 2014). These toxic heavy metal concentrations must be controlled because they can reach the human body via food and water. The intake of water directly can allow these toxic metals into our bodies. Heavy metal contamination in food might lead to biomagnification of toxicity as the food chain advances through the trophic levels, leading to a multiplication of toxicity as well as its influence on foremost users. As a result, sediments have evolved as an important indicator of the quality of water, and their investigation is critical for understanding the impacts on riverine ecosystems. Many heavy metal investigations (Singh and Pandey 2022; Jaiswal and Pandey 2018; Siddiqui et al. 2019; Aggarwal et al. 2022, 2023, 2024, 2025) have been released for bed sediment of Ganga River which have highlighted the higher concentrations of some toxic metals particularly in the stretch between Kanpur and Prayagraj. The current research seeks to identify the variability of metal concentrations in the Ganga River sediments collected from the 210 km middle section of the Ganga River within Kanpur and Prayagraj in Uttar Pradesh, considering all of the aforementioned factors. The quality of sediment is assessed using the pollution indicators namely Enrichment Factor, Contamination Factor, Geoaccumulation Index, and Pollution Load Index to classify polluted locations and sources of contamination (Aggarwal et al. 2022, 2023, 2024, 2025).

## 2 Materials and methods

### 2.1 Selection of Site

Ten (10) sampling locations – Jajmau Bridge (JMU), Maharajpur (MJR), Sarsaul (SRL), Dalmau-Fatehpur Ganga Bridge (DFB), Village Unchahar (UHR), Village Nawabganj (NWJ), Manikpur Ghaat (MNR), Shringverpur Ghaat (SVR), Phaphamau Ganga Bridge(PMU) and Sangam (SGM) were chosen along 210 km stretch of the river (Fig. 1) to determine the metal concentrations in Ganga's sediments. The sites were selected based on their accessibility and were roughly equidistant from one another. All sites are populated because these stretches have lately been impacted by tourists, industrial growth, agricultural activities, immigration, and the gradual extension of neighboring towns.

### 2.2 Sample Collection

During the summer of 2021, sediment samples were grabbed from the top 10–15 cm layer of the river bed at the banks of each site. Using the standard procedure, a total of Three (3) grab samples of sediments were collected from each location and the samples were then placed in cleaned sealed polythene packaging for subsequent study. Before collecting the next sample, the sampling instruments were washed and dried.

### 2.3 Analysis

The moisture in samples was removed with the help of a preheated oven set at 60 degrees Celsius. The dried sediments were then crushed and ground to obtain fractions of 63 µm size. 1 g of samples were acid digested using 10 mL of a 1:2 composition of $HClO_4$ and $HNO_3$ and heated for 1 h at low heat, followed by 2-hour heating at a high temperature of 120 °C, till the fumes began to emerge and clear

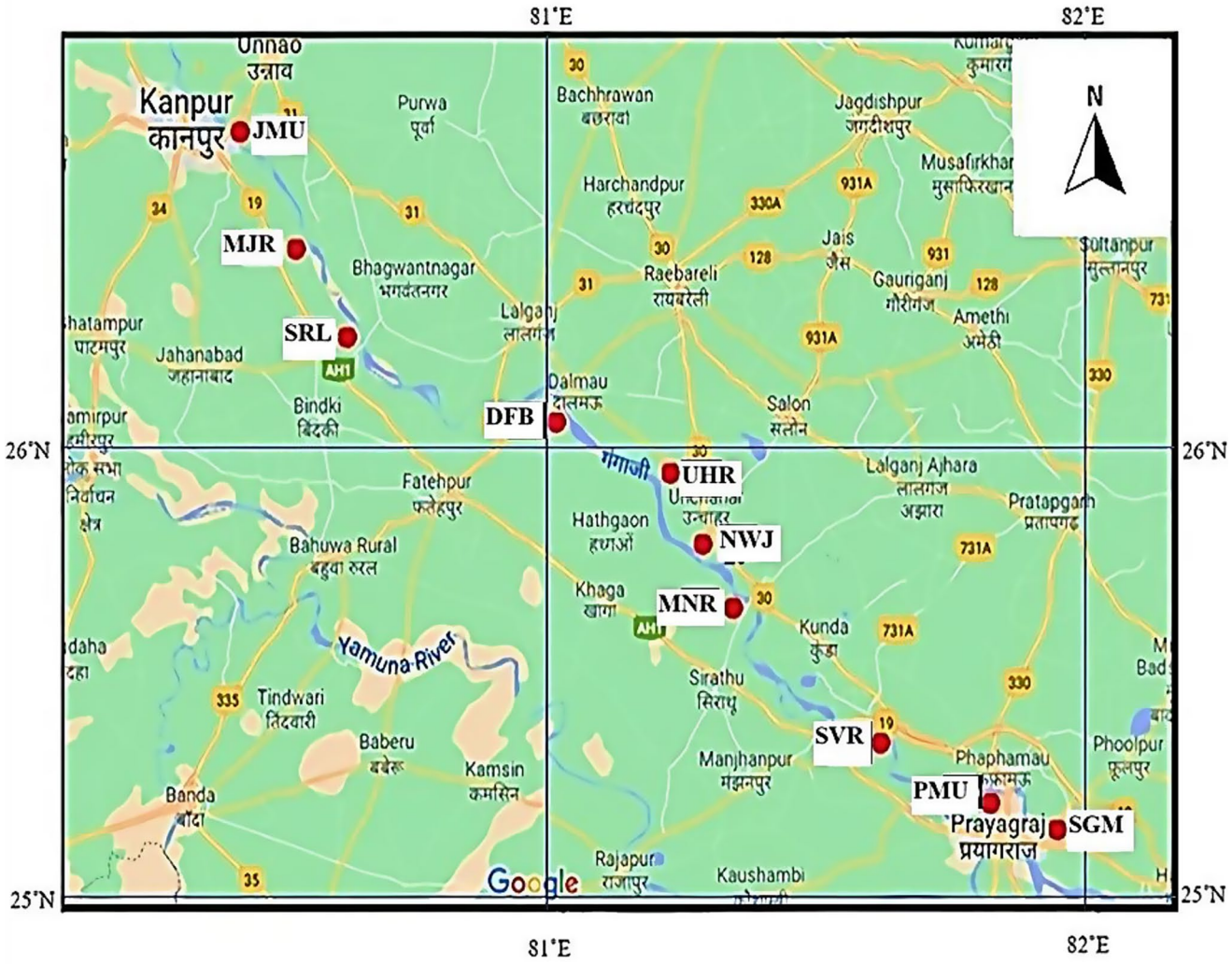



**Fig. 1** Study area map and sampling locations within Kanpur and Prayagraj, India

and transparent solution is obtained. After that, digested samples are filtered through Whatman filter paper 42. The filtrate is then diluted with 0.1 N $HNO_3$ until it reaches 50 mL of the final volume. Lastly, the concentrations of Cd, Cr, Cu, Fe, Mn, Ni, Pb, and Zn metals in digested samples are determined with the help of atomic absorption spectroscopy (AAS) (Shimadzu AA-6300). MS Excel program was used to calculate the mean concentration. The Metal Enrichment Factor is used to investigate the anthropogenic effect on sediments. The formula is as follows:

$$Enrichment\ Factor\ = \frac{\left[\left(Mean\ concentration\ of\ x\ metal\right) \div \left(Mean\ concentration\ of\ \ referred\ metal\right)\right]_{\ sample}}{\left[\left(Baseline\ concentration\ of\ x\ metal\right) \div \left(Baseline\ line\ concentration\ of\ referred\ metal\right)\right]_{\ background}} \quad (1)$$

Three other indices were computed based on the metal concentrations to conclude the level of contamination of the sediments and to identify the source of metal pollution. Using the following formula, the Geo-accumulation Index (GI) is computed:

$$\mathrm{GI}\ =\ \mathrm{Log}_2\ \frac{\mathrm{Cn}}{1.5*\mathrm{Bn}} \quad (2)$$

where Cn is the observed concentration of metal in the current study and Bn is its concentration recorded in the past. GI classification is categorized into 7 classes, ranging from 0 to 6 (Muller 1969) as mentioned in Table 1.

The following formula is used to compute Pollution Load Index (PLI):

$$PLI\ =\ (CF1 \times\ CF2 \times CF3 \ldots . \times CFn)^{1/n} \quad (3)$$

**Table 1** GI class and pollution levels corresponding to GI values

| GI value | GI Class | Level of Pollution |
|---|---|---|
| Less than 0 | 0 | Uncontaminated |
| Between 0 and 1 | 1 | Uncontaminated to moderate contamination |
| Between 1 and 2 | 2 | Moderate contamination |
| Between 2 and 3 | 3 | Moderate to strong contamination |
| Between 3 and 4 | 4 | Strong contamination |
| Between 4 and 5 | 5 | Strong to Extreme contamination |
| Greater than 5 | 6 | Extreme contamination |

Here, CF is the contaminator factor for 'n' metal and is the ratio of that metal's present concentration (Cn) and its background concentration (Bn).

## 3 Results and Discussion

### 3.1 The Concentration of Heavy Metals in Sediments

Figure 2 depicts the longitudinal patterns of studied metals found in the collected sediment of the Ganga River. Fe metal concentration varied from 12,088 ppm (Sarsual) to 24,014 ppm (Nawabganj). Overall, it was noticed that the concentration of Fe increased initially and further reduced as we proceeded toward downstream locations. Concentrations of Cr varied from 26.69 ppm (Jajmau) to 85.17 (Maharajpur). Sangam location recorded the minimum concentration for Mn, Cu, and Zn (190.2, 2.47, and 14.30 ppm, respectively), whereas the highest values were observed as 373.5 ppm at Nawabganj, 11.12 ppm at Nawabganj and 30.21 ppm at Manikpur, respectively. The concentration of Cd varied between 1.18 and 2.85 ppm at Phaphamau and Nawabganj, respectively. The lowest concentrations of Ni and Pb were observed at Sarsaul with a value of 18.35 ppm and 25.37 ppm respectively while the highest was observed at Manikpur with a value of 24.28 ppm and 34.06 ppm respectively. Based on the results, it was determined that the concentration of all metals increased first as one moved downstream and then reduced till reaching the last locations. Several anthropogenic activities like effluents from industries, increase in tourism, sewage from the domestic area, runoff from agricultural areas, etc. are affecting river sediments, especially at middle sites i.e., Nawabganj and Manikpur. Purushothaman and Chakrapani (2007) found similar discoveries, revealing increased levels of metal pollution in lower parts of the Ganga due to anthropogenic activity.

### 3.2 Enrichment Factor

Determining the Metal Enrichment Factor is among the most prominent used approaches for assessing the anthropogenic impact on sediments. Because Earth's crust has an abundance of metal Fe, it could be presumed that its concentration is not so altered by human activity and is commonly used as a metal for reference (Haritash and Kaushik 2007). The main challenge in quantifying the enrichment of heavy metals in sediments is determining the baseline levels of metals.). The average shale concentration should be taken as baseline concentrations as per Turekian and Wedepohl (1961. Singh et al. (2003) recently discovered the baseline concentration of various metals in sediments of the Ganga. The geometric mean values supplied by Singh et al. (2003) were utilized as the baseline concentration for studied metals in this study. Cd, Cr, Cu, Fe, Mn, Ni, Pb, and Zn have baseline concentrations of 0.58, 147, 55, 40,346, 1764, 47, 22, and 105 ppm, respectively. In this investigation, reference metal is considered as Fe. Table 2 displays the computed values for metal enrichment factors.

The Enrichment factor calculated above 1 indicates an increase in metal concentrations due to anthropogenic sources (Haritash and Kaushik 2007). Applying this criterion, it is possible to state that the concentration of Mn, Zn, and Cu metals in the sediments of all locations, and to some extent of Cr and Ni metals, are within the limits ($<1.5$). However, Pb and Cd enrichment exists in all locations' sediments. The enrichment of metals can be linked to the fact that anthropogenic activity in selected locations is far greater as compared to the upper stretches of the river. As a result, wastewater from domestic areas, effluents from industries, runoff from agricultural regions, and other pollutants damage the selected sites (Kansal et al. 2013). The existence of many tourist attractions around these locations is also putting strain on Ganga's water quality.

### 3.3 Geo-Accumulation Index

Muller (1969) proposed the Geoaccumulation Index (GI) as a quantitative indicator of heavy metal pollution of sediments if the concentration of hazardous metal is equal to 50% or greater than its lithogenic background values. The GI for studied heavy metals is depicted in Table 3.

The sediments of all selected locations are found to be uncontaminated by Cr, Fe, Mn, Cu, Ni, and Zn; sediments of Manikpur Ghaat are uncontaminated or have moderate contamination of Pb and moderate contamination of Cd; sediments of Jajmau Bridge, Shringverpur Ghaat, Phaphmau Ganga Bridge and Triveni Sangam are indicating the pollution level of uncontaminated to moderate contamination of Cd, and sediments of Maharajpur, Sarsaul, Dalmau-Fatehpur Bridge, Unchahar and Nawabganj are moderately contaminated with Cd. The observed values for GI correspond to the anthropogenic sources of identified metals detected for each site.

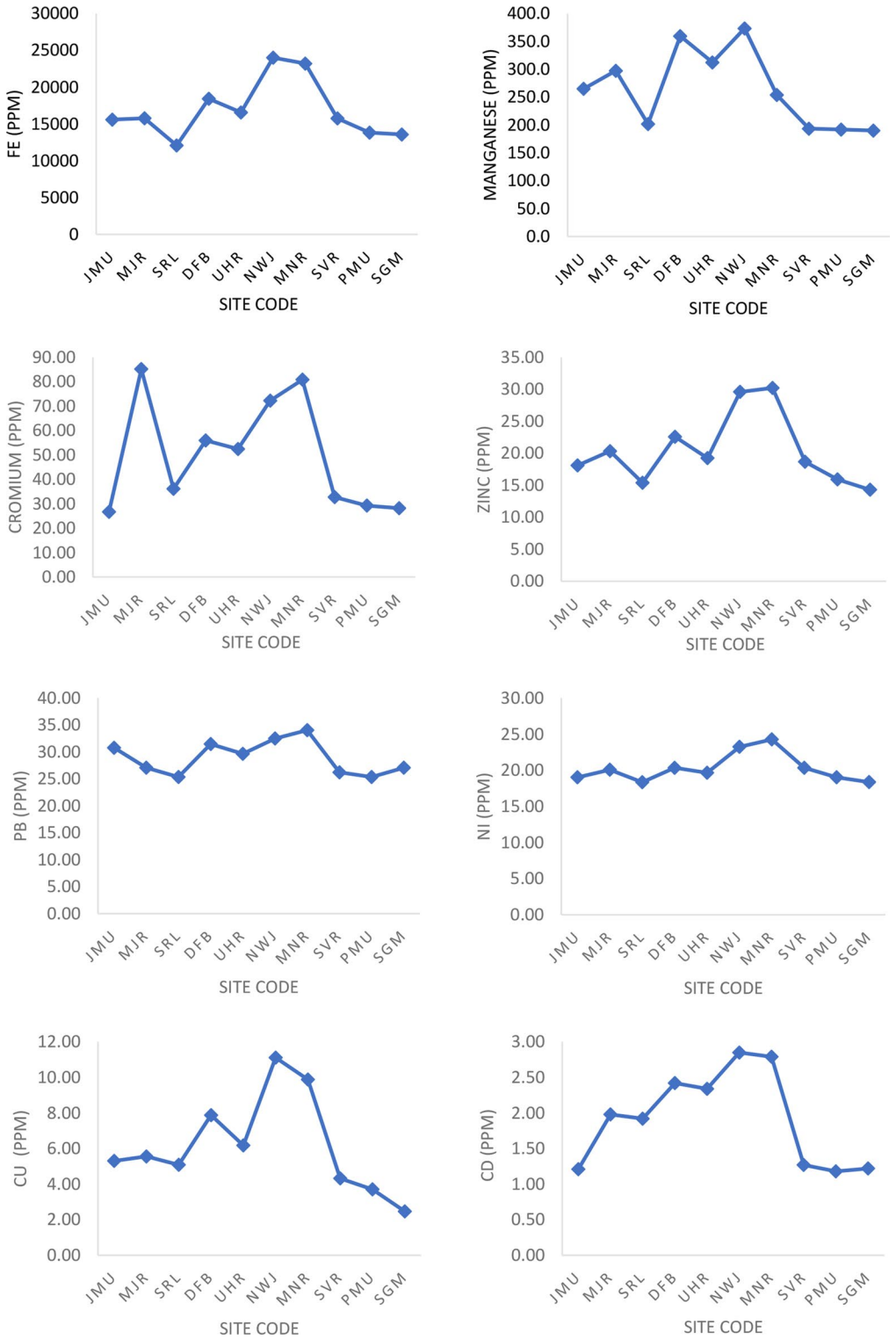



**Fig. 2** Variations in concentrations of heavy metals in river sediments along the selected stretch of Ganga, India

### 3.4 Pollution Load Index

Another pollution indicator was developed by Tomlinson et al. (1980) named as Pollution Load Index (PLI) and is used to assess the sediment quality. PLI values greater than one show an increase in the degradation of the quality of sediment (Tang et al. 2014). Table 4 displays the PLI value for several sample locations.

**Table 2** Enrichment factors of sediment sampled from selected locations of river Ganga

| Location | Code | Mn | Cr | Zn | Pb | Ni | Cu | Cd |
|---|---|---|---|---|---|---|---|---|
| Jajmau Bridge | JJU | 0.39 | 0.47 | 0.45 | 3.62 | 1.05 | 0.25 | 5.39 |
| Maharajpur | MJR | 0.43 | 1.48 | 0.50 | 3.15 | 1.09 | 0.26 | 8.73 |
| Sarsaul | SRL | 0.38 | 0.82 | 0.49 | 3.85 | 1.30 | 0.31 | 11.05 |
| Dalmau-Fatehpur Ganga Bridge | DFB | 0.45 | 0.83 | 0.47 | 3.13 | 0.95 | 0.31 | 9.13 |
| Village Unchahar | UHR | 0.43 | 0.87 | 0.45 | 3.28 | 1.02 | 0.27 | 9.82 |
| Village Nawabganj | NWJ | 0.36 | 0.83 | 0.47 | 2.48 | 0.83 | 0.34 | 8.26 |
| Manikpur Ghaat | MNR | 0.25 | 0.96 | 0.50 | 2.69 | 0.90 | 0.31 | 8.36 |
| Shringverpur Ghaat | SVR | 0.28 | 0.57 | 0.46 | 3.05 | 1.11 | 0.20 | 5.61 |
| Phaphamau Ganga Bridge | PMU | 0.32 | 0.58 | 0.44 | 3.36 | 1.18 | 0.20 | 5.94 |
| Triveni Sangam | SGM | 0.32 | 0.57 | 0.40 | 3.66 | 1.16 | 0.13 | 6.25 |

**Table 3** The Geo-accumulation index (GI) of studied metals in sediments samples of river Ganga

| Location | Code | Fe | Mn | Cr | Zn | Pb | Ni | Cu | Cd |
|---|---|---|---|---|---|---|---|---|---|
| Jajmau Bridge | JJU | -1.96 | -3.32 | -3.05 | -3.12 | -0.10 | -1.89 | -3.96 | 0.48 |
| Maharajpur | MJR | -1.94 | -3.15 | -1.37 | -2.95 | -0.28 | -1.81 | -3.89 | 1.19 |
| Sarsaul | SRL | -2.32 | -3.71 | -2.61 | -3.36 | -0.38 | -1.94 | -4.02 | 1.14 |
| Dalmau-Fatehpur Ganga Bridge | DFB | -1.72 | -2.88 | -1.98 | -2.80 | -0.07 | -1.79 | -3.39 | 1.48 |
| Village Unchahar | UHR | -1.87 | -3.08 | -2.07 | -3.03 | -0.15 | -1.84 | -3.74 | 1.43 |
| Village Nawabganj | NWJ | -1.33 | -2.82 | -1.61 | -2.41 | -0.02 | -1.60 | -2.89 | 1.71 |
| Manikpur Ghaat | MNR | -1.38 | -3.38 | -1.45 | -2.38 | 0.05 | -1.54 | -3.06 | 1.68 |
| Shringverpur Ghaat | SVR | -1.94 | -3.77 | -2.75 | -3.08 | -0.33 | -1.79 | -4.25 | 0.55 |
| Phaphamau Ganga Bridge | PMU | -2.13 | -3.79 | -2.91 | -3.31 | -0.38 | -1.89 | -4.48 | 0.44 |
| Triveni Sangam | SGM | -2.16 | -3.80 | -2.97 | -3.46 | -0.28 | -1.94 | -5.06 | 0.49 |

**Table 4** Values of contamination factor and pollution load index of collected sediment samples of Ganga river

| Location Code | Contamination Factor | | | | | | | | Pollution Load Index (PLI) 2021 | Pollution Load Index (PLI) 2019* |
|---|---|---|---|---|---|---|---|---|---|---|
| | Fe | Mn | Cr | Zn | Pb | Ni | Cu | Cd | | |
| JJU | 0.387 | 0.150 | 0.182 | 0.172 | 1.399 | 0.405 | 0.096 | 2.086 | 0.346 | 0.018 |
| MJR | 0.391 | 0.169 | 0.579 | 0.194 | 1.231 | 0.427 | 0.101 | 3.414 | 0.438 | 0.187 |
| SRL | 0.300 | 0.114 | 0.246 | 0.147 | 1.153 | 0.391 | 0.093 | 3.310 | 0.338 | 0.016 |
| DFB | 0.457 | 0.204 | 0.380 | 0.215 | 1.431 | 0.433 | 0.143 | 4.172 | 0.480 | 0.143 |
| UHR | 0.411 | 0.177 | 0.357 | 0.183 | 1.349 | 0.418 | 0.112 | 4.034 | 0.432 | 0.154 |
| NWJ | 0.595 | 0.212 | 0.491 | 0.282 | 1.478 | 0.495 | 0.202 | 4.914 | 0.579 | 0.660 |
| MNR | 0.575 | 0.144 | 0.550 | 0.288 | 1.548 | 0.517 | 0.180 | 4.810 | 0.555 | 0.964 |
| SVR | 0.391 | 0.110 | 0.223 | 0.178 | 1.192 | 0.433 | 0.079 | 2.190 | 0.333 | 0.035 |
| PMU | 0.343 | 0.109 | 0.199 | 0.152 | 1.153 | 0.405 | 0.067 | 2.034 | 0.303 | 0.007 |
| SGM | 0.337 | 0.108 | 0.192 | 0.136 | 1.231 | 0.391 | 0.045 | 2.103 | 0.285 | 0.012 |

*Values taken from Aggarwal et al. (2022)

Based on PLI values, the locations indicate the following increasing order: Sangam < Phaphamau < Shringverpur < Sarsaul < Jajmau < Unchahar < Maharajpur < Dalmau Fatehpur Bridge < Manikpur < Nawabganj. It was noticed that PLI values at the middle locations of the selected stretch of Ganga River were calculated to be higher as compared to the first, third, and last 3 locations indicating the constantly rising burden of metals as anthropogenic activities move downstream with Ganga's stretch. The PLI is less than one for all selected locations with the highest value obtained for Nawabganj (0.579). This indicates that the sediments of the selected stretch of river are currently safe from deterioration.

PLI values observed in 2021 are more uniform as compared to the values obtained in 2019 (Aggarwal et al. 2022) which showed huge variations (Refer to Table 4). PLI values for all sites except NWJ and MNR have increased drastically in 2021. The PLI values for MNR have decreased by 0.409 and for NWJ by 0.081. In 2019 MNR was the most polluted site followed by NWJ, whereas in 2021 NWJ was the most polluted location followed by MNR. This shows that the input of metals in the bed sediment at the local level has reduced but the spikes in the values of PLI at most locations show the input of metals from upstream areas has increased thus leading to their uniform spread and fewer variations in the PLI values as compared to that in 2021.

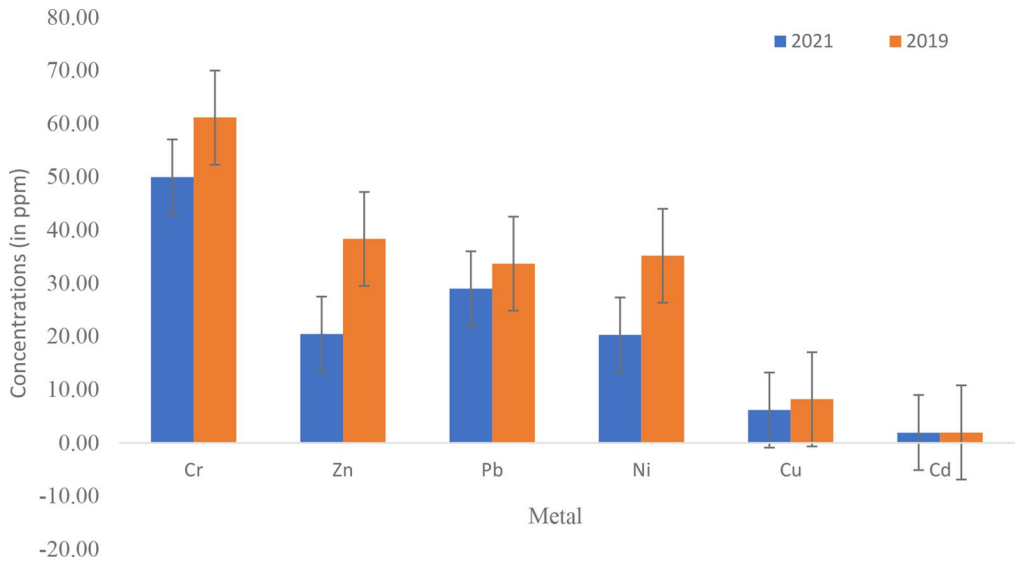



**Fig. 3** Average concentrations of metals recorded in 2019 & 2021

### 3.5 Relative Mobility of Metals

The Metal/Fe ratio is used to calculate the relative mobility of various metals as Fe metal is relatively abundant within the Earth's crust, Fe has been chosen as a conservative element for examination. As a result, their susceptibility to being impacted by the activities of humans is low. It is often considered that Fe has achieved a stable state and is no longer collected by the soil layer and is only obtained by erosion of land. The metals' relative mobility was noticed in the following decreasing order: Mn>Cr>Pb>Zn>Ni>Cu>Cd. This demonstrates that, of all heavy metals, Mn from sediments has the highest ability to solubilize in water under suitable conditions. Cd has the lowest propensity to dissolve in sediment. Jain et al. (2005) noticed a comparatively similar pattern in the relative mobility of metals in the sediments of the Hindon River. Copper and Cadmium become much less mobile under conditions if sufficient iron-rich precipitates are present and if the pH for optimal sorption is attained.

On comparing the average concentrations of metals obtained in this study with the average concentrations of metals recorded in the year 2019 by Aggarwal et al. (2022) as shown in Fig. 3, it was investigated that average concentrations of Fe, Mn, Cr, Zn, Pb, Ni, Cu and Cd reduced by 19.86, 46.25, 18.27, 46.70, 14.00, 42.33, 24.73 and 0.88% respectively. The maximum decline in concentrations was recorded for metals Mn, Zn, and Ni whereas Cd metal showed a negligible reduction in concentrations after 2 years. It indicates that Cd metal is strongly accumulated in the sediment of the Ganga River as compared to other metals. Countrywide lockdown due to COVID-19 and the shutdown of industries in the years 2020 and 2021 is the main reason for this reduction in metal concentrations. Further, the increasing popularity of Prayagraj and Kanpur city as tourist destinations, the expansion of road infrastructure, industries, and agricultural activities, and the expansion of cities along the banks of the Ganga River might all be responsible for the rising level of toxic metals. Given the religious and ecological importance of water, as well as the huge population fed by the Ganga River, it is past time to implement quick actions in order to prevent the further increase in pollution levels of Ganga's sediments and water, as well as the toxification of exposed creatures along the studied stretch.

## 4 Conclusion

According to the findings of this investigation, the mid locations of the selected length of the Ganga, namely Nawabganj and Manikpur, had worse sediment quality than the other places. The sequence of metal occurrence varied geographically. Fe and Mn were the most extensive metals found in the sediments of the selected area, with crustal origins. Cd had the lowest concentration of any metal. The concentration of heavy metals varied in the following decreasing order: Fe>Mn>Cr>Zn>Pb>Ni>Cu>Cd. Metal mobility was noticed in the following decreasing order: Mn>Cr>Pb>Zn>Ni>Cu>Cd. Furthermore, increasing relative mobility of metals such as Cu and Cd may exacerbate toxicity issues. The indices used to determine sediment quality found that sediments are significantly polluted by Cd and Pb metal and that their quality is deteriorating. The sediments of the selected length of the Ganga are therefore deteriorating, and while the situation is not as terrible, it still

needs care, particularly in recognized hotspots of Manikpur and Nawabganj. It is past time to implement advantageous control techniques to limit additional metal pollution in the Ganga while also preserving its ecology.

**Author Contributions** All authors contributed to the study conception and design. Sample collection and preparation, testing and analysis were performed by Mohit Aggarwal. The first draft of the manuscript was written by Mohit Aggarwal. All authors read and approved the final manuscript.

**Funding** The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Data Availability** Data will be made available on request.

## Declarations

**Ethics Approval** This is an observational study. No ethical approval is required.

**Consent to Participate** No participant was involved in this study, so no consent is needed.

**Consent to Publish** My manuscript does not contain any individual person's data in any form (including any individual details, images or videos). So no consent is needed.

**Competing Interests** The authors have no relevant financial or non-financial interests to disclose.

## References

Aggarwal M, Anbukumar S, Kumar TV (2022) Heavy metals concentrations and risk assessment in the sediment of Ganga River between Kanpur and Prayagraj, UP, India. Sādhana 47(4):195. https://doi.org/10.1007/s12046-022-01972-6

Aggarwal M, Anbukumar S, Kumar TV (2023) Analysis and pollution assessment of heavy metals in suspended solids of the middle stretch of river Ganga between Kanpur to Prayagraj, U.P., India. Sadhana 48(4):257. https://doi.org/10.1007/s12046-023-02325-7

Aggarwal M, Anbukumar S, Kumar TV (2024) Measurement of heavy metals content in suspended sediment of ganges river using atomic absorption spectrometry. MAPAN 39(4):913–930. https://doi.org/10.1007/s12647-024-00771-0

Aggarwal M, Anbukumar S, Kumar TV (2025) Assessment of pollution risks posed by trace elements in the riverbed sediment of the Ganges River in Eastern Uttar Pradesh, India. Water Supply 25(3):628–646. https://doi.org/10.2166/ws.2025.027

Ansari AA, Singh IB, Tobschall HJ (1998) Status of anthropogenically induced metal pollution in the Kanpur-Unnao industrial region of the gang plain. Environ Geol 38(1):25–33. https://doi.org/10.1007/s002540050397

Buaisha M, Balku S, Ozalp-Yaman S (2020) Heavy metal removal investigation in conventional activated sludge systems. Civil Eng J 6(3):470–477. https://doi.org/10.28991/cej-2020-03091484

Chander DVR, Venkobacher C, Raymahashay BC (1994) Retention of flyash derived copper in sediments of the Pandu river near Kanpur, India. Environ Geol 24:133–139. https://doi.org/10.1007/BF00767886

Chitrakshi, Haritash AK (2022) Appraisal of hydrochemistry and suitability assessment for water in an agriculture-dominated aquatic ecosystem of Rajasthan, India. Rend Fis Acc Lincei 33:851–866. https://doi.org/10.1007/s12210-022-01107-3

Dasgupta SP (1984) The Ganga basin, part II. Central Board for Prevention and Control of Water Pollution, New Delhi

Haritash AK, Kaushik CP (2007) Assessment of seasonal enrichment of heavy metals in respirable suspended particulate matter of a sub-urban Indian City. Environ Monit Assess 128(1):411–420. https://doi.org/10.1007/s10661-006-9335-1

Haritash AK, Gaur S, Garg S (2014) Assessment of water quality and suitability analysis of river Ganga in Rishikesh. Appl Water Sci 6:383–392. https://doi.org/10.1007/s13201-014-0235-1

Jain CK, Singhal DC, Sharma MK (2005) Metal pollution of sediment and water in the river hindon. Environ Monit Assess 105:193–207. https://doi.org/10.1007/s10661-005-3498-z

Jaiswal D, Pandey J (2018) Impact of heavy metal on activity of some microbial enzymes in the riverbed sediments: ecotoxicological implications in the Ganga river (India). Eco Environ Saf 150:104–115. https://doi.org/10.1016/j.ecoenv.2017.12.015

Kansal A, Siddiqui N, Gautam A (2013) Assessment of heavy metals and their interrelationships with some physicochemical parameters in eco-efficient rivers of Himalayan region. Environ Monit Assess 185:2553–2563. https://doi.org/10.1007/s10661-012-2730-x

Karaca O, Cameselle C, Reddy KR (2017) Acid pond sediment and mine tailings contaminated with metals: physicochemical characterization and electrokinetic remediation. Environ Earth Sci 76:408. https://doi.org/10.1007/s12665-017-6736-0

Khan KI, Mata MN, Martins J, Nasir A, Dantas RM, Correia AB, Saghir US (2022) Impediments of green finance adoption system: linking economy and environment. Emerg Sci J 6(2):217-237. https://doi.org/10.28991/ESJ-2022-06-02-02

Milliman JD, Meade RH (1983) Worldwide delivery of river sediments to ocean. J Geol 91:1–19. https://www.jstor.org/stable/30060512

Mirra R, Ribarov C, Valchev D, Ribarova I (2020) Towards energy efficient onsite wastewater treatment. Civil Eng J 6(7):1218–1226. https://doi.org/10.28991/cej-2020-03091542

Muller G (1969) Index of geoaccumulation in sediments of the rhine river. GeoJournal 2:108–118

Osakwe JO, Adowei P, Jnr MH (2014) Evaluation of heavy metal species in bottom sediments from Imo river system southeastern Nigeria. Res J Chem Sci 4(6):23–30. http://www.isca.in/rjcs/Archives/v4/i6/4.ISCA-RJCS-2014-64.pdf

Purushothaman P, Chakrapani GJ (2007) Heavy metals fractionation in Ganga river. Environ Monit Assess 132:475–489. https://doi.org/10.1007/s10661-006-9550-9

Satarug S, Garrett SH, Sens MA, Sens DA (2010) Cadmium, environmental exposure, and health outcomes. Environ Heal Perspect 118:182–190. https://doi.org/10.1289/ehp.0901234

Siddiqui E, Verma K, Pandey U, Pandey J (2019) Metal contamination in seven tributaries of the Ganga river and assessment of human health risk from fish consumption. Arch Environ Contam Toxicol 77:263–278. https://doi.org/10.1007/s00244-019-00638-5

Singh M, Pandey J (2022) COVID-19 lockdown–driven changes in the Ganga river ecosystem in response to human perturbations. Environ Monit Assess 194(12):858. https://doi.org/10.1007/s10661-022-10507-9

Singh M, Müller G, Singh IB (2003) Geogenic distribution and baseline concentration of heavy metals in sediments of the Ganges river, India. J Geochem Explor 80(1):1–7. https://www.sciencedirect.com/science/article/abs/pii/S0375674203000165?via%3Dihub

Subramanian V, Grieken RV, Dack LV (1987a) Heavy metals distribution in the sediments of Ganges and Brahmaputra rivers. Environ Geol Water Sci 9:93–103. https://doi.org/10.1007/BF02449940

Subramanian V, Sitasawad R, Abbas N, Jha PK (1987b) Environmental geology of the Ganga river basin. J Geol Soc India 30:335–355. https://www.geosocindia.com/index.php/jgsi/article/view/65931

Tang W, Shan B, Zhang H, Zhang W, Zhao U, Ding Y, Rong N, Zhu X (2014) Heavy metal contamination in the surface sediments of representative limnetic ecosystems in Eastern China. Sci Rep 4(1):7152. https://doi.org/10.1038/srep07152

Tomlinson DC, Wilson JG, Harris CR, Jeffrey DW (1980) Problems in the assessment of heavy metal levels in estuaries and the formation of a pollution index. Helgel Meeresuntters 33:566–575. https://doi.org/10.1007/BF02414780

Turekian KK, Wedepohl KH (1961) Distribution of elements in some major units of the Earth's crust. Bull Geol Soc Am 72:175–192. https://pubs.geoscienceworld.org/gsa/gsabulletin/article/72/2/175/5262/Distribution-of-the-Elements-in-Some-Major-Units

**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.

2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT) | 979-8-3315-4335-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/DICCT64131.2025.10986481

# Keystroke Dynamics Based User Authentication Focusing on Fixed Text Approaches

Anurag Kamra
*Dept. of Computer Engineering*
*Delhi Technological University*
New Delhi, India
anuragkamra7@gmail.com

Sanya Khurana
*Dept. of Computer Engineering*
*Delhi Technological University*
New Delhi, India
sanya14n@gmail.com

Anurag Goel
*Dept. of Computer Engineering*
*Delhi Technological University*
New Delhi, India

***Abstract*—Protecting sensitive information has become highly critical in the digital age as traditional authentication methods like passwords and PINs are susceptible to theft and brute-force attacks. Even though physical biometric systems, such as fingerprints, iris scans, and facial recognition, offer better security, they usually demand expensive hardware and remain prone to spoofing techniques, such as fake fingerprints, voice recordings, and deepfakes. Behavioral biometrics, which analyzes user interaction patterns such as keystroke dynamics, provides a more secure and user-friendly alternative by focusing on unique typing behaviors. This paper introduces a hybrid CNN-LSTM architecture designed to effectively capture both spatial patterns and temporal dependencies in keystroke dynamics data. The proposed model aims to enhance authentication accuracy while maintaining usability, offering a robust solution to modern security challenges.**

***Index Terms*—Behavioral biometrics, Keystroke dynamics, CNN-LSTM**

## I. Introduction

In the field of global cybersecurity, protecting sensitive data and computer systems remains a persistent challenge. Unauthorized access to such information poses significant risks, potentially leading to financial losses or the unintended disclosure of highly confidential data. Ensuring information security involves with user authentication, which serves as the critical first line of defense against unauthorized access. Achieving a balance between easy access for authorized users and preventing unauthorized intrusions is crucial. For years, traditional authentication methods, predominantly based on username and password combinations, have been the foundation of cybersecurity systems. However, these methods are plagued by significant weaknesses, such as password sharing, shoulder surfing, brute force attacks, dictionary attacks, guesswork, phishing attempts, and other evolving threats. [1]

Despite efforts to increase awareness about these risks, a significant number of individuals remain indifferent to password security. Studies show that 24% of users rely on variations of the same eight commonly used passwords. Additionally, password reuse is widespread; half of IT professionals admit to reusing passwords, the average employee reuses a single password approximately 13 times, and nearly half of all individuals use the same password across all their accounts.

To address these limitations, biometric technology has emerged as a reliable alternative for authentication and verification [2]. Biometric authentication methods are generally categorized into two types. The first, known as physical biometrics, relies on physiological traits such as fingerprints, facial recognition, and voice identification. However, this approach has several challenges, including the high cost of additional system requirements [3], vulnerability to spoofing methods like fake fingerprints, voice recordings, and deepfakes, as well as potential changes due to aging, injuries, or illnesses. These drawbacks highlight the need for a shift toward behavioral-based authentication systems, which leverage unique individual traits such as typing patterns, swipe motions, or mouse dynamics[2].

Keystroke dynamics-based user authentication is a very effective way to enhance security in a practical manner. This method involves analyzing the typing behavior of a user to identify individuals based on their unique rhythmic patterns. It can be used as an alternative to or complement one-time codes, causing minimal inconvenience to users. Keystroke dynamics features are derived by examining the timing details of KeyPress, KeyRelease, and Hold events[4].

Keystroke verification techniques are broadly categorized into two types: static and dynamic [5]. Static verification, also known as fixed-text approach, which will be the major focus for us throughout this paper, analyzes typing patterns during specific events like login, registration, or enrollment. On the other hand, dynamic verification, also known as free-text approach continuously monitors and evaluates a user's typing behavior throughout their session. A key advantage of keystroke dynamics is its ability to detect imposters by analyzing unique typing rhythms, and it is relatively inexpensive to implement [1]. However, this method faces challenges such as user fatigue, significant variations in typing styles, medical conditions, typing proficiency, and differences in keyboard hardware [6].

This research paper investigates a hybrid architecture, CNN-LSTM in the present application field of keystroke dynamics-based user authentication. The suggested architecture is based on the space feature extraction capabilities of the CNNs and the sequential pattern recognition capabilities of Long Short-Term Memory (LSTM) networks to yield strong performances in keystroke behavior identification uniquely. The application

979-8-3315-4335-8/25/$31.00 ©2025 IEEE



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:56:45 UTC from IEEE Xplore. Restrictions apply.

of CNNs for feature extraction and LSTMs for temporal modeling will enable the model with the ability to learn local as well as global patterns in keystroke dynamics.

## II. Related Work

Research in the area of user authentication using keystroke dynamics has been a major area of interest in recent years because of the increasing requirement for efficient and user-friendly authentication systems that are highly secure. Various methods have been discussed to increase the efficiency and accuracy of keystroke dynamics-based systems, which have been further categorized into fixed-text and free-text techniques.

Arsh et al. (2024) discussed various methods of authenticating using keystroke dynamics, highlighting the efficiency of hybrid machine learning methods in handling user behavior variability, improving the classification accuracy, etc. [1]. Risto and Graven, 2024 highlighted the fixed-text dataset and feature extraction in making the system reliable [4]. Elliot et al. (2019) compared Decision Tree and Gradient Boosting wherein algorithm selection has been highly emphasized [6].

Neural network techniques have also been explored extensively. Shanmugapriya and Padmavathi (2010) discussed the application of neural networks in keystroke dynamics authentication. The study proved the versatility of neural networks in detecting slight typing pattern differences to enhance the accuracy of authentication [5]. AbdelRaouf et al. (2023) discussed an efficient CNN-based technique in improving the accuracy of authentication, employing convolutional layers for spatial feature extraction from keystroke data [8].

Free-text keystroke dynamics have also been a popular area of research in recent years. Li et al. (2021) proposed a deep learning system for free-text authentication, delivering stable results by employing contextual typing features [9]. Wang et al. (2022) combined keystroke and mouse dynamics using scene-independent features, highlighting an impressive multimodal approach for hybrid scenarios [11].

The necessity of keystroke-based authentication has been further reinforced by research in password security and user behavior. Google and Harris Poll's (2019) reports emphasized the public's attitudes toward passwords and the shocking statistics regarding poor password practices. Likewise, the LastPass State of the Password Report (2020) offered key observations on password usage patterns, emphasizing the necessity for alternative authentication techniques such as keystroke dynamics.

Existing research has laid a solid groundwork for the creation of keystroke dynamics-based authentication systems. Sophisticated machine learning models, neural networks, and high-quality data sets are further driving innovation in this area by overcoming the challenges of user variability and data security.

## III. Proposed Work

In this section, a hybrid CNN-LSTM-based architecture is presented for user authentication based on keystroke dynamics. The model combined is feature extraction based on convolutional layers and sequential pattern analysis based on LSTM layers. The system must be able to capture the temporal and spatial properties of keystroke data, thus being robust and highly adaptable for various fixed-text scenarios. Every subsection discusses important parts of the proposed methodology. The dataset employed for the Fixed Text task is CMU Keystroke Benchmark Dataset [20], an industry standard.

### A. Preprocessing Pipeline

The preprocessing pipeline formats and prepares raw keystroke data to be fed into the CNN-LSTM model.

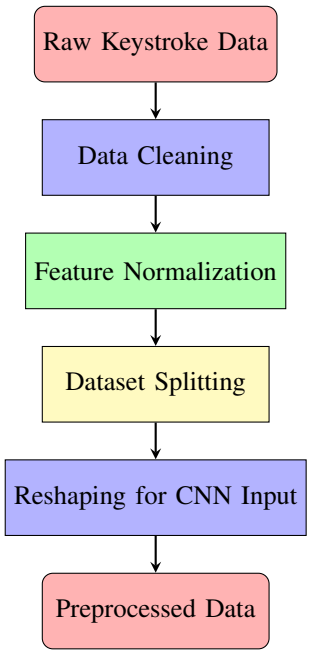



Fig. 1: Data Preprocessing Pipeline

*1) Data Cleaning:* Redundant columns are eliminated from the dataset. All identifiers, for instance, subject and sessionIndex, and other non-feature data are removed to ensure that only the respective input features reach the model. Timing features such as dwell times and flight times are normalized using StandardScaler from sklearn to have a zero mean and unit variance. The processes ensure a fair contribution by all the features toward making the training process faster [4][8].

*2) Split the Datasets:* The preprocessed data is shuffled and divided into training and testing subsets with the standard 80:20 ratio to prevent overfitting. Also, during training, a part of the training data is held back as validation data to monitor model performance robustly [10][13].

*3) Reshaping for CNN Input:* The reshaped input features are of 3D tensor with dimensions as (samples, timesteps, features) that fulfill the requirement of Conv1D layers' input. This reshaping process is very critical for capturing the temporal sequence intrinsic in keystroke dynamics [12]. These preprocessing steps are made to allow the model efficiently extract



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:56:45 UTC from IEEE Xplore. Restrictions apply.

features while adapting to a user-specific keystroke pattern due to fixed-text inputs.

### B. Model Architecture

The hybrid CNN-LSTM architecture employs the best aspects of convolutional and recurrent neural networks to model keystroke dynamics.

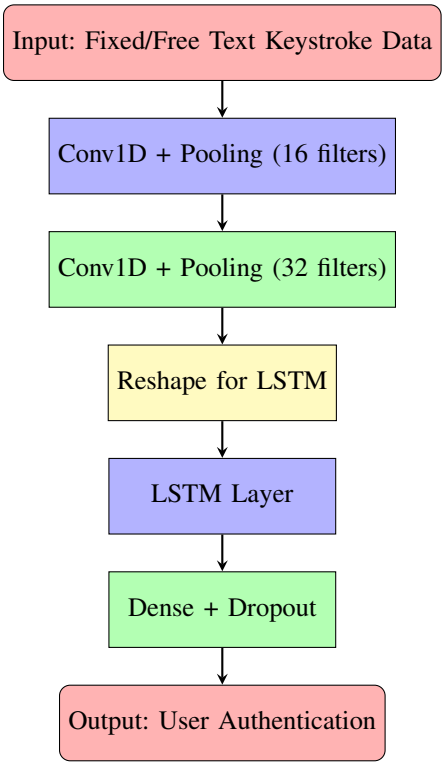



Fig. 2: Proposed Model Architecture

*1) Convolutional Layers:* The model is initiated with a Conv1D layer of 16 filters, kernel size of 3, followed by AveragePooling1D to reduce the dimensionality of the features while keeping important local features. A batch normalization and a ReLU activation function are employed for training stability and introducing non-linearity, respectively. Dropout is also applied to prevent overfitting [8][12]. A second convolutional block consists of a Conv1D with 32 filters (kernel size = 3), followed by another AveragePooling1D. The block also employs batch normalization, ReLU activation, and dropout to further enhance the local feature extraction process [5][12].

*2) Reshaping for LSTM Input:* Following convolutional processing, feature maps are reformed into a 2D sequence format, maintaining temporal dependencies in keystroke data prior to being sent to the LSTM layer [9][13].

*3) LSTM Layer:* The application of a single LSTM layer which identifies long-term dependencies in the sequential data employs 64 units. The layer is given a tanh activation function and is merged with batch normalization and dropout, making the model robust and capable of generalizing from diverse typing patterns [8][12].

*4) Fully connected layers:* The output of the LSTM is sent to a Dense layer with 64 units and followed by another dropout layer. Lastly, the output layer performs the ultimate classification using a sigmoid activation function, which gives a binary classification as to whether the user is real or an imposter [1][10].

*5) Training Details:* The model is trained using the Adam optimizer with a learning rate of 0.001 and a binary cross-entropy loss function, ensuring efficient convergence. Evaluation metrics include accuracy, precision, recall, and AUC, providing a comprehensive measure of the system's performance [13][12].

The model architecture and the algorithm used can be condensed into the following block of pseudo code.

**Algorithm 1** CNN-LSTM Model Pipeline

**Data Preparation:**
Load and normalize dataset.
Shuffle, split, and reshape data to $(timesteps, 1)$.
**Model Building:**
**Input:** Define input shape $(timesteps, 1)$.
**Conv Block 1:** Apply Conv1D $\rightarrow$ AvgPool $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ Dropout.
**Conv Block 2:** Apply Conv1D $\rightarrow$ AvgPool $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ Dropout.
Reshape output for LSTM.
**LSTM Block:** Apply LSTM $\rightarrow$ Tanh $\rightarrow$ BatchNorm $\rightarrow$ Dropout.
**Dense Layers:** Apply Dense (ReLU) $\rightarrow$ Dropout $\rightarrow$ Output Dense (sigmoid).
**Compile:** Use Adam optimizer and binary crossentropy loss.
**Train:** Fit model with training data and validation split; use checkpoint to save best model.
**Evaluate:** Load best model and evaluate on test data.

After the training and testing of the dataset, results are analyzed followed by comparison against the existing methodologies applied in the domain; Here, the subsequent sections discuss an analysis of results obtained from testing the dataset.

## IV. Result and Discussions

### A. Dataset Description

The dataset used throughout this study is the CMU Benchmark Keystroke dataset [20], which is a fixed-text, i..e.. static keystroke dataset, and is the most used dataset in the field of keystroke dynamics.
It is a publicly available dataset gathered by Carnegie Mellon University for the analysis of keystroke dynamics-based authentication. It consists of keystroke timing data of users typing predefined strings of text. The data were gathered by



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:56:45 UTC from IEEE Xplore. Restrictions apply.

requiring users to type a given text repeatedly in a series of repeated sessions, using standard keyboards without any special equipment.

The graph below (Fig. 3) plots the latency (flight time) between two adjacent key presses, computed from the ”DD” (down-down) latencies in the data, which is the interval between one key press and the next on the y-axis; against the sequential position of keystrokes within a typing session on the x-axis, for 6 classes,, i.e., subjects (or users), who participated in the dataset preparation. Each point on the x-axis is mapped to a specific key press in the typed string (1st key, 2nd key, etc.).

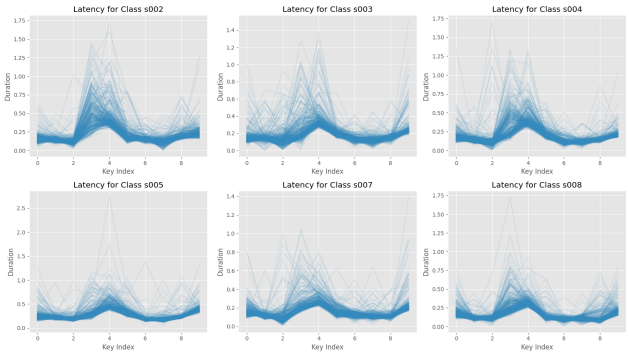



Fig. 3: Keystroke Latency Patterns across Different Users

The dataset captures the most essential features of hold time (dwell time), i.e., key hold time, and flight time, i.e., the release of a key and the press of the next one, capturing precise timestamps of key press and release for each keystroke action. Having many samples gathered from each user, the dataset easily captures differences among the typing behaviors of people. Because of its structured form and high-quality annotated data, it is very much used in keystroke dynamics analysis, especially in statistical and machine learning-based authentication methods.

### *B. Result Analysis and Comparison*

The performance and outcome of the proposed hybrid CNN-LSTM model, in comparison to existing fixed-text keystroke dynamics authentication models is reported in Table-II. This table puts more focus on the excellent progress our approach has taken.

The test metrics achieved by our model on the chosen dataset are shown in Table-I.

TABLE I: Proposed Model Result Metrics

| Metric | Value |
|---|---|
| Accuracy | 0.99 |
| Precision | 0.86 |
| Recall | 0.78 |
| AUC | 0.96 |
| FPR | 0.19 |
| TPR | 0.99 |

These results demonstrate that the proposed model not only achieves high accuracy and AUC but also maintains a balanced performance between precision and recall. The high AUC indicates that the model is robust across different decision thresholds, which is crucial for real-world authentication scenarios [4][13].

Figures 4 and 5 plot the ROC and Precision-Recall curves, which demonstrate the effectiveness of adding convolutional layers for spatial feature extraction into an LSTM architecture that recognizes temporal patterns and positions this work above the existing RNN-based methods and other fixed-text models [8][11].

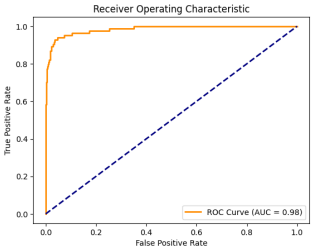



Fig. 4: ROC Curve

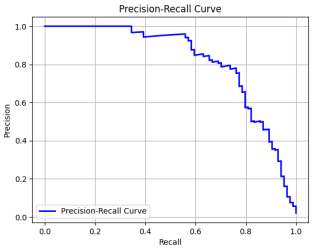



Fig. 5: Precision-Recall Curve

It is evident in the case of the ROC curve that there exists a high true positive rate with a low false positive rate (Fig. 4), whereas the Precision-Recall curve kept the precision higher with any value of the recall within those limits (Fig. 5).

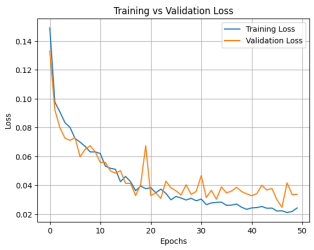



Fig. 6: Loss Curve

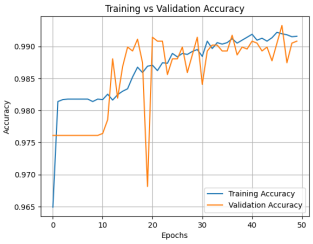



Fig. 7: Accuracy Curve

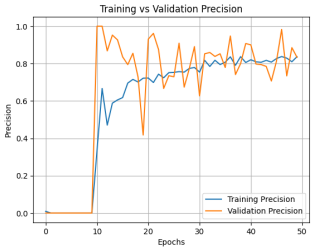



Fig. 8: Precision Curve

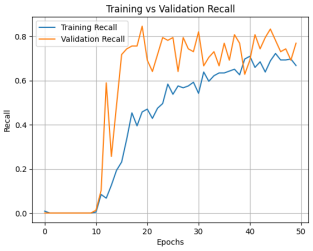



Fig. 9: Recall Curve

Fig. 6 and 7 present the Loss and the Accuracy Curves respectively, showing how the model adapts well and learns meaningful keystroke patterns. The small fluctuations of validation accuracy appear to indicate a certain sensitivity towards the variations present in the test data, these may be compen-



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:56:45 UTC from IEEE Xplore. Restrictions apply.

sated by adding more regularization techniques or by using even richer training data.

Fig. 8 and 9 show the Precision and Recall curves. The steep rise in precision in early epochs indicates the model's capacity to learn discriminative features very quickly. The curves show an upward trend, which means that the model continues improving its metrics, making it apt for authentication tasks requiring high confidence in positive classifications.

The set of figures shows that the proposed model is appropriate to represent keystroke patterns that have complex dynamics.

TABLE II: Comparison of Model Result Metrics

| Reference Paper | Model Used | Accuracy |
|---|---|---|
| Thakare et. al [2] | ANN | 91.8% |
| Graham et. al [6] | NN | 95% |
| T. Ramu et. al [18] | SVM | 95% |
| Risto et. al [4] | SVM+KNN | 97% |
| Iyer et. al [12] | CNN-RNN | 98.8% |
| **Proposed** | **CNN-LSTM** | **99.04%** |
| Rahman et. al [17] | RnF | 99.9% |
| Graham et. al [6] | RnF | 100% |

The above table (Table-II) compares the results obtained by the model used in our work with the existing models that have been applied to this task.

The primary reason for the superior performance of our model is the novel integration of the CNN and LSTM architectures. The CNN layers are capable of extracting discriminative local features from the raw keystroke data, while the LSTM layers capture long-term temporal dependencies, providing a comprehensive representation of the user's typing behavior [8][11]. This combination empowers our approach to bridge the gaps in the existing pure models being applied in the field of keystroke based user authentication and the traditional RNN-based models with better metrics as shown in Table-II.

Overall, the results and discussions confirm that the proposed hybrid CNN-LSTM model offers a robust and efficient solution for fixed-text keystroke dynamics authentication, with promising implications for enhancing security in practical applications.

## V. Conclusion

A considerable improvement is presented over fixed-text keystroke dynamics-based user authentication with the proposed hybrid CNN-LSTM model. With the model effectively linking the use of convolutional layers on features to LSTM layers that capture temporal dependencies, individuals' unique typing patterns are learned. Thus, it is shown that our experimental results accurately yield significant performance for the system, with high accuracy, precision, recall, and high AUC on the CMU DSL-StrongPassword dataset [20]. These results demonstrate that the proposed approach can achieve more secure and efficient authentication than current password-based mechanisms and even previously developed biometric systems.

## VI. Future Work

The future work for the current study and findings involves focusing on expanding the scope of the system and also making it robust. An important step is dataset compatibility, achieved by adding more samples and introducing data from a variety of sources. This should help capture the widest range of user behaviors, improving the generalizability of the model. Efforts can also be made in order to better the precision of our work, which can be achieved via various methods, including using a deeper cnn layer with more filters, using self-attenuation or transformer layers, or by using focal loss in place of binary cross-entropy loss function.

Another important direction is to extend the current work to free-text keystroke data. Free-text inputs are more variable than fixed-text, and thus pose new challenges. By adapting the CNN-LSTM model to handle free-text sequences, the system can be used in more dynamic and continuous authentication settings. Additional efforts will also look into other deep learning methods, including attention mechanisms, to enhance feature extraction and further improve model performance. Such advancements will seek to make the authentication framework more flexible and scalable for real-world applications. Some free text datasets which this study can be extended to, are Clarkson University Keystroke Dataset [19] and Mendeley Free Text Dataset [16].

## References


[1] A. Arsh, N. Kar, S. Das, and S. Deb, "Multiple Approaches Towards Authentication Using Keystroke Dynamics," *Procedia Computer Science*, vol. 235, pp. 2609–2618, 2024, doi: 10.1016/j.procs.2024.04.246.

[2] A. Thakare, S. Gondane, N. Prasad, and S. Chigale, "A Machine Learning-Based Approach to Password Authentication Using Keystroke Biometrics," in *Advances in Communication, Devices and Networking*, 2021, pp. 409–419, doi: 10.1007/978-981-16-0289-4_30.

[3] J. Moody, "Public Perceptions of Biometric Devices: The Effect of Misinformation on Acceptance and Use," *Issues in Informing Science and Information Technology*, vol. 1, 2004, doi: 10.28945/775.

[4] H. Risto and O. Graven, "Fixed-text keystroke dynamics authentication data set—collection and analysis," *Annals of Telecommunications*, vol. 79, 2024, doi: 10.1007/s12243-024-01039-z.

[5] V. S. Shanmugapriya and G. Padmavathi, "Keystroke dynamics authentication using neural network approaches," in *Information and Communication Technologies: International Conference, ICT 2010, Kochi, India*, Springer, 2010, pp. 686–690.

[6] K. Elliot et al., "A comparison of machine learning algorithms in keystroke dynamics," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2019, pp. 127–132.

[7] H. Lee, J. Hwang, D. Kim, S. Lee, S.-H. Lee, and J. S. Shin, "Understanding Keystroke Dynamics for Smartphone Users Authentication and Keystroke Dynamics




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:56:45 UTC from IEEE Xplore. Restrictions apply.

on Smartphones Built-In Motion Sensors," *Security and Communication Networks*, vol. 2018, pp. 1–10, 2018, doi: 10.1155/2018/2567463.

[8] H. AbdelRaouf et al., "Efficient Convolutional Neural Network-Based Keystroke Dynamics for Boosting User Authentication," *Sensors*, vol. 23, no. 10, p. 4898, 2023, doi: 10.3390/s23104898.

[9] J. Li, H.-C. Chang, and M. Stamp, "Free-Text Keystroke Dynamics for User Authentication," *arXiv*, 2021, doi: 10.48550/arXiv.2107.07009.

[10] Y. B. W. Piugie, J. Di Manno, C. Rosenberger, and C. Charrier, "Keystroke dynamics-based user authentication using deep learning neural networks," in *2022 International Conference on Cyberworlds (CW)*, 2022, pp. 220–227.

[11] X. Wang et al., "User Authentication Method Based on Keystroke Dynamics and Mouse Dynamics with Scene-Irrelated Features in Hybrid Scenes," *Sensors (Basel)*, vol. 22, no. 17, p. 6627, Sep. 2022, doi: 10.3390/s22176627.

[12] K. N. Iyer, H. K. Upadhyay, and R. S. Vaghela, "Enhancing User Authentication via Deep Learning: A Keystroke Dynamics Approach," *International Journal of Next-Generation Computing*, vol. 15, no. 2, 2024, doi: 10.47164/ijngc.v15i2.1706.

[13] D. Escobar Grisales, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, and J. R. Orozco-Arroyave, "Identity verification in virtual education using biometric analysis based on keystroke dynamics," *TecnoLógicas*, vol. 23, no. 47, pp. 197–211, 2020.

[14] S. Krishnamoorthy, L. Rueda, S. Saad, and H. Elmiligi, "Identification of user behavioral biometrics for authentication using keystroke dynamics and machine learning," in *Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications*, 2018.

[15] K. Shekhawat and D. P. Bhatt, "Recent Advances and Applications of Keystroke Dynamics," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 2019.

[16] N. González and E. Calot, "Dataset of Human-written and Synthesized Samples of Free-Text Keystroke Dynamics to Evaluate Liveness Detection Methods," *Mendeley Data, V2*, 2022, doi: 10.17632/mzm86rcxxd.2.

[17] A. Rahman et al., "Multimodal EEG and keystroke dynamics based biometric system using machine learning algorithms," *IEEE Access*, vol. 9, pp. 94625–94643, 2021.

[18] T. Ramu, K. Suthendran, and T. Arivoli, "Machine learning based soft biometrics for enhanced keystroke recognition system," *Multimedia Tools and Applications*, vol. 79, pp. 10029–10045, 2020.

[19] L. Tappert, L. Villani, and S. Cha, "Clarkson University Keystroke Dataset," Clarkson University, Potsdam, NY, USA. [Online]. Available: https://www.clarkson.edu/research/keystroke-dataset.

[20] F. Monrose and A. D. Rubin, "Keystroke Dynamics Benchmark Dataset," Carnegie Mellon University, 1997. [Online]. Available: http://www.cs.cmu.edu/ keystroke.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:56:45 UTC from IEEE Xplore. Restrictions apply.

# Machine Learning for Advancing Electric Vehicles Security Leveraging CICEVSE2024

Anmol Geetesh Kumar
*Department of Software Engineering*
*Delhi Technological University*
New Delhi, India
anmolanmol981@gmail.com

Amandeep
*Department of Software Engineering*
*Delhi Technological University*
New Delhi, India
amandeepkhajan@gmail.com

Sonika Dahiya
*Department of Software Engineering*
*Delhi Technological University*
New Delhi, India
sonika.dahiya@dtu.ac.in

2025 International Russian Smart Industry Conference (SmartIndustryCon) | 979-8-3315-1124-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/SMARTINDUSTRYCON65166.2025.10986082

***Abstract*—The emergence of Electric Vehicles (EVs) and related charging infrastructure marks a significant shift toward sustainable transportation solutions. However, the increasing deployment of Electric Vehicle Supply Equipment (EVSE) has introduced a new landscape for cybersecurity challenges. To address this critical issue, we utilize machine learning techniques for enhancing EVSE security systems leveraging CICEVSE2024, a new and comprehensive dataset tailored to advancing research on EVSE security. The dataset provides detailed records of EVSE activities, capturing both benign operations and attacks like Reconnaissance, Backdoor, Cryptojacking, Denial of Service (DoS), etc. in multiple states. Our research focuses on analyzing the performance of machine learning algorithms for securing EVSE systems. This study uses two techniques, namely Host Anomaly Detection Model (HADM) and Power Anomaly Detection Model (PADM), wherein machine learning algorithms are used to identify and classify attack scenarios with high performance. This research emphasizes the prospects of machine learning technologies in fortifying EVSE systems against cybersecurity threats, thereby contributing to the broader goal of ensuring the resilience of clean energy powered transportation solutions.**

***Keywords—electric vehicle, datasets, cybersecurity, privacy, detection, machine learning***

## I. INTRODUCTION

The adoption of Electric Vehicles (EVs) and related charging infrastructure denotes a major shift for sustainable transportation by reducing emissions and promoting clean energy solutions. Central to this shift is the deployment of Electric Vehicle Supply Equipment (EVSE), which provides essential charging services to EVs. However, the integration of EVSE systems into smart grids and other infrastructure introduces significant cybersecurity risks. As charging stations become more connected, they expose themselves to an array of threats, including cyberattacks targeting communication networks and system vulnerabilities [1]. The cybersecurity risk landscape includes vulnerabilities like unsecured interfaces, interconnected assets, and integration into Building Management Systems [2], which can unintentionally create vulnerabilities [3]. Moreover, EVSE authorization systems that rely on technologies such as RFID, Bluetooth, and Wi-Fi present common entry points for attackers, with the potential to compromise networks and disrupt power grids [4]. To mitigate these risks, comprehensive cybersecurity strategies, technical risk assessments, and proactive countermeasures must be established. Cybersecurity threats affecting EVSE systems are multifaceted, with common attack types such as reconnaissance, denial of service (DoS), backdoor exploits, cryptojacking, etc. These attacks exploit vulnerabilities in communication protocols, upstream services, and device configurations, impacting power systems and infrastructure [4]. A major point of concern is the Open Charge Point Protocol (OCPP) 1.6, which relies on unencrypted WebSocket communications and is vulnerable to machine-in-the-middle attacks, remote code execution, and DoS threats [5]. Addressing these challenges requires a mix of IT and OT best practices, strategic threat detection, and machine-learning-based monitoring approaches [6]. To support cybersecurity research in this domain, this paper utilizes CICEVSE2024, a new dataset comprising both benign and malicious activity patterns during idle and charging states. It captures various attack scenarios like reconnaissance, DoS, backdoor, cryptojacking, etc. This dataset fills a significant gap in publicly available resources and enables the evaluation of machine learning models for anomaly detection.

Our research focuses on two distinct approaches: the Host Anomaly Detection Model (HADM) and Power Anomaly Detection Model (PADM). These techniques employ machine learning algorithms to analyze system behaviors and power data to find anomalies efficiently. These systemic approaches offers new insights into securing EVSE systems, contributing to the development of proactive strategies for counter emerging cyber threats. As the world tackles the twin problem of Climate Change & Global Warming, the significance of Electric Vehicles is only going to increase in the near future, by utilizing machine learning solutions our research provides valuable insights which are essential for advancing cybersecurity solutions and ensuring the resilience of EV infrastructure.

## II. RELATED WORKS

Recent works have extensively focused on enhancing cybersecurity and anomaly detection in electric vehicle (EV) systems, addressing key vulnerabilities, intrusion detection mechanisms, and advanced methodologies. For instance, Mavikumbure et al. [7] introduced the ResNet Autoencoder-based Cyber-Physical Anomaly Detection System *(*Cy-Phy ADS*)* to identify anomalies in EV Controller Area Network

2025 International Russian Smart Industry Conference (SmartIndustryCon) | 979-8-3315-1124-1/25/$31.00 ©2025 IEEE



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:36 UTC from IEEE Xplore. Restrictions apply.

(CAN) communication, outperforming baseline models. Similarly, Kern et al. [8] developed a hybrid intrusion detection system combining regression-based forecasting and anomaly detection, which was evaluated using real-world attack data. Tanyildiz et al. [9] explored machine learning-based intrusion detection for EV supply equipment security, emphasizing GAN-augmented models. Comparative studies demonstrated that ResNet Autoencoders outperformed physics-based methods for detecting anomalies in EV charging stations [10]. Additionally, recent works have also explored DDoS attacks with the CICEV2023 dataset simulating these events in smart grid infrastructures [11] and studied text-based anomaly detection for event trigger analysis across 30 event types [12]. Mishra et al. [13] proposed using one-class SVMs to detect anomalies in CAN bus communications. Notably, He et al. [14] provided real-world EV charging datasets for assessing battery health and anomaly detection. Threats to EV charging infrastructure have been examined, particularly focusing on vulnerabilities in Open Charge Point Protocol (OCPP). Studies identified multiple attack vectors like machine-in-the-middle and remote code execution [15], with proposed countermeasures such as secure shell tunnels and adherence to the IEC 62351 standard [16]. Combining AI and ML approaches, Almazrouei et al. [17] highlighted the growing role of AI and ML in securing EV systems. Collectively, these diverse methodologies, datasets, and insights collectively address security vulnerabilities, attack detection, and anomaly prediction for EV systems and smart grids.

## III. Background

- **Isolation Forest:** Isolation Forest is an algorithm for detecting anomalies in an unsupervised manner that utilizes randomized splits to recursively partition data for identifying anomalies. It is particularly effective for huge datasets because it takes use of the fact that anomalies are simpler to isolate because they are uncommon and behave differently from normal points.
- **AutoEncoder:** It is a type of neural network that focuses on learning effective data representations by compressing input data into a lower dimensional latent space through an encoder and then reconstructing it to its original form using a decoder. It finds usage in dimensionality reduction, noise removal, unsupervised learning, etc.
- **XGBoost:** It is a potent, scalable gradient-boosting machine learning algorithm. By employing strategies like regularization, parallel processing, and handling missing inputs, it enhances performance and leverages decision trees as foundation learners.
- **Local Outlier Factor:** LOF is a method employed for identifying irregularities in data without needing labeled information. It works by comparing how densely a data point is surrounded by its neighbours. A point is deemed an outlier if its density is significantly lower than that of neighbouring points.
- **OC-SVM:** It represents a specialized form of Support Vector Machine designed for anomaly detection. Instead of requiring labeled instances from both normal and abnormal classes, OC-SVM trains exclusively on normal data to create a boundary that encapsulates typical instances in the feature space.
- **Gaussian Mixture:** It is a probability based model that shows an input space as a mixture of several Gaussian distributions, each with its own covariance and mean. It assumes that data points are generated from these underlying normal distributions, allowing for the modeling of complex, multimodal data patterns.
- **Transformer:** It is a machine learning model that enlists transformer architecture to classify input data into one of two categories. A transformer backbone for processing data is followed by a classification layer that outputs probabilities for each class using a softmax function.

## IV. Methodology

### A. Dataset

The CICEVSE2024, encompasses total 3 subsets of data, namely HPC-Kernel Events, Power Consumption and Network Traffic. For our research, we've focused on HPC-Kernel & Power Consumption data. HPC and Kernel Events contains system kernel event logs, state transitions, and attack indicators, providing critical insights for detecting Denial-of-Service (DoS) events by capturing anomalous patterns within the system's kernel activity. Power Consumption data focuses on bus voltage, shunt voltage, current, and power measurements, which offers valuable insights into energy patterns that may signal irregularities or anomalies associated with malicious activities.

### B. Data Preprocessing

Data cleaning phase involved filtering corrupted entries, renaming columns, removing NaN values, and deriving new features for DoS detection. Unnamed columns were dropped, and column types were validated. Power Consumption data required converting timestamps to datetime, handling invalid rows, normalizing categorical variables, removing null values, and ensuring type compatibility. Data normalization was performed using Min-max scaling and standardization. A binary target variable was created by encoding 'attack' labels.

### C. Training and Testing Partitions

Data was divided in an 80-20 manner across training and testing sets, where 80% of the data was allocated for training the models and the rest 20% of the data was being used for the testing phase. Fig. 1. represents the complete workflow flowchart.

## V. Results and Discussion

### A. Performance Metrics

Following metrics were used to evaluate the efficacy of the aforementioned system:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:36 UTC from IEEE Xplore. Restrictions apply.

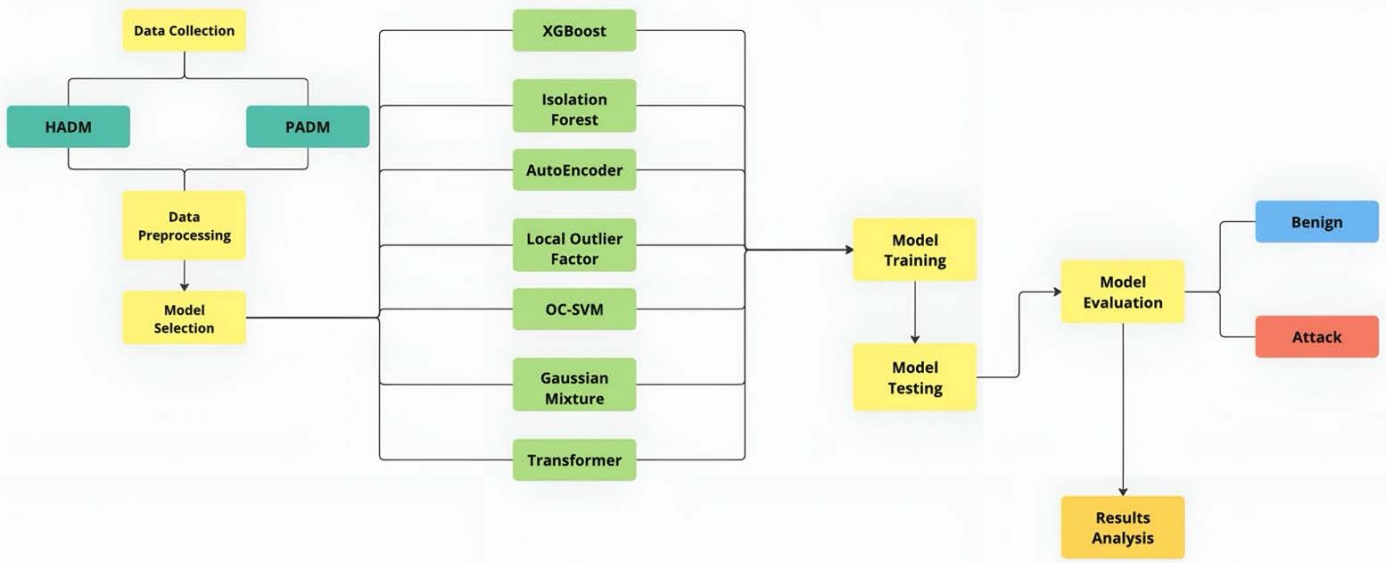



Fig. 1. Flowchart illustrating the workflow.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F\text{-}1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

## B. Evaluation

While conducting our research, evaluation focused on two key models: HPC Anomaly Detection Model (HADM) and Power Anomaly Detection Model (PADM). HADM utilizes high-dimensional HPC and kernel events such as CPU instruction cycles, memory access patterns, interrupt frequencies, etc. Fig. 2. Showcases the Performance Metrics for both the models. The results show that every classifier performed considerably well. In terms of accuracy, almost every technique had above 90%, with F1-score readings indicating differences among the techniques. Models like Isolation Forest and OC-SVM showed moderate performance, which can possibly be boosted even more by investigating additional hyperparameter arrangements or feature engineering tactics. Meanwhile, LOF underperformed substantially, limiting its usage, as it is undoubtedly sensitive to the arrangement and density of the multifaceted HPC-Kernel Events input.

The sustained functioning of HADM was made attainable in a major way by the richness of the data input, containing state shifts and system kernel event logs. These features enable models to disentangle subtle adversarial patterns from normal operational noise, particularly when leveraging architectures like the Transformer, which exploit attention mechanisms to correlate the events across hardware subsystems.

In the case of PADM, which uses only two input features (current and power) derived from the power consumption dataset, the evaluation revealed a different performance trend. As a result of this simple input space, most of the techniques considerably lagged behind HADM, emphasizing the challenges of detecting anomalies with limited feature sets. For instance, a stealthy load-altering attack which subtly manipulates charging current to destabilize grid frequency may evade detection by these models due to the absence of contextual features (e.g., voltage phase angles or harmonic distortions). The Transformer and XGBoost excelled in terms of F1-score and accuracy, showcasing their ability to handle low-dimensional data effectively.

However, models such as Isolation Forest and LOF exhibited poor performance, which can be attributed as an outcome of their dependence on distance based measures which have trouble understanding the power consumption data's sparse and unclear feature space. These results could be enhanced with further model optimization, refined hyperparameter tuning, and the inclusion of other inferred features like shunt or bus voltage to give anomaly detection additional context. Moreover, exploring mixed models or ensemble approaches that merge the benefits of several methods might boost PADM's potential of recognizing irregularities relating to power.

## C. Results

Our study evaluates the Electric Vehicle Supply Equipment (EVSE) security against anomalies by utilizing various machine learning models applied to HPC and power consumption datasets. The Transformer model consistently outperformed other approaches, achieving the highest accuracy and F1-scores across both HADM and PADM. XGBoost and GMM also exhibited strong performance, while models like Isolation Forest and LOF struggled in detecting anomalies effectively.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:36 UTC from IEEE Xplore. Restrictions apply.




Fig. 2. Performance Metrics for HADM & PADM.

The analysis of key performance metrics and visual illustrations highlights the scope of these advanced models in anomaly detection for EVSE systems. These results underline the substantial promise for enhancing EVSE security with machine learning and addressing challenges associated with ever-changing cyber threats. Table 1 presents all the metrics obtained across various approaches.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:36 UTC from IEEE Xplore. Restrictions apply.

TABLE I. PERFORMANCE EVALUATION OF DIFFERENT PROPOSED APPROACHES

| Model | Metrics | Isolation Forest | AutoEncoder | Transformer | LOF | GMM | OC-SVM | XGBoost |
|---|---|---|---|---|---|---|---|---|
| HADM | Accuracy | 0.9554 | 0.9133 | 0.9903 | 0.8436 | 0.9887 | 0.9206 | 0.9862 |
| | Recall | 0.7151 | 0.3636 | 0.9939 | 0.0983 | 0.9769 | 0.8497 | 0.9697 |
| | Precision | 0.9365 | 0.9677 | 0.9371 | 0.3148 | 0.9441 | 0.6712 | 0.9302 |
| | F1-score | 0.8109 | 0.5286 | 0.9648 | 0.1498 | 0.9602 | 0.7500 | 0.9496 |
| PADM | Accuracy | 0.1385 | 0.3702 | 0.8715 | 0.1570 | 0.4912 | 0.3598 | 0.9468 |
| | Recall | 0.0107 | 0.2846 | 0.9995 | 0.0455 | 0.4691 | 0.2719 | 0.9841 |
| | Precision | 0.823 | 0.9713 | 0.8717 | 0.7960 | 0.8970 | 0.9727 | 0.9560 |
| | F1-score | 0.0211 | 0.4402 | 0.9307 | 0.0859 | 0.6161 | 0.4250 | 0.9699 |

### *D. Discussion on Outcomes*

The findings reveal that machine learning classifiers like Transformer and XGBoost excels in detecting anomalies when compared to other approaches, showcasing their potential in securing EVSE systems. The Transformer achieved the highest accuracy and F1-score across both HADM and PADM, displaying their ability to capture complex patterns and relationships within high-dimensional and simplified data spaces. This exceptional capability stems from its self-attention mechanisms, which dynamically weigh the significance of sequential and contextual relationships within multivariate time-series data.

Similarly, XGBoost performed consistently well with metrics values close to the Transformer, indicating robust and adaptive nature of the technique. This performance is due to XGBoost's substantial ability to tolerate with missing values, prioritizing relevant attributes, which further highlights its adaptability to both raw, high-dimensional datasets and feature-engineered PADM environments. At the same time, the Transformer employs its attention mechanisms to recognize intricate correlations, making it suited for anomaly identification in dynamic contexts.

On the contrary, approaches like LOF and Isolation Forest showed limited applicability particularly in PADM, where their precision and recall dropped by 15–20% compared to supervised counterparts. Both models also suffered from high false-positive rates in dynamic contexts, such as detecting stealthy attacks mimicking legitimate user behavior—a critical shortcoming given the adversarial sophistication of modern EVSE cyber threats. A possible cause for this underperformance is likely due to their sensitivity to data distribution and inadequate generalization in both high-dimensional and minimal datasets. The exceptional performance of Transformer and XGBoost models allows them to be suitable for anomaly detection in EV charging systems. Their ability to identify anomalies consistently even in complex scenarios, making them ideal for cybersecurity applications. However, the results emphasize the need for further refinement of models like LOF and Isolation Forest through hyperparameter tuning and feature engineering to improve their reliability and applicability in anomaly detection tasks. Additionally, hybrid frameworks—such as ensemble models combining Transformer-derived embeddings with XGBoost classifiers could further bolster detection accuracy while mitigating individual algorithmic weaknesses.

## VI. CONCLUSION AND FUTURE SCOPE

In our evaluation we discovered that the Transformer and XGBoost approaches had the best performance metrics in both HADM and PADM, making them most effective in detecting anomalies in the dataset. However, models like LOF and Isolation Forest performed inadequately, emphasizing how crucial it is to select and fine-tune appropriate approaches for anomaly detection in EVSE systems. Our research reaffirms the involvement of machine learning in improving the security of EVSE infrastructure. Future work should focus on addressing the limitations of underperforming models by employing advanced optimization techniques, data preprocessing strategies and exploring other algorithms. Furthermore, integration with IoT networks, anomaly detection on real-time basis and scalable nature of the model could be explored for larger datasets.

Overall, this research focused on providing valuable insights about leveraging machine learning for anomaly detection, emphasizing its potential to enhance the safety of EVSE systems in the continuously changing electric vehicle ecosystem.

## REFERENCES

[1] J. Johnson et al., “Cybersecurity for electric vehicle charging infrastructure,” SAND2022-9315, 1877784, 708580, Jul. 2022.

[2] J. Johnson, “Securing vehicle charging infrastructure,” SAND--2020-11971R, 1706221, 691697, Nov. 2020.

[3] R. Varriale, R. Crawford, and M. Jaynes, “Risks of electric vehicle supply equipment integration within building energy management system environments: A look at remote attack surface and implications,” in National Cyber Summit (NCS) Research Track 2021, vol. 310, K.-K. R. Choo, T. Morris, G. Peterson, and E. Imsand, Eds., in Lecture Notes in Networks and Systems, vol. 310, Cham: Springer International Publishing, 2022, pp. 163–173.

[4] R. Gottumukkala, R. Merchant, A. Tauzin, K. Leon, A. Roche, and P. Darby, “Cyber-physical system security of vehicle charging stations,” in 2019 IEEE Green Technologies Conference(GreenTech), Lafayette, LA, USA: IEEE, Apr. 2019, pp. 1–5.

[5] D. Elmo, G. Fragkos, J. Johnson, K. Rohde, S. Salinas, and J. Zhang, “Disrupting EV Charging sessions and gaining remote code execution



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:36 UTC from IEEE Xplore. Restrictions apply.

with DoS, MITM, and Code Injection Exploits using OCPP 1.6," in 2023 Resilience Week (RWS), National Harbor, MD, USA: IEEE, Nov. 2023, pp. 1–8.

[6] D. Ronanki and H. Karneddi, "Electric vehicle charging infrastructure: Review, cyber security considerations, potential impacts, countermeasures, and future trends," IEEE J. Emerg. Sel. Topics Power Electron., vol. 12, no. 1, pp. 242–256, Feb. 2024.

[7] H. S. Mavikumbure et al., "Cy-Phy ADS: Cyber physical anomaly detection framework for EV charging systems," IEEE Trans. Transp. Electrific., pp. 1–1, 2024.

[8] D. Kern, C. Krauß, and M. Hollick, "Detection of anomalies in electric vehicle charging sessions," in Annual Computer Security Applications Conference, Austin TX USA: ACM, Dec. 2023, pp. 298–309.

[9] H. Tanyıldız, C. Batur Şahin, and Ö. Batur Dinler, "Enhancing cybersecurity through GAN-augmented and hybrid feature selection machine learning models: A case study on EVSE data," NATURENGS MTU Journal of Engineering and Natural Sciences Malatya Turgut Ozal University, Jun. 2024.

[10] H. S. Mavikumbure et al., "Physical anomaly detection in EV charging stations: Physics-based vs ResNet AE," in 2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE), Helsinki, Finland: IEEE, Jun. 2023, pp. 1–7.

[11] Y. Kim, S. Hakak, and A. Ghorbani, "DDoS attack dataset (CICEV2023) against EV authentication in charging infrastructure," in 2023 20th Annual International Conference on Privacy, Security and Trust (PST), Copenhagen, Denmark: IEEE, Aug. 2023, pp. 1–9.

[12] H. Man Duc Trong, D. Trong Le, A. Pouran Ben Veyseh, T. Nguyen, and T. H. Nguyen, "Introducing a new dataset for event detection in cybersecurity texts," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online: Association for Computational Linguistics, 2020, pp. 5381–5390.

[13] V. Mishra and S. Kadam, "A systematic review on security mechanism of electric vehicles," in Intelligent Systems Design and Applications, vol. 717, A. Abraham, S. Pllana, G. Casalino, K. Ma, and A. Bajaj, Eds., in Lecture Notes in Networks and Systems, vol. 717, Cham: Springer Nature Switzerland, 2023, pp. 587–598.

[14] H. He et al., "EVBattery: A large-scale electric vehicle dataset for battery health and capacity estimation," 2022, arXiv:2201.12358v3.

[15] C. Alcaraz, J. Cumplido, and A. Triviño, "OCPP in the spotlight: Threats and countermeasures for electric vehicle charging infrastructures 4.0," Int. J. Inf. Secur., vol. 22, no. 5, pp. 1395–1421, Oct. 2023.

[16] J. E. Rubio, C. Alcaraz, and J. Lopez, "Addressing security in OCPP: protection against man-in-the-middle attacks," in 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), Paris: IEEE, Feb. 2018, pp. 1–5.

[17] N. Mohamed, S. K. Almazrouei, A. Oubelaid, M. Bajaj, F. Jurado, and S. Kamel, "Artificial Intelligence (AI) and Machine Learning (ML)-based information security in electric vehicles: A review," in 2023 5th Global Power, Energy and Communication Conference (GPECOM), Nevsehir, Turkiye: IEEE, Jun. 2023, pp. 108–113.

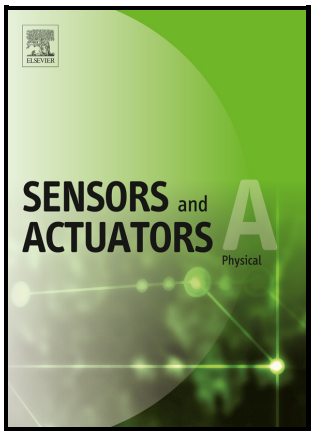


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:36 UTC from IEEE Xplore. Restrictions apply.

Journal Pre-proof

# Mitigating Thermal Effects in MEMS Gyroscopes: A Novel Substrate Integration Approach

Shaveta, R.K. Bhan, S.S. Lamba, Vinita Gond, Rishu Chaujar




PII: S0924-4247(25)00539-4

DOI: https://doi.org/10.1016/j.sna.2025.116733

Reference: SNA116733

To appear in: *Sensors and Actuators: A. Physical*

Received date: 18 March 2025
Revised date: 11 May 2025
Accepted date: 24 May 2025

Please cite this article as: Shaveta, R.K. Bhan, S.S. Lamba, Vinita Gond and Rishu Chaujar, Mitigating Thermal Effects in MEMS Gyroscopes: A Novel Substrate Integration Approach, *Sensors and Actuators: A. Physical,* (2025) doi:https://doi.org/10.1016/j.sna.2025.116733

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

© 2025 Published by Elsevier B.V.

Journal Pre-proof

# Mitigating Thermal Effects in MEMS Gyroscopes: A Novel Substrate Integration Approach

**Shaveta[1,3], R. K. Bhan[2], S.S. Lamba[1], Vinita Gond[1] and Rishu Chaujar[3]**

[1]Solid State Physics Laboratory, Lucknow Road, Timarpur, Delhi-110054, India
[2]Institute of Defence Scientists and Technologists, Majumdar Marg, Timarpur, Delhi-110054, India
[3]Delhi Technological University, Bawana Road, Rohini, Delhi 110042, India

**Abstract**

This research introduces a novel approach to improve the thermal performance of MEMS gyroscopes significantly. By strategically integrating a thermally optimized substrate at the packaging level, we establish a controlled temperature offset between the sensor and the package base. This innovative technique effectively reduces heat transfer to the sensor, mitigating the detrimental effects of temperature-induced performance degradation. Furthermore, our meticulously engineered gyroscope design minimizes heat flow into the sensor's sensitive areas, such as the sense mass. This synergistic approach leads to substantial enhancements in thermal stability across key performance parameters, including sense deflection, sensitivity, noise, and bandwidth. Through a thermally optimized substrate and engineered design, a substantial device temperature reduction of approximately 25°C was achieved. This enhanced thermal management, validated by transient thermal analysis, along with favourable outcomes from stress analysis indicating improved structural integrity, collectively led to significantly decreased temperature-dependent variations in sense deflection and a reduction in scale factor temperature sensitivity. Notably, sense deflection was reduced from 2.7 $\times 10^{-6}$ μm/$^0$C to 8.6$\times 10^{-8}$ μm/$^0$C, while scale factor temperature sensitivity was reduced from 232 ppm/°C to 6 ppm/°C. Additionally, we observed notable improvements in noise reduction and bandwidth stabilization. A comparative analysis with existing literature validates the superior thermal performance achieved through our proposed method. The etching experiment to study the feasibility of the Thermal Substrate is also reported. The ability to effectively manage thermal effects in MEMS gyroscopes has profound implications for enhancing their reliability and accuracy in various applications.

**Word count**: 6,165 words, excluding references.

**Keywords:** MEMS gyroscope, Thermal Substrate, Thermal resistance, Taper angle, Temperature coefficient of frequency, Scale factor temperature coefficient.

**Author Contributions:** Shaveta: Analysis, Conceptualisation, Writing - Original Draft Preparation, R.K. Bhan: Editing, Guidance, Final Proofreading, S.S. Lamba: Experimental Work Planning, Vinita Gond: Experimental Work Dr. Rishu Chaujar: Guidance and Supervision.
Corresponding Author: Dr. Rishu Chaujar, chaujar.rishu@dtu.ac.in

## 1. INTRODUCTION

MEMS gyroscopes have become ubiquitous because of their small size, low power consumption, and high accuracy [1–4]. Their critical applications span diverse sectors, including automotive, aerospace, defence, industrial automation, health care, consumer electronics, and virtual and augmented reality [5–8]. The accurate output from the MEMS gyroscope ensures reliable performance of all the systems under varying environmental conditions (temperature, pressure, and humidity). Consequently, understanding and compensating for the behaviour of the device under different conditions is of paramount importance.

The primary focus of this study was to analyze the impact of temperature variations on key performance parameters. Numerous studies have focused on developing effective active temperature compensation and passive thermal isolation techniques to ensure reliable and accurate operation. Common approaches to temperature compensation include integrating external circuits to actively compensate for temperature-induced changes, employing sophisticated algorithms to process sensor data and correct for temperature effects, and adjusting the frequency of the drive and sense modes to mitigate temperature-related variations. Ref. [9] proposed an active temperature compensation method using external electronic hardware integrated with an ASIC and a gyroscope. Ref. [10] employed a multiparameter fusion compensation method with a weighted linear combination of resonant frequency, frequency split, primary mode amplitude, and quadrature error to obtain the stable scale factor temperature coefficient. In addition, Ref. [11] utilized a thermal deformation suppression chip to reduce the thermal out-of-plane deformation for thermal stabilization. Ref. [12] introduced a peripheral monitoring circuit for temperature compensation. Ref. [13] implemented frequency mismatch

control to reduce the effect of temperature on the gyroscope structure. Ref. [14] utilised a self-calibration method to control the scale factors by detecting non-linear components. Ref. [15] employed in-run scale-factor error compensation using algorithms for temperature compensation.

The other approaches to thermal isolation include utilising materials or structures to isolate the gyroscope from temperature fluctuations and enclosing the gyroscope in a temperature-controlled environment.

For example, Ref. [16] investigated porous silicon as an effective material for thermal isolation Ref. [17] employed a powder MEMS wafer-level process to create porous microstructures. In, Ref. [18], an oven-controlled gyroscope was used for thermal stabilization. Recently, Ref.[19] proposed a thermoelectric-cooler-controlled gyroscope design. These studies collectively demonstrate the diverse strategies available to effectively compensate for the temperature variations in MEMS gyroscopes and ensure their reliable performance in various applications.

The difficulty with the above-mentioned approaches is that they may not completely eliminate the temperature dependence of the performance parameters, which may result in residual misalignments, instability, drift, and high-power consumption, particularly at high temperatures.

Recently, we proposed a novel thermal-design approach for gyroscope sensors [20]. Here, we present a novel approach to manage heat transfer at the packaging level to mitigate temperature-induced performance degradation, offering a complementary strategy to both compensation and isolation; hence, one may not need the complex hardware or software approaches discussed in the literature for temperature compensation. The proposed approach can be applied to most MEMS sensors and gyroscopes.

The paper covers the thermal design and analysis of the new advanced thermal substrate (TS), which is verified through finite element analysis, and discusses the integration of the proposed MEMS gyroscope with the thermal substrate. This also analyzes the sensitivity, bandwidth, and noise trends across varying temperatures using thermal substrates, followed by validation with the existing literature.

## 2. Novel Thermal Design Approach

Heat transfer to the device occurs mainly through three mechanisms: conduction, convection, and radiation, with conduction being the most dominant [21]. The heat from the environment and associated electronics is transmitted through the back surface of the device through the package base, which serves as the mounting surface during packaging. Subsequently, this heat was applied to the proof mass of the sensor via the spring beams of the device. Two strategies were proposed to mitigate heat transmission to the device:

a) The device is isolated from the package using an independent Thermal Substrate (TS) that limits heat transmission, thereby reducing the temperature of the sensor.
b) Reducing the heat transmitted to the sensor through innovative design modifications of the sensor beams serving as primary heat conduits for heat flow.

The results and analysis demonstrated that these approaches significantly improved the thermal performance of the device at the design level. The device consisted of a MEMS gyroscope integrated with a TS mounted on the package. The TS design was optimized through finite element modeling using the SOLIDWORKS 2024 software, enabling an in-depth analysis of the temperature behaviour within the TS. Second, the novel and improved thermal beam design of the gyroscope and key performance parameters, including sensitivity, noise, and bandwidth, are evaluated under varying temperature conditions using the Coventorware MEMS+ software. This study culminates in developing a thermally stable MEMS gyroscope, achieved by applying the optimized TS.

### *2.1 Design of Thermal Substrate (TS)*

The primary heat transfer mechanism between the MEMS device and the package base is conduction due to their direct thermal contact. The package base attains a steady temperature influenced by the ambient environment and the heated drive electronics. To enhance the sensor's temperature-dependent performance, we propose a novel approach: utilizing an independently optimized thermal substrate (TS). The TS acts as a thermal barrier, effectively reducing heat flow from the package base to the sensor. This thermal isolation improves the sensor's temperature stability and overall performance. Additionally, the TS serves as an electrode, eliminating the need for additional adhesives and reducing potential errors.

Various TS designs are possible, including multiple legs and circular or rectangular shapes. However, the optimal TS design should prioritize the ease of fabrication, mechanical stability, and compatibility with the bonding process. By carefully selecting the TS design, the thermal performance and reliability of MEMS sensors can be further enhanced. Four distinct TS, namely- Planar Silicon with oxide (Type-I), 2-pillar Silicon with oxide (Type-II), 4-pillar Silicon with oxide (Type-III), and 4-pillar tapered silicon with oxide (Type-IV), are considered here for comparison based on the fabrication compatibility and mechanical strength. Rectangular and square columns are standard microfabrication features that can be reliably produced with maximum utilisation of Silicon real estate post-dicing of the sensor chips, resulting in higher yields. Figure 1 (a, b, c, d) presents isometric views illustrating four distinct TS. Each figure consists of a TS mounted on a package base (mostly ceramic or alumina), with a uniform

oxide layer at the bottom. The Type I TS is a planar silicon structure. The Type II TS consists of two rectangular pillars etched on the backside of the silicon. Type III TS is similar to type II TS but with four square pillars instead of two. Type IV TS has four pillars, and the base is tapered along the thickness by an angle. Tapered TS Type IV can be realized using either wet or dry etching. The anisotropic wet etching of (100) silicon resulted in walls with a $54.75^0$ sloping angle [22]. Reactive-ion etching results in different taper angles using different etchant gas ratios and RF bias powers [23]. It is assumed that the top surface length $L_{TS}$ and width $W_{TS}$ are 2.2 mm and substrate thickness $T_{TS}$ is 500 µm. These dimensions are the same for all TS.

Further optimization of the structure involves the etching process of the four support columns from the backside, which yielded only a marginal improvement of approximately 1°C in thermal isolation primarily because of large thermal contact area yielding maximum heat transfer. Considering the limited thermal improvement with backside etching optimization and the increased fabrication complexity and mounting difficulties associated with non-standard support shapes, we focused our optimization efforts on the distribution and dimensions of the rectangular support columns, which offer a good tradeoff between limited heat transfer and mechanical support and practical implementation.

Steady-state thermal simulations of four distinct Target Structures (TS) were performed utilizing the SOLIDWORKS Simulation software, as depicted in Figure 1. Boundary conditions were established as follows: The package base temperature was fixed, the ambient temperature was specified, A uniform heat flux was applied to the model, and contact conditions between the TS and the package base were defined to model heat transfer through conduction accurately. The model was subsequently meshed, and the thermal gradient across the thickness of the TS along the edges was analyzed.

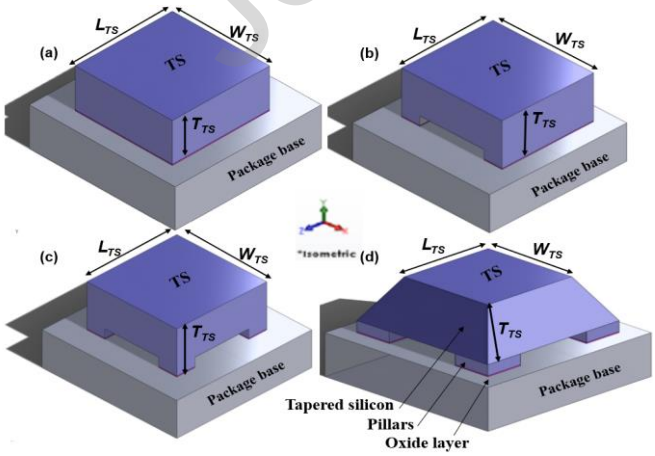



**Figure 1.** Three-dimensional views of various thermal substrates (TS). (a) Planar Silicon with oxide (Type-I). (b) 2-pillar Silicon with oxide (Type-II). (c) 4-pillar Silicon with oxide (Type-III). (d) 4-pillar tapered silicon with oxide (Type-IV).

It was assumed that the package base was at a temperature of $100^0$C. Any temperature fluctuations in the base induced changes in the MEMS gyroscope, both across its thickness and along its length. We estimated the amount of heat transmitted across the TS along its thickness.

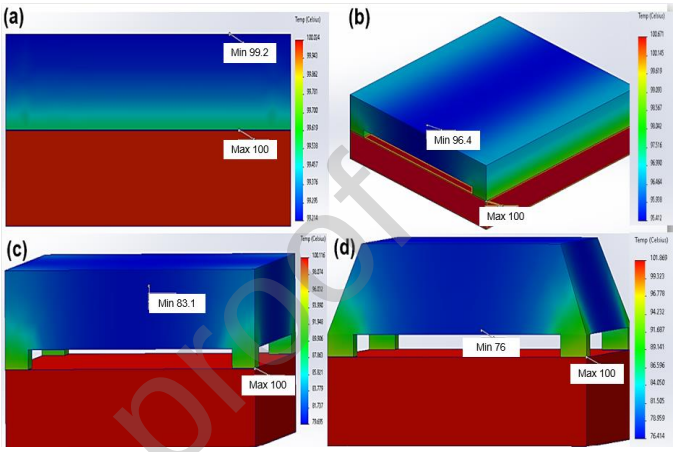



**Figure 2.** Temperature distribution profile for different thermal substrates (a) Planar Silicon with oxide (Type-I). (b) 2-pillar Silicon with oxide (Type-II). (c) 4-pillar Silicon with oxide (Type-III). (d) 4-pillar tapered silicon with oxide (Type-IV).

Figure 2 illustrates the simulated temperature distribution within four different TS. The colour gradients represent temperature variations, with blue indicating cooler regions and red indicating hotter regions. The structures are subjected to a uniform heat flux at the base, and the temperature distribution along the thickness of the TS is analyzed. The planar structure in Figure 2a exhibits a relatively uniform temperature distribution, with heat spreading evenly throughout the device. The two-pillar structure in Figure 2b shows a higher temperature concentration near the base, indicating that heat is being conducted primarily through the pillars. The four-pillar structure in Figure 2c also exhibits a higher temperature concentration near the base. The tapered structure in Figure 2d demonstrates a significant reduction in temperature near the top of the device. In Figure 2, the maximum temperature value on the scale is 100 °C. The minimum temperature values for Figures 2a to 2d, as indicated by the temperature distribution scale, are 99.2 °C, 96.4 °C, 83.1 °C, and 76 °C, respectively.

The simulation results demonstrate that the geometry of the MEMS structure significantly influences the temperature distribution. Tapered structures, such as Type-IV, can effectively reduce the temperature at the top of the device, which is crucial for applications where temperature sensitivity is a concern. Understanding these thermal characteristics is

essential for designing MEMS devices that can operate reliably under various conditions.

### *2.2 Thermal Resistance*

Understanding the thermal resistance characteristics of MEMS structures is essential for designing devices that can effectively manage heat and maintain desired operating conditions. The data presented in Table 1 highlight the importance of considering all three heat transfer mechanisms (conduction, convection, and radiation) when analyzing the thermal performance of MEMS devices. The simplified equations for thermal resistance by conduction, convection, and radiation are given by [21]

$$R_{t,cond} = \frac{T_{TS}}{kA_{TS}} \quad ,R_{t,conv} = \frac{1}{hA_{TS}} \quad ,R_{t,rad} = \frac{1}{h_r A_{TS}} \qquad (1)$$

$L_{TS}$, $W_{TS}$, and $T_{TS}$ are the length, width, and thickness of the top surface of TS, respectively. $k$ is the thermal conductivity of the material used for TS, h is the convection heat transfer coefficient, and $h_r = \varepsilon\sigma(T_{TSs} + T_{TS0})(T_{TSs}^2 + T_{TS0}^2)$ is the radiation heat transfer coefficient, ε is the emissivity, σ is the Stefan–Boltzmann constant, $T_{TSs}$ is the temperature on the surface of the thermal substrate, $T_{TSo}$ is ambient temperature.

| TS Type | $R_{cond}$ (K/W) | $R_{conv}$ (K/W) | $R_{rad}$ (K/W) |
|---|---|---|---|
| Planar Silicon with oxide (Type-I). | 0.8 | $2.0\times10^6$ | $1.2\times10^6$ |
| 2-pillar Silicon with oxide (Type-II). | 1.5 | $1.7\times10^6$ | $1.0\times10^6$ |
| 4-pillar Silicon with oxide (Type-III). | 5.7 | $2.0\times10^6$ | $1.1\times10^6$ |
| Tapered with oxide (Type-IV). | 33 | $31\times10^6$ | $18\times10^6$ |

**Table 1** Thermal Resistance for Thermal Substrates (Type-I TS to Type-IV TS)

Table 1 shows the consolidated thermal resistances of all four types of TS (Type I to Type IV) after putting TS dimensions and heat transfer coefficients used in the equations. For all four types of MEMS structures, the conductive thermal resistance ($R_{cond}$) was significantly lower than the convective ($R_{conv}$) and radiative ($R_{rad}$) resistances. This indicates that heat transfer occurs primarily through conduction, whereas convection and radiation play less significant roles. The type of MEMS structure (types I to IV) significantly affects the thermal resistance. The type-IV structures exhibited the highest overall thermal resistance, suggesting better thermal isolation from the surroundings.

### *2.3 Thermal Substrate (TS) temperature variation*

The simulation results of the Thermal Substrate (TS) temperature variation across the thickness and lateral dimensions are discussed in Figures 3a and 3b, respectively. For reference, the temperature distribution trends at a typical base temperature of 100 $^0$C were depicted. The base thickness varied from 0 to 500 µm, while the lateral dimension varied by 1.1 mm left and right from the middle. The analysis also included conventional silicon to demonstrate the temperature drop at a particular thickness for reference. Furthermore, introducing a thermally grown oxide layer of 1 µm results in a noticeable temperature decrease across the thickness and along the length as the base thickness increases.

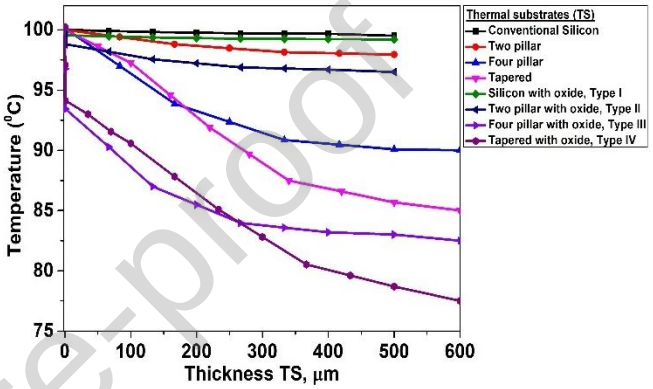



**Figure 3a.** Plot shows the temperature variation with and without the oxide intermediate layer, along with base modifications across the thickness, is also included.

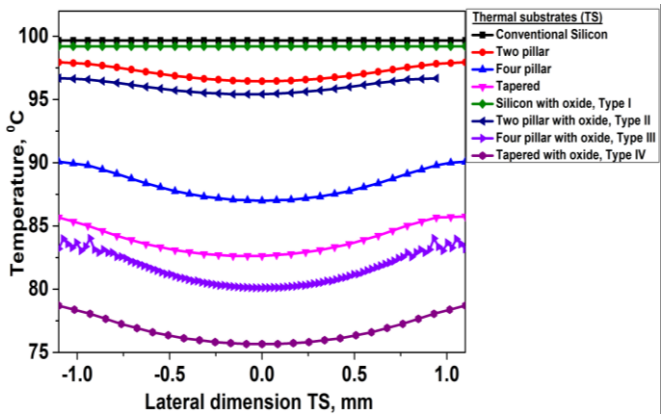



**Figure 3b.** Plot of temperature variation without and with oxide layer with base modifications along the length of the base.

Figure 3a illustrates the temperature distribution profile among different TS configurations for the thermal management of MEMS devices using SolidWorks software. Here, we mainly consider the contribution of heat conduction from the package's base to the thermal substrate (TS) because the actual sensor will be subsequently mounted on this base, as will be discussed later. Silicon with oxide (Type I) shows a uniform higher temperature near the base and a lower temperature at the top. The presence of an oxide layer offers some thermal isolation; however, the temperature variation is not highly pronounced. Further, by increasing the number of

Journal Pre-proof

pillars from two to four, the temperature distribution showed a significant reduction from the base to the TS. Tapered TS with oxide (type-IV) provided the most substantial thermal isolation. The temperature distribution profile shows a steep gradient, indicating that the tapered shape combined with the oxide layer effectively minimizes the heat transfer, making it the most effective design for minimizing the temperature-induced performance variation in MEMS gyroscopes. Figure 3a shows that incorporating pillars and tapering the base improves thermal isolation by reducing the temperature transfer from the package base. Among all TS, for the 4-pillar tapered silicon with an oxide (Type IV), the temperature is reduced across the thickness by approximately 25 $^0$C. Hence, this is the most effective method to reduce the temperature as the base thickness increases.

Figure 3b shows the temperature variations along the lateral dimensions or length of the base for different thermal substrate (TS) configurations. Regarding the lateral temperature distributions, all TS configurations exhibited the lowest temperature profile in the center. Introducing an oxide layer (Types I, II, III, and IV) consistently led to lower temperatures than the corresponding substrate without an oxide layer. This reinforces the insulating effect of the oxide layer on heat dissipation. A maximum temperature drop of approximately 3 $^0$C is seen at the center of the structure compared to the outer edge of the tapered structure with four pillars. The different base modifications (two-pillar, four-pillar, and tapered) resulted in varied temperature profiles, but the decreasing trend remained. The type IV configuration exhibits lower temperatures, suggesting potentially better heat dissipation. The conventional silicon substrate showed a relatively stable lateral temperature profile, indicating a consistent heat dissipation behaviour. These results suggest that careful design of the base, including its length and shape, can contribute to improved thermal management. The results indicated that the tapered TS yielded the best thermal management results. We intend to investigate the dependence of the taper angle on the temperature drop in detail for a type IV TS.

Transient thermal analysis was conducted to evaluate the temporal temperature response of the various Thermal Substrate (TS) designs, a critical factor in their performance. As depicted in Figure 4, the baseline Silicon and Type I structures exhibit near-instantaneous temperature increases to approximately 100°C and 99°C, respectively, indicating minimal thermal isolation. While the Type II (2-pillar silicon with oxide) design demonstrates marginally improved isolation, it rapidly achieves its steady-state temperature of ~96 °C (within ~60 µs, as shown in the inset for the initial 100 µs). In contrast, the Type III (4-pillar) and Type IV (4-pillar tapered) structures display substantially enhanced thermal isolation, requiring significantly longer durations to reach their considerably lower steady-state temperatures of approximately ~83°C and ~76°C, respectively. This behaviour confirms that the increased thermal resistance engineered into the Type III and, particularly, the Type IV designs effectively lower the final equilibrium temperature and extends the time required to reach it.

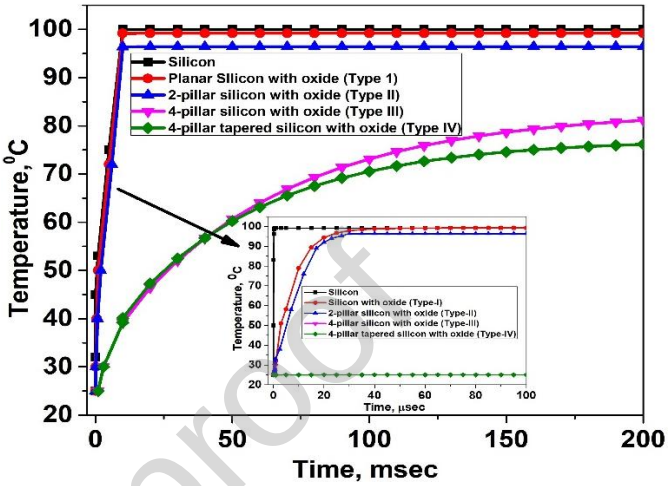



**Figure 4.** Transient response of thermal substrates.

*2.4 Stress analysis*

A MEMS device's stress tolerance before significant performance degradation happens varies based on several factors. These include the specific material's mechanical properties, the structural design influencing stress distribution, and the types of stress experienced (residual, thermal, or mechanical from external loads). The intended application, such as gyroscopes, also dictates acceptable stress levels. Minimizing internal and external stresses is crucial for ensuring long-term reliability and stable performance, as high or non-uniform stresses can lead to detrimental deformation, fracture, fatigue, and undesirable shifts in critical device parameters like resonant frequencies.

To ensure the structural integrity and prevent permanent deformation or failure of Thermal Substrate (TS) devices, the induced stress within any part should ideally remain well below the ultimate or fracture strength of the constituent materials. In this study, the von Mises stress resulting from Coefficient of Thermal Expansion (CTE) mismatch between Gyroscope sensor and Thermal Substrate (TS) was simulated for four distinct TS designs across a temperature range of -40°C to 125°C.

Figure 5 illustrates thermal stress contours in the Type-IV TS, with stress maximally concentrated near Si/$SiO_2$ interfaces and geometric discontinuities. The arrows depicted at the bottom of the structure in each sub-figure indicate the fixed boundary conditions applied to the substrate during the simulation. This arises from CTE mismatch, Si (2.6 ppm/$^0$C)> $SiO_2$(0.5 ppm/$^0$C), leading to compressive stress in $SiO_2$ after fabrication. Cooling below room temperature typically increases this compression and overall stress. Conversely,

heating, while potentially reducing $SiO_2$ compression, can still elevate the overall von Mises stress in the composite structure due to complex interactions, as observed. The arrows at the bottom show the fixed boundary applied to the substrate.

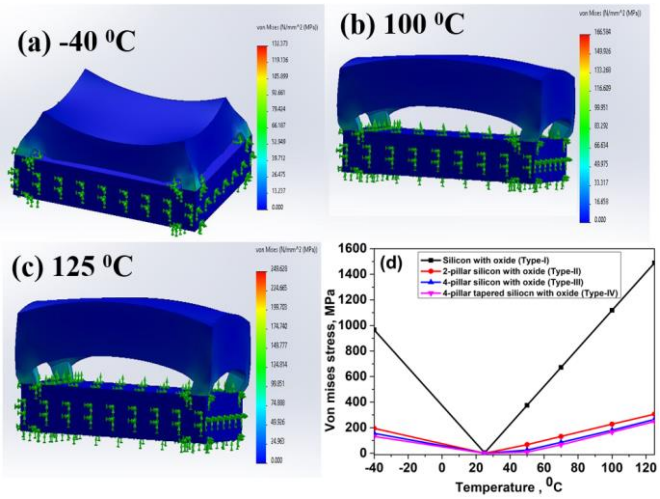



**Figure 5.** Thermal stress contours in the Type-IV TS at temperature (a) -40 $^0$C, (b) 100 $^0$C, (c)125 $^0$C, (d) Stress plot.

| Temperature ($^0$C) | Mises stress (MPa) | | | |
|---|---|---|---|---|
| | Planar Silicon with oxide (Type-I) | 2-pillar Silicon with oxide (Type-II) | 4-pillar Silicon with oxide (Type-III) | 4-pillar tapered silicon with oxide (Type-IV). |
| -40 | 965 | 195 | 156 | 132 |
| 25 | 4 | 0.48 | 0.45 | 0.43 |
| 100 | 1118 | 228 | 230 | 166 |
| 125 | 1489 | 306 | 255 | 249 |

**Table 2.** The consolidated results

**Table 2** Shows the Stress values of Gyroscope and Thermal substrate (TS) due to Coefficient of Thermal Expansion (CTE) mismatch.

Furthermore, as the Temperature is either increased or decreased from the Room Temperature, Maximum stress increases approximately linearly. (Figure 5d). In short, it is concluded that stress induced due to incorporation of TS is well under tolerable level.

*2.5 Taper angle dependence*

The change in temperature through the thermal substrate is given by equation [21].

$$\Delta T = \frac{T_{TS}}{kA_{TS}cos\theta} Q \qquad (2)$$

where θ is the taper angle, and Q is the heat flow rate. Figure 6 illustrates the simulation results of the taper angle of the Type IV TS structure and the resulting temperature. As the taper angle of the Type IV TS increased, the temperature consistently decreased. The relationship between the taper angle and temperature appears to be non-linear; for smaller taper angles, the temperature drop is more gradual. Conversely, at larger taper angles, the temperature exhibits a steeper decrease. Tapering effectively reduces the heat conduction pathways, thus maintaining the cooling operating temperature for the MEMS gyroscope. From Equation (2), the temperature change is a trade-off between the TS taper angle and the TS base area. It may be stated here that there is a limit to the taper angle because as the taper angle is increased, the effective area of the sensor will be significantly reduced. However, the bottom of the TS should be compatible with the package where we want to put the sensor, and the top should be approximately 10 % larger than the sensor footprint.

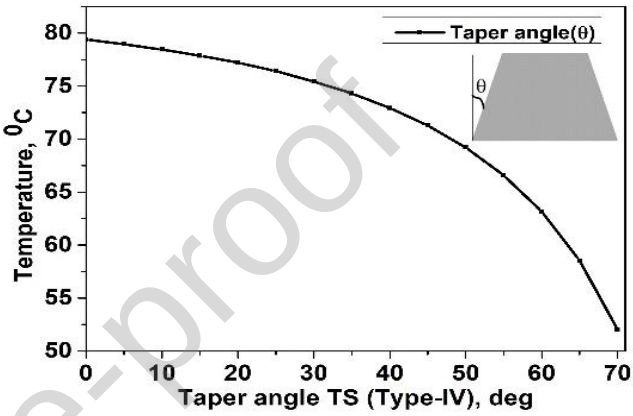



**Figure 6.** Plot of temperature dependence on taper angle for the Type IV TS structure.

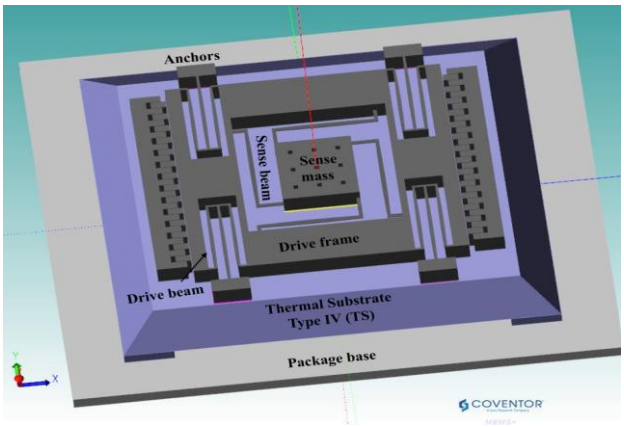
## 3. Improved Thermal Design of Gyroscope

Figure 7 presents a 3D model and side view of a MEMS gyroscope integrated into a Type IV TS in the Coventorware MEMS+ software. The central component is a proof or sense mass of the MEMS Gyroscope, responsible for sensing angular velocity. The drive beam anchors are away from the corners to take advantage of a slightly lower temperature. The thermal Substrate (TS), a tapered structure with four pillars, is designed to dissipate heat, as discussed earlier. Because the material of the TS is silicon, it also acts as the bottom plate to sense the change in the output in response to the input. This prevents the use of additional adhesives to paste the sensor onto the insulator substrate, which has been explored in the literature [11][24][25]. It also avoids the misalignment errors caused by the use of epoxy. In fact, the TS is directly wafer-bonded with the central wafer, in which the sense mass is etched. The foundation on which the entire assembly is mounted is a package base from which heat flows into the sensor and raises the temperature of the sensor.

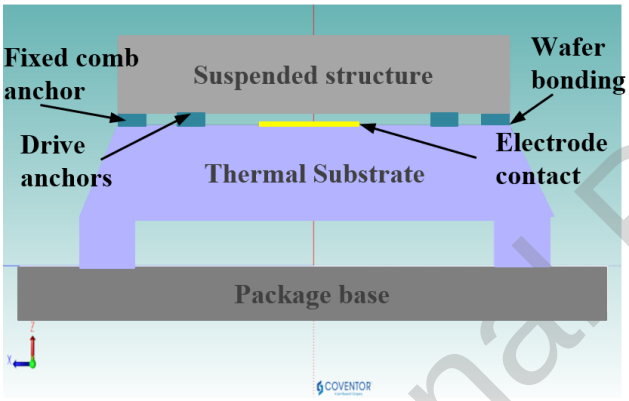
Regarding thermal management, the Type IV TS, which features a tapered design with pillars, will likely be engineered to enhance heat dissipation from the MEMS gyroscope. This is crucial for maintaining optimal performance and preventing device failure owing to overheating. Additionally, the TS and package base offer mechanical support to the gyroscope, shielding it from external shocks and vibrations. The

Journal Pre-proof

gyroscope integration with the TS is proposed here as a novel design approach to optimize thermal and mechanical performance. The subsequent discussion addresses how integration affects the gyroscope's sensitivity and other performance parameters.




(a)




(b)

**Figure 7.** MEMS Gyroscope mounted on a Type IV (tapered with four pillars) Thermal substrate (TS) (a) 3D view (b) Side view. The integrated hybrid structure is mounted on the package base.

### *3.1 Sense displacement with different thermal Substrate (TS) bases*

Previously, a new, improved thermal design of MEMS gyroscopes using modified tapered beams was introduced to reduce the heat flow from the beams to the sensing mass of the gyroscope. The revised gyroscope design featuring these tapered beams is shown in Figure 8. A comprehensive analysis was presented in [20]. Heat flow is restricted to the centre of the beam, which prevents temperature rise in the sense mass.

This discussion presents the overall improvement in thermal performance, emphasizing the combined impact of the Thermal Substrate (TS) and tapered beams for both the sense and drive directions. The TS effectively reduces the temperature transmitted from the environment and package base, while the tapered beams minimize the temperature dependence of key performance parameters such as sense displacement, sensitivity, noise, and bandwidth. The following section discusses these aspects. Conventional design and tapered beam design without TS and with TS are considered for further studies.

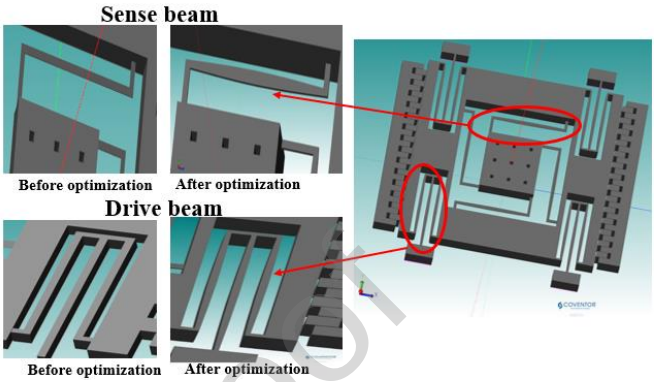



**Figure 8.** 3D figure of complete Gyroscope with taper beam design with Zoomed inset of the sense and drive beam.

Figure 9 illustrates the variation of sense displacement with temperature for different gyroscope configurations at a constant rotation rate of 1 rad/s. It is crucial to note that the sense mass is supported by beams from all four sides, resulting in a constrained motion where the beam's centre undergoes rotational motion without significant lateral displacement. This constraint leads to a non-linear dependence of sense displacement on temperature, even under constant rotation.

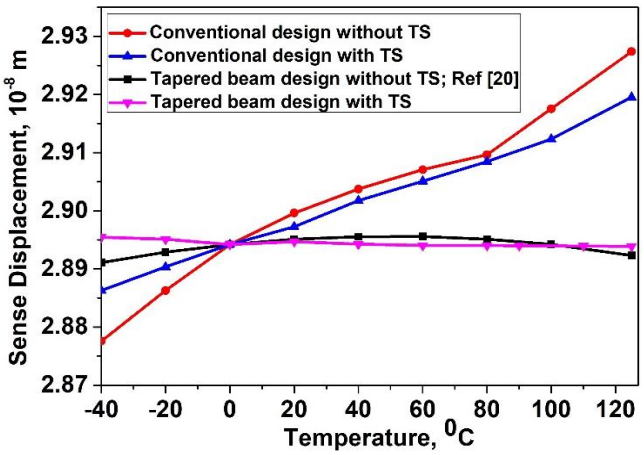



**Figure 9.** Sense displacement as a function of Temperature for various types of Gyroscopes like Conventional design without TS, Conventional design with TS, Tapered beam design without TS, and Tapered beam design with TS (Type IV).

As evident from Figure 9, a consistent trend of increasing sense displacement with rising temperature is observed across all gyroscope configurations. There is a positive correlation between temperature and sense displacement. At low temperatures, adding a base to the conventional gyroscope design (original with base) slightly increases the sensed displacement, whereas the trend is opposite at the upper end

Journal Pre-proof

of temperature. From the slope of conventional design without TS configuration, the change in sense displacement is $2.7\times10^{-6}$ µm/$^0$C. Introducing a TS (Type IV) further amplifies the sense displacement at lower temperatures and reverses this trend at the upper temperature. The crossover point is the ambient temperature of 0 deg (273 K), which is the default setting in the software. All structures are designed to have the same sense displacement at the default temperature value. The Tapered beam design with TS (Type IV) configuration, which likely incorporates additional modifications, exhibits an almost constant sense displacement of $8.6\times10^{-8}$ µm/$^0$C across the entire temperature range. This is highly desirable for the temperature-independent performance of the Gyroscope.

Furthermore, it suggests that one can tailor the deflection slope vs. temperature curve by using our proposed thermal management approaches at the design level. The gyroscope output (sense deflection) is significantly influenced by temperature. This sensitivity could potentially lead to inaccuracies in measurements if not properly accounted for. The tapered base configurations seem to reduce the sensitivity of the gyroscope to temperature. For performance considerations, the Type IV configuration, with an almost flat temperature dependence, might offer advantages in certain applications.

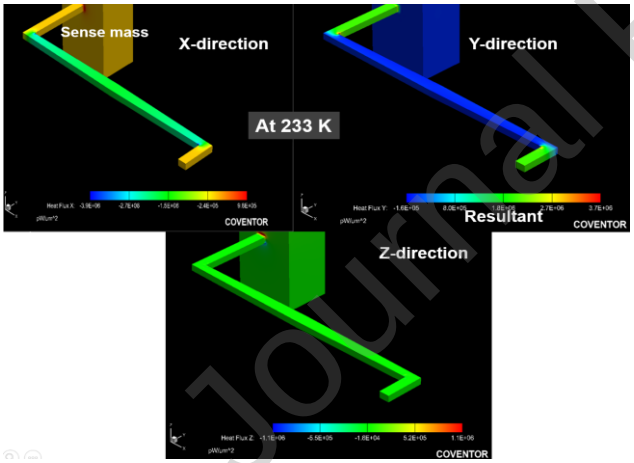



**Figure 10.** The heat flow in x, y, z directions at 233 K for the tapered beam.

Understanding the physical mechanisms responsible for the said improvements is discussed earlier [20]. Figure 10 show the heat flow in the tapered beam at 233 K. The heat flux in the x direction is controlled by tapering the beam from the centre for a longer beam length at any temperature. In this direction, the heat is more concentrated at the centre than at the ends of the beams. The heat flux in x direction reduced by 10 % as compared to the rectangular beam. For y and z directions the heat flow is almost uniformly distributed in the beam region. For y and z direction heat flux is reduced by ~3 % and 15 % respectively. So, it is concluded that with the tapered beam the heat flow in the x, y, z directions can be controlled. Additionally, it is also observed that the stress is also reduced from 38 MPa for the rectangular beam to 25 MPa for the tapered beam. Similar trend is observed at higher temperatures too.

In summary, the conical beam design enhances thermal stability and reduces sensing displacement errors through a multi-faceted approach. Its tailored geometry mitigates differential thermal expansion within the beam, minimizing temperature-induced displacement of the sensing mass. Furthermore, optimized stiffness decouples the sensing mode from thermal stresses, reducing measurement errors. Finally, its heat dissipation characteristics promote efficient heat transfer and uniform temperature distribution, minimizing localized thermal gradients and improving overall accuracy in varying thermal environments.

The implications of these findings for gyroscope design and operation must be carefully evaluated, especially in applications where temperature stability is critical.

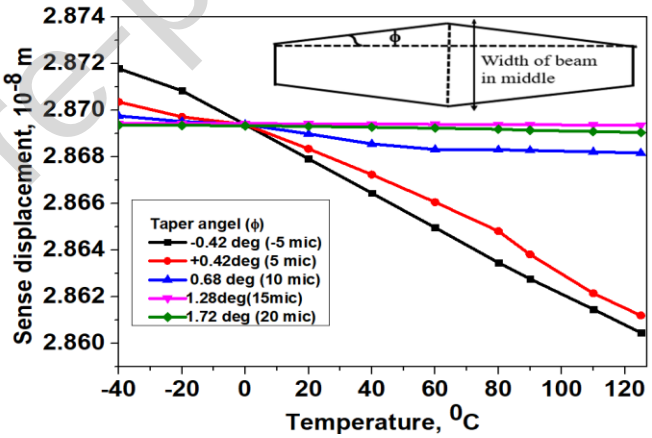


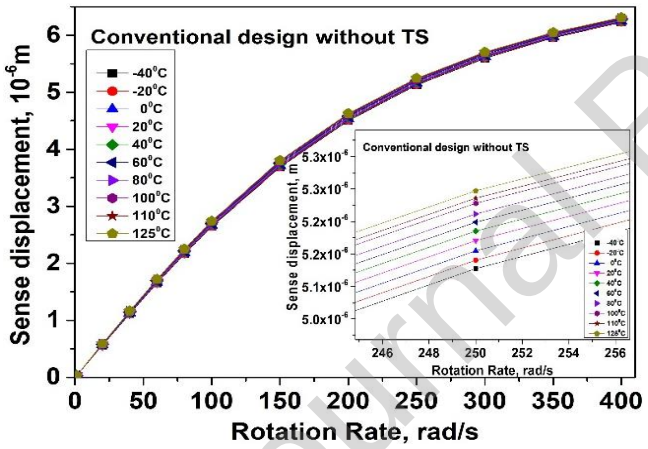

**Figure 11.** Plot of sense displacement versus temperature for different taper angles.

Figure 11 illustrates the effect of the taper angle (ϕ) on the temperature-dependent sensing displacement of the MEMS gyroscope. As shown in the subset, the beams were tapered, varying in width from 5–20 µm from the middle; accordingly, the angle will change. It was observed that the sense displacement consistently decreased with increasing temperature for all taper angles. Taper angles of -0.42° and 0.42° exhibited a more pronounced decrease in sense displacement compared to larger angles (0.68° to 1.72°). Taper angles between 0.68° and 1.72° show minimal or no discernible changes in sense displacement across the temperature range. These findings suggest that the chosen taper angle of 1.28° (15 µm middle portion width) provides a suitable balance between sensitivity and temperature stability, which is considered in the present study.

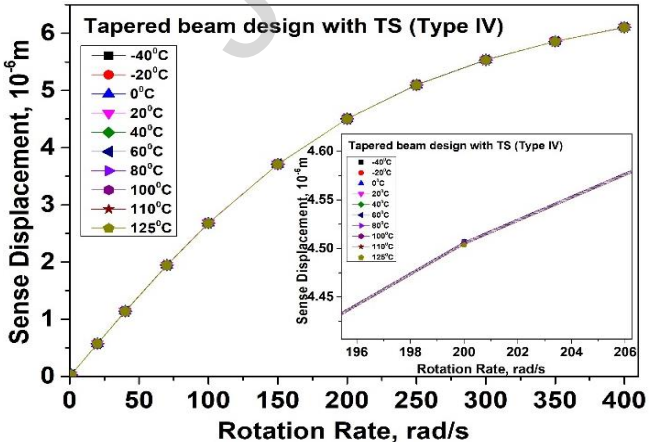
### *3.2 Scale factor temperature sensitivity*

Journal Pre-proof

The scale factor sensitivity is defined as the slope of the sense displacement versus rotation rate curve at different temperatures. Figures 12a and 12b show a plot for different temperatures from -40 to 125 $^0$C for the conventional design without TS and tapered beam design with TS (Type IV). Only these two cases were considered to show a clear comparison between the designs. For both plots, the data points for each temperature indicated a direct proportionality between the sense displacement and rotation rate. The relationship is nonlinear for each temperature at higher rotation rates. This suggests a positive correlation between the temperature and sense displacement. These results suggest that temperature-compensation mechanisms may be necessary to account for the varying sensitivities at different temperatures to ensure accurate measurements. The temperature dependence of the sense displacement is attributed to the thermal expansion coefficients or other material properties of the components. The slope for each curve was estimated using a piecewise linear approximation to estimate the scale factor sensitivity. The results up to 100 rad/s were considered first, as the plot was approximately linear up to this rotation. The scale factor sensitivity at each temperature was evaluated and plotted from the slopes of the plots in Figures 12a and 12b.




(a)




(b)

**Figure 12.** Plot of sense displacement versus rotation rate for (a) Conventional design without TS (b) Tapered beam design with TS (Type IV) at different temperatures ranging from -40 $^0$C to 125 $^0$C.

Figure 13 illustrates the effect of temperature on scale factor sensitivity (SFS) for two different MEMS gyroscope designs: a conventional design without TS and a new design using a tapered beam with Type IV TS. For conventional designs without TS, the SFS increases linearly from 2.64 × 10$^{-8}$ to 2.74 × 10$^{-8}$ as the temperature rises from -40 $^0$C to 125 $^0$C. However, the SFS remained constant for a tapered beam design with Type IV TS across the entire temperature range. The scale factor temperature coefficient (SFTC) is given by [26].

$$SFTC = \frac{k_{1(\Delta T=\Delta T)} - k_{1\,(\Delta T=0)}}{k_{1(\Delta T=0)}\,\Delta T} \qquad (3)$$

Where $k_1$ is the scale factor, and ΔT is the temperature difference. The Scale factor temperature coefficient (SFTC) for the conventional design without TS is 232 ppm/$^0$C, whereas it is significantly lower at 6 ppm/$^0$C for a tapered beam design with TS. The new tapered beam design demonstrated superior thermal stability compared with the conventional design. Additionally, at higher rotation rates, a three-point linear approximation reveals that the SFTC for the conventional design without TS is 77 ppm/$^0$C, whereas, for the tapered beam design, it is drastically reduced to 1.89 ppm/$^0$C.

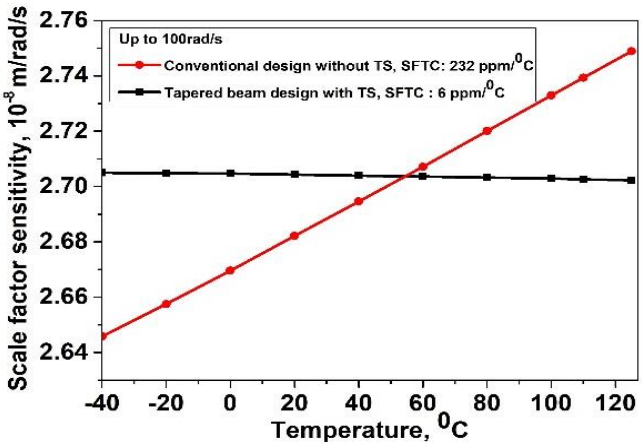



**Figure 13.** Plot of scale factor sensitivity for conventional design without TS and tapered beam design with TS Type IV up to 100 rad/s at temperatures ranging from -40 $^0$C to 125 $^0$C.

The non-uniform deformation, driven by differential thermal expansion, results in a net change in capacitance with temperature. Our simulations inherently capture this effect. To quantify this, our simulations indicate an absolute change in capacitance with temperature (ΔC/°C) of 2.8 × 10$^{-18}$ F/°C for the design without the TS, reduced significantly to 1.5 × 10$^{-20}$ F/°C for the design incorporating the TS. The resolution of the proposed readout circuit is 50 zF (50 × 10$^{-21}$ F). The change in capacitance is substantially larger than the readout resolution,

confirming that these temperature-induced capacitance changes are indeed resolvable. Here we want to add that the goal of designer is to make ΔC/°C much smaller so that capacitance readout output change due to temperature is much smaller than due to Gyroscope input angular rotation. Hence resolving ΔC/°C is not of primary concern. Only point of concern is that it should not dominate scale factor sensitivity ΔC/Ω where Ω is input angular rate. However, for further context, if we consider a nominal sense capacitance ($C_0$) of 100 fF ($1 \times 10^{-13}$ F), the relative change in capacitance per degree Celsius ($\Delta C/C_0$ per °C) is reduced from $2.8 \times 10^{-5}$ /°C without the TS to $1.5 \times 10^{-7}$ /°C with the TS.

Using a Thermal Substrate (TS) and tapered beam effectively mitigates the influence of temperature variations on the gyroscope's scale factor sensitivity (SFS) at all rotation rates, making it reliable for applications that require consistent performance across a wide temperature range.

### *3.3 Noise temperature dependence*

Thermomechanical noise is another key parameter that demonstrates the behaviour of a device. Thermomechanical (TM) noise is expressed by the following equation[27]:

$$TM\ Noise = \sqrt{\frac{k_B T\, \omega_s(T)\, BW(T)}{m_s \omega_d^2(T)\, A_{dm}^2\, Q_s(T)}} \qquad (4)$$

$k_B$ is the Boltzmann constant, $T$ is the absolute temperature, $BW$ is the bandwidth, $A_{dm}$ is the area of the drive mass, $\omega_s = \sqrt{k_s/m_s}$ is the sense mode frequency, $f_s = \frac{\omega_s}{2\pi}$ is the sense resonant frequency, $Q_s$ is the sense quality factor, and $m_s$ is the sense mass excited by the Coriolis force.

This equation clarifies the dependence of noise on temperature. The behaviour of each parameter with respect to temperature was studied to understand the effect of temperature. As mechanical noise has major impact on the structure, hence other sources of noise are neglected as far as the optimum temperature tolerant design of structure is concerned. The variation in the resultant thermomechanical noise with the temperature was studied.

Figure 14 shows multiple plots depicting the variation of important parameters in Equation 4 with temperature for one of the designs as an example. The parameters analyzed were drive frequency ($f_d$), sense frequency ($f_s$), drive displacement ($x_d$), quality factor ($Q_s$), and bandwidth (BW). The first two curves from the bottom show a steady decrease in the frequencies with increasing temperature owing to the variation in material properties such as Young's modulus [28]. The sense frequency change rate was more pronounced than the drive frequency. This is because of the different beam shapes; the thermal expansion causes more changes in the dimensions of the sense beam than in the drive beam. The curve for the drive displacement ($x_d$) shows a non-linear trend in the drive displacement, with a slight decrease at lower and higher temperatures. This variation indicates a complex interaction between the thermal effects and potential resulting stresses that potentially modify the beam dimensions and critical gaps for the drive beam. The interplay of these factors—primarily the variation in material stiffness, resonant frequency, and damping (Q-factor)—collectively determines the drive displacement achieved under constant drive conditions, leading to the observed non-linear profile. The bandwidth (BW) curve decreased slightly with temperature. Because BW=Δf×0.54 [27], Δf represents the drive and sense frequency difference. As seen in the bottom two curves, the drive and sense frequencies decrease with temperature, and a decrease in BW is expected with temperature change. The quality factor ($Q_s$) takes into consideration the squeeze film damping variation, thermoelastic dissipation and anchor losses. The curve for the quality factor ($Q_s$) gradually increased with temperature. This is due to the lower damping losses at higher temperatures. Overall, the plots highlight how temperature variation can significantly affect different parameters. These variations potentially add to the overall noise level of the system. All changes in these parameters lead to a change in the overall thermomechanical noise of the MEMS gyroscope, which is discussed below.

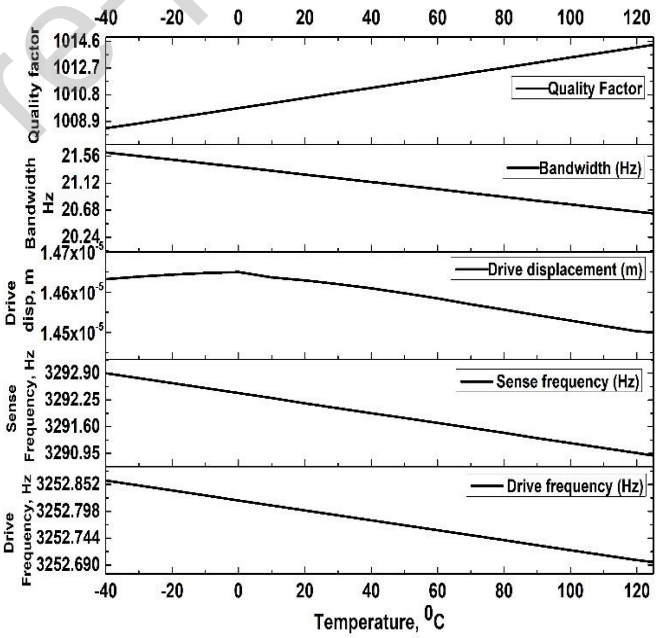



**Figure 14.** The plot shows the variation of various parameters, namely $f_d$, $f_s$, $x_d$, BW, and $Q_s$, with temperatures ranging from -40 $^0$C to 125 $^0$C.

Figure 15 shows the MEMS gyroscope's overall thermomechanical noise as a temperature function for the conventional design without TS and tapered beam design with TS at temperatures ranging from -40 $^0$C to 125 $^0$C. For both designs, the noise increased almost linearly with temperature. As the temperature increased, the thermal agitation became more intense, leading to an overall increase in noise. Moreover, this type of variation in noise is the commutative effect of the various parameter changes observed in the previous curves in Figure 14. These parameter variations

collectively contribute to the gyroscope's sensitivity to noise, influencing the overall performance stability across temperature changes. For conventional design, the noise increases from 5.0 × $10^{-4}$ rad/s at -40 $^0$C to 6.5 × $10^{-4}$ rad/s at 125 $^0$C. but for a tapered beam with TS, the TM noise is approximately at 1.0 × $10^{-5}$ rad/s for the entire temperature range.

The noise trend in Figure 15 appears quasi-linear, differing from the fundamental $T^{0.5}$ thermal noise dependence represented in Equation 4. However, the overall noise metric plotted incorporates multiple temperature-dependent parameters from the comprehensive noise model, including Quality factor (Q), resonant frequencies (ω), and bandwidth (BW), whose individual thermal characteristics are presented in Figure 12. Fitting the simulation data in Figure 13 yields effective dependencies approximated by $K_1*T^{0.465}$ (conventional design without TS) and $K_2*T^{0.9}$ (Tapered design with TS), where $K_1$ and $K_2$ are proportionality constants. While the $T^{0.9}$ trend appears nearly linear, it represents the combined outcome of multiple factors, consistent with the underlying theory. This trend shows that the tapered beam design with TS is more effective in reducing thermomechanical noise over a broad range of temperatures.

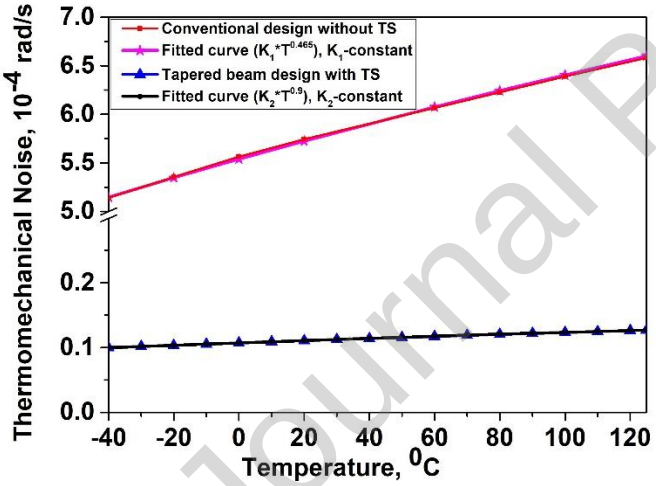



**Figure 15.** The plot of Thermomechanical Noise variation for conventional design without TS and tapered beam design with TS at temperatures ranging from -40 $^0$C to 125 $^0$C.

The TS acts as an insulating layer that limits heat transfer from the environment, whereas the tapered beam design enhances mechanical stability and reduces the heat conduction pathway. This makes it a preferred choice for applications in which thermal stability and minimal noise are crucial, such as precision navigation and inertial measurement systems.

*3.4 Bandwidth Temperature dependence*

The bandwidth (BW) refers to the frequency at which the MEMS gyroscope can accurately detect the rotational signal. A stable BW is crucial for ensuring performance in various applications.

Figure 16 illustrates the variation in BW for the two designs at different temperatures. For conventional designs without TS, at -40 $^0$C, the BW is approximately 21.6 HZ and reduces to approximately 20.6 Hz at 125 $^0$C. This decrease is due to the change in the mechanical properties (such as stiffness) of the gyroscope's structural material, which affects the resonance frequency and, thus, the overall BW. This makes it less reliable in applications where performance consistency over a wide temperature range is critical. However, for a tapered beam design with TS Type IV, the BW hovers at approximately 20.4 Hz with minimal fluctuation.

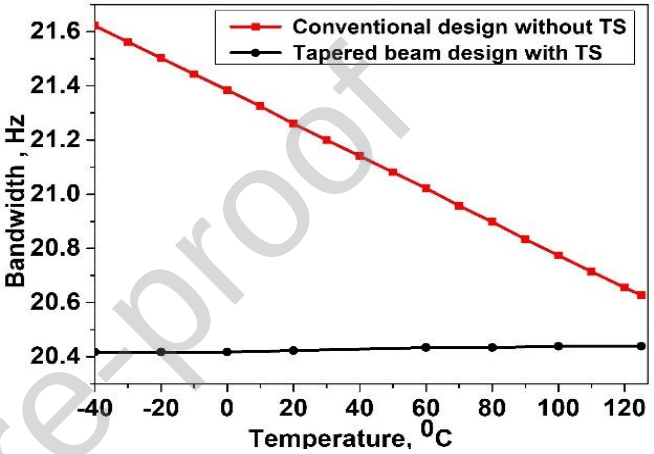



**Figure 16.** Plot shows the bandwidth for conventional design without TS and Tapered beam design with TS Type IV temperature ranging from -40 $^0$C to 125 $^0$C.

The bandwidth (TCB) temperature coefficient was reduced to 158 ppm/$^0$C for the tapered beam with TS from 6060 ppm/$^0$C for the conventional design without TS. The temperature coefficient of frequency (TCF) for the drive and sense structure of a tapered beam with TS is 0.9 mHz/$^0$C and 12 mHz/$^0$C. This is likely due to the combined effect of TS, which acts as a thermal buffer, and tapered beam design, which enhances the structural rigidity and reduces the sensitivity to thermal expansion.

## 4. Results validation

Table 3 presents a comparative analysis of the temperature compensation techniques employed in previous studies and in the current work. The scale factor temperature coefficient and temperature frequency coefficient were considered over a specified temperature range. The main techniques discussed in the literature are frequency mismatch or mode match control, self-calibration, scale factor optimization, fusion compensation, and thermoelectric coolers. Although previous studies have demonstrated effective temperature compensation, the present work offers several advantages.

*a) Wider Temperature Range:* The current work operates over a significantly wider temperature range (-40°C to 125°C), demonstrating its robustness in harsh environments where accuracy and reliability are critical.

Journal Pre-proof

*b) Improved Scale Factor stability:* The present study demonstrates a significant improvement in scale factor stability compared to previous studies, achieving a temperature coefficient of 6 ppm/°C. This is notably lower than the majority of the reported values in only two studies [16] and [29], thereby achieving comparable results. [16] and [29] employed scale factor self-calibration techniques involving extensive external hardware and advanced algorithms and are not efficient, in addition to being cost intensive. [29] incorporated two additional error-correction methods. In addition, these approaches generally require higher power consumption owing to the use of external electronics. By contrast, this study focuses on structural modifications at the design level, enabling effective temperature compensation without relying on external hardware and software, with no additional cost. This approach offers a promising avenue for the development of energy-efficient high-performance MEMS gyroscopes.

*c) Improved frequency coefficient:* The frequency coefficient data for drive and sense mode are available for only a few references; the present work exhibits superior results with TCF for drive and sense mode of 0.9 mHz/$^0$C and 12 mHz/$^0$C, respectively, which are better than other cases as can be seen from this table.

Overall, the results presented in this paper highlight the effectiveness of temperature compensation techniques employed to achieve enhanced sensor performance and reliability. However, Table 3 presents a comparative analysis of the temperature compensation techniques employed in previous studies and in the current work. It provides a benchmark of typical performance values for existing MEMS gyroscopes. This context helps to illustrate the potential impact and significance of achieving improved thermal stability through our design approach, even though a direct validation of our specific simulated structure through experimentation is beyond the scope of this initial study. Future work will focus on the fabrication and experimental verification of the proposed design to directly correlate simulation results with real-world performance.

| **Ref.** | **Technique used or Operation mode** | **Temperature** | | **Temperature coefficient of frequency (TCF)** | | **Scale factor temperature coefficient (SFTC) (ppm/ $^0$C)** |
|---|---|---|---|---|---|---|
| | | Range | ΔT ($^0$C) | Drive mode (Hz/°C) | Sense mode (Hz/°C) | |
| [10] | Fusion compensation method | -30 $^0$C to 70 $^0$C | 100 | Not available | Not available | 8 |
| [12] | Periphery control circuit | -40 $^0$C to 60 $^0$C | 100 | Not available | Not available | 250 |
| [13] | Frequency mismatch control | -40 $^0$C to 60 $^0$C | 100 | Not available | Not available | 282.12 |
| [14] | Scale factor self-calibration | −30 °C to 30 °C | 60 | Not available | Not available | 5.02 |
| [15] | Coefficient optimization | −20 °C to 40 °C | 60 | Not available | Not available | 10 |
| [19] | Thermoelectric Cooler | 0 $^0$C to 60 $^0$C | 60 | Not available | Not available | 210 |
| [29] | Scale factor calibration, Dynamic phase error correction, In-phase error calibration | 15 $^0$C to 45 $^0$C | 30 | Not available | Not available | 4 |
| [30] | Split mode | -40 $^0$C to 60 $^0$C | 100 | Not available | Not available | 27 |
| [31] | Mode matched | -10 $^0$C to 60 $^0$C | 70 | Not available | Not available | 87 |
| [32] | Mode matched | −40 °C to 80 °C | 120 | $41.5 \times 10^{-3}$ | $16.5 \times 10^{-3}$ | Not available |
| [33] | Frequency mismatch control | −40 °C to 80 °C | 120 | 1.61 | 1.31 | Not available |
| **Present study** | **Thermal substrate, beam modification** | **-40 $^0$C to 125 $^0$C** | **165** | **$0.9 \times 10^{-3}$** | **$12\times 10^{-3}$** | **6** |

**Table 3** Presents a comparative analysis of the temperature compensation techniques employed in previous studies and in the current work.

## 5. Fabrication Experiment

Fabrication experiments were conducted on test structures to evaluate the feasibility of fabricating the proposed thermal substrate (TS) design. Figure 15 (a-d) presents scanning electron microscope (SEM) images of the fabricated test structures on Si (110), demonstrating the successful implementation of the design. The wet etching is carried out in 40 wt% KOH solution at 120 °C [34]. Figures 15(a) and 15(c) provide top-down views of the fabricated TS, revealing an array of square features representing individual thermal elements or heat sinks. This consistent patterning across the substrate indicates reliable and consistent processing. Figures 15(b) and 15(c) showcase cross-sectional views, highlighting the slanted sidewalls characteristic of wet etching processes with surface roughness of 50.1 nm. Figure 15 (d) shows the pillars of the thermal substrate etched on the backside of the TS. These images provide visual evidence of the successful fabrication of the TS design and demonstrate the effectiveness of the wet etching techniques employed. Wet etching's inherent isotropy and crystallographic sensitivity limit uniformity and precision for complex, high-aspect-ratio features. Conversely, while offering superior control, dry

etching as an alternative approach typically involves higher costs and slower processing speeds than wet etching. An estimated range for the bonding interface thermal resistance for direct bonding is 0.1 to 10 K mm²/W.

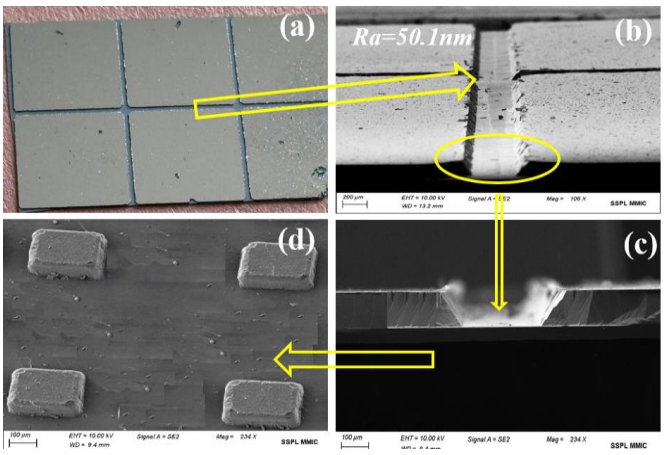



**Figure 17** (a) Top view of the TS, (b) SEM image showing channel between the TSs where dicing will be carried out, (c) Magnified view of one of the channels, (d) Backside of the TS showing etched pillars.

## 5. Conclusions

This study demonstrated a novel and effective strategy for enhancing the thermal performance of MEMS gyroscopes. By integrating a thermally optimized Thermal Substrate (TS) and a specifically engineered design, Transient thermal and stress analyses confirmed improved thermo-mechanical characteristics, which directly translated into significantly decreased temperature-dependent variations of parameters. This has significantly reduced the heat transfer to the sensor, resulting in substantial improvements in key performance parameters, including sense deflection, sensitivity, noise, and bandwidth. The case study results confirm a temperature reduction of 25 $^0$C and a substantial decrease in the temperature-dependent variation of the critical parameters. The present work demonstrates TCF for drive and sense mode as 0.9 mHz/$^0$C and 12 mHz/$^0$C, respectively, and SFTC of 6 ppm/$^0$C up to 100 rad/s and 1.7 ppm/$^0$C at higher rotation rates. The proposed method offers a promising alternative to conventional temperature compensation techniques, eliminating the need for complex external circuitry and algorithms. A comparative analysis with existing literature confirms the exceptional thermal performance of our proposed method. Our results demonstrate a superior temperature coefficient of frequency (TCF) and scale factor temperature coefficient (SFTC) compared to previous studies. The SEM images of the TS show the feasibility of the fabrication using the wet etching technique. The versatility of this approach makes it applicable to a wide range of MEMS sensors, paving the way for advancements in thermal management and the development of high-performance sensor technologies.

**Conflict of interest statement**

The authors declare no competing financial interests or personal relationships that could influence the work reported in this manuscript.

**Compliance with Ethical Standards**

This ensures that the accepted principles of ethics and professional conduct are followed. The authors declare no conflict of interest in the work presented in the manuscript. This study did not involve human or animal participants.

**Funding statement**

No funding agency is involved in the research work of this manuscript.

**Acknowledgementss**

The authors would like to extend their sincere gratitude to the Director of the Solid State Physics Laboratory (SSPL) and Delhi Technological University (DTU) for their generous support and permission to publish this study.

**References**

[1] Harindranath A and Arora M 2024 A systematic review of user - conducted calibration methods for MEMS-based IMUs Measurement (Lond) 225

[2] Huang W, Yan X, Zhang S, Li Z, Hassan J N A, Chen D, Wen G, Chen K, Deng G and Huang Y 2023 MEMS and MOEMS Gyroscopes: A Review Photonic Sensors 13

[3] Gill W A, Howard I, Mazhar I and McKee K 2022 A Review of MEMS Vibrating Gyroscopes and Their Reliability Issues in Harsh Environments Sensors 22

[4] Zhang H, Zhang C, Chen J and Li A 2022 A Review of Symmetric Silicon MEMS Gyroscope Mode-Matching Technologies Micromachines (Basel) 13

[5] Bhatt G, Manoharan K, Chauhan P S and Bhattacharya S 2019 MEMS Sensors for Automotive Applications: A Review Energy, Environment, and Sustainability (Springer Nature) pp 223–39

[6] Passaro V M N, Cuccovillo A, Vaiani L, De Carlo M and Campanella C E 2017 Gyroscope technology and applications: A review in the industrial perspective Sensors (Switzerland) 17

[7] Xiao Y, Meng J, Yan H, Wang J, Xin K and Tao T 2023 Research on the Application of MEMS Gyroscope in Inspecting the Breakage of Urban Sewerage Pipelines Water (Switzerland) 15

[8] Liechti R 2024 Equivalent electrical circuits for electroacoustic MEMS design: a review Journal of Micromechanics and Microengineering 34

[9] Chiu S R, Teng L T, Chao J W, Sue C Y, Lin C H, Chen H R and Su Y K 2014 An integrated thermal compensation system for MEMS inertial sensors Sensors (Switzerland) 14 4290–311

[10] Cui J, Yan G and Zhao Q 2019 Enhanced temperature stability of scale factor in MEMS gyroscope based on multi parameters fusion compensation method Measurement (Lond) 148

[11] Xing B, Zhou B, Zhang X, Zhang W, Hou B, Wei Q, Zhang T and Zhang R 2020 Thermal Deformation Suppression Chip Based on Material Symmetry Design for Single Center Supported MEMS Devices IEEE Access 8 43314–24

[12] Cao H, Li H, Sheng X, Wang S, Yang B and Huang L 2013 A novel temperature compensation method for a MEMS gyroscope oriented on a periphery circuit Int J Adv Robot Syst 10

[13] Li W, Xiao D, Wu X, Su J, Chen Z, Hou Z and Wang X 2017 Enhanced temperature stability of sensitivity for MEMS gyroscope based on frequency mismatch control Microsystem Technologies 23 3311–7

[14] Wang P, Li Q, Zhang Y, Xi X, Wu Y, Wu X and Xiao D 2022 Scale Factor Self-Calibration of MEMS Gyroscopes Based on the High-Order Harmonic Extraction in Nonlinear Detection IEEE Sens J 22 21761–8

Journal Pre-proof

[15] Jia J, Ding X, Qin Z, Ruan Z and Li H 2021 In-Run Scale Factor Compensation for MEMS Gyroscope without Calibration and Fitting IEEE Sens J 21 7316–25
[16] Galiouna Nassiopoulou A and Kaltsas G 2000 Porous Silicon as an Effective Material for Thermal Isolation on Bulk Crystalline Silicon
[17] Behrmann O, Lisec T and Gojdka B 2022 Towards Robust Thermal MEMS: Demonstration of a Novel Approach for Solid Thermal Isolation by Substrate-Level Integrated Porous Microstructures Micromachines (Basel) 13
[18] Wang Y, Cao R, Li C and Dean R N 2021 Concepts, Roadmaps and Challenges of Ovenized MEMS Gyroscopes: A Review IEEE Sens J 21 92–119
[19] Wang Z, Zhou Y, Jiang B, Zhang J, Zhou T and Su Y 2024 A Compact Temperature Controller for MEMS Vibratory Gyroscopes Using Thermoelectric Cooler IEEE Trans Electron Devices 71 3888–94
[20] Shaveta, Bhan R K and Chaujar R 2024 Tapered-Based Novel Thermal Compensation Techniques for Enhanced Performance of MEMS-Based Gyroscope 2nd International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2024 (Institute of Electrical and Electronics Engineers Inc.)
[21] Frank P. et.al. Fundamentals-of-heat-and-mass-transfer-6th-edition JOHN WILEY & SONS
[22] Pannaga N and Rajesh R P Review: MEMS Fabrication Technology
[23] Saffih F, Con C, Alshammari A, Yavuz M and Cui B 2014 Fabrication of silicon nanostructures with large taper angle by reactive ion etching Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena 32
[24] Marinis T, Soucy J, Marinis T F, Soucy J W and Johansson B D 2007 High Performance MEMS Inertial Instruments Fabricated on LTCC Substrates
[25] Allegato G, Corso L and Valzasina C 2022 Inertial Sensors Silicon Sensors and Actuators: The Feynman Roadmap 439–75
[26] He J, Xie J, He X, Du L and Zhou W 2016 Analytical study and compensation for temperature drifts of a bulk silicon MEMS capacitive accelerometer Sens Actuators A Phys 239 174–84
[27] Bao M-H 2000 Handbook of Sensors and Actuators 8 Micro Mechanical Transducers Pressure Sensors, Accelerometers and Gyroscopes
[28] Nandanpawar M L and Rajagopalan S 1978 Wachtman's equation and temperature dependence of bulk moduli in solids J Appl Phys 49 3976–9
[29] Burak Eminoglu, Mitchell H. Kline, Igor Izyumin, Yu-Ching Yeh and Bernhard E. Boser 2014 Background Calibrated MEMS Gyroscope (SENSORS, IEEE)
[30] Zhao Y, Zhao J, Wang X, Xia G M, Shi Q, Qiu A P and Xu Y P 2018 A Sub-0.1°/h Bias-Instability Split-Mode MEMS Gyroscope with CMOS Readout Circuit IEEE J Solid-State Circuits 53 2636–50
[31] Norouzpour-Shirazi A and Ayazi F 2018 A Dual-Mode Actuation and Sensing Scheme for In-Run Calibration of Bias and Scale Factor Errors in Axisymmetric Resonant Gyroscopes IEEE Sens J 18 1993–2005
[32] Bashir U, Bazaz S A, Saleem M M, Shakoor R I, Tariq M O and Kumar P 2024 Design and analysis of mode-matched decoupled mass MEMS gyroscope with improved thermal stability Meas Sci Technol 35
[33] Tran T H, Nguyen Q V, Lo V H, Nesterenko T G and Nguyen H C 2023 Frequency Characteristics of a Micro Electromechanical Gyroscope under Temperature Variations International Journal of Engineering Trends and Technology 71 178–85
[34] Dutta S, Imran M, Kumar P, Pal R, Datta P and Chatterjee R 2011 Comparison of etch characteristics of KOH, TMAH and EDP for bulk micromachining of silicon (110) Microsystem Technologies 17 1621–8

## Author's Biography



Prof. Rishu Chaujar received her B.Sc. (Hons.), M.Sc., and Ph.D. degrees in Electronics from the University of Delhi, India, in 2003, 2005, and 2009, respectively. For her academic excellence in M.Sc. (Electronics), she was awarded a Gold Medal and a talent appreciation award. Currently, she is a faculty member in the Department of Physics at the University of Delhi. Her primary area of specialization is Semiconductor Device Modeling and Simulation, focusing on the analysis of Novel Device Structures, FinFETs, and Nanowire FETs. Prof. Chaujar has authored or co-authored over 300 publications in international journals. Dr. Chaujar is a Life Member of the Semiconductor Society of India, an Executive Member of the IEEE EDS Delhi Chapter (New Delhi, India), and a member of the IEEE Communication Society, the International Association of Engineers (Hong Kong), and IEEE.



Shaveta completed her B.Tech in Electronics and Communication Engineering (ECE) from NIT, Hamirpur in 2004, and her M.Tech in Nanoscience and Technology from Delhi Technological University in 2020. Currently, she is working as Scientist 'F' at the Solid State Physics Laboratory (SSPL), DRDO. Her area of interest is MEMS-based inertial sensors. She is the recipient of various awards for the development of MEMS technology at SSPL

Journal Pre-proof

Journal Pre-proof



R.K. Bhan is a retired Scientist 'H' / Outstanding Scientist from the Solid State Physics Laboratory (SSPL), DRDO, and currently works with the Institute of Defence Scientists and Technologists (IDST), DRDO. His career has centered on semiconductor and microelectronic devices. He has made significant contributions to the fields of MEMS sensors, infrared detector development, Charge Coupled Devices (CCD), CMOS devices, and infrared detector modelling and characterisation. Mr. Bhan has authored or co-authored over 100 publications in international journals and has also published several technical reports, notably on infrared detector characterisation and MEMS sensors. He is a Senior Member of both the Materials Research Society of India and the Semiconductor Society of India and is listed in biographical references such as *Marquis Who's Who* and the *Dictionary of International Biography*.

## Author's statement

On behalf of all co-authors, I, the submitting author, hereby declare that this manuscript, titled "Mitigating Thermal Effects in MEMS Gyroscopes: A Novel Substrate Integration Approach" is our original work, has not been published previously, and is not under consideration for publication elsewhere in any language. We confirm that all authors have contributed significantly to the work and have approved the final version of the manuscript. We also confirm that all necessary permissions for any third-party material included in the manuscript have been obtained and properly acknowledged. We have adhered to the ethical guidelines of the journal and relevant research practices. We have no conflicts of interest to disclose, or if any exist, they have been clearly stated in the manuscript.

Sincerely,

**Dr. Rishu Chaujar**
Microelectronics Research Lab
Department of Applied Physics
Delhi Technological University
Email: chaujar.rishu@dtu.ac.in

**Shaveta**
Solid State Physics Laboratory
Defence Research and Development Organization
Email: shavetarajial@gmail.com

Journal Pre-proof

**Declaration of interest**

**Conflict of Interest**: The authors declare that they have NO affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.

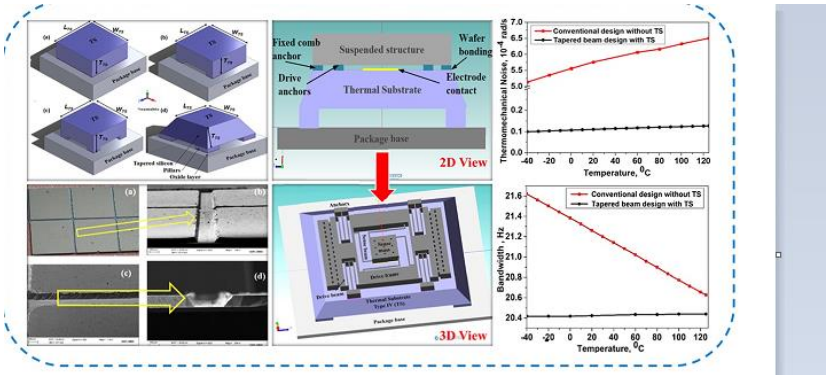
**Funding Statement**: No funding agency is involved in the research work of this manuscript.

**Ethical Compliance**: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee.

**Data Access Statement**: Research data supporting this publication are available.

**Acknowledgements:** The authors would like to thank the Directors of SSPL, IDST, and DTU, Delhi, for providing permission to publish this work on the design of this MEMS Gyroscope.

**Declaration of generative AI: None of** AI and AI-assisted technologies are used in the writing of the manuscript.




## Highlights

Here are the key highlights of the manuscript:

- **Novel Thermal Management Approach:** Introduces a new method using a thermally optimized substrate at the packaging level to control temperature offset and reduce heat transfer to the MEMS gyroscope sensor.
- **Synergistic Design Optimization:** Combines the substrate approach with an engineered gyroscope design that minimizes heat flow to sensitive areas like the sense mass.
- **Significant Temperature Reduction:** The combined approach resulted in a substantial 25°C temperature reduction at the sensor level.
- **Enhanced Thermal Stability:** Demonstrates significant improvements in key performance parameters:
    - Sense deflection reduced from $2.7 \times 10^{-6}$ µm/°C to $8.6 \times 10^{-8}$ µm/°C.
    - Scale factor temperature sensitivity reduced from 232 ppm/°C to 6 ppm/°C.
    - Notable improvements in noise reduction and bandwidth stabilization were observed.
- **Validation through Comparison:** Superior thermal performance of the proposed method is validated against existing literature.
- **Feasibility Study:** Reports on etching experiments conducted to assess the feasibility of the thermal substrate.
- **Profound Implications:** Highlights the significant impact of effective thermal management on the reliability and accuracy of MEMS gyroscopes for diverse applications.

Journal Pre-proof

2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3315-2169-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/IATMSI64286.2025.10984886

# Online Proctoring Systems - A Hardware Implementation using Enhanced Eye Detection and Gaze Tracking

Ashish Chotani[1] and Rohit Kumar[1]

[1]*Electronics and Communications Department, Delhi Technological University*, India-110042

***Abstract*—Online Proctoring Systems(OPS) have been on the rise due to the increasing dependence on video communication. With an important application in exam monitoring, these systems ensure maintenance of academic integrity by utilizing Artificial Intelligence and Computer Vision to detect suspicious behavior and automate the process of proctoring. Each OPS consists of 4 processes in the following order - Feature Extraction, Gaze Tracking and Estimation. Environment modeling and finally Inferential/ Prediction models for behavioral analysis. In this paper, we discuss the two most integral parts of any OPS, namely, Feature Extraction and Gaze Tracking. We compare some state-of-the-art methods available in both aspects of an OPS and attempt to clone them in a hardware setting. We indicate how deep learning-based methods offer more robust detection and tracking compared to template-matching-based classification. We show that CNN-based Deep Learning methods tend to exhibit higher accuracy and robustness to noise in detecting eye-features such as the pupil and iris as compared to Haar-based Cascading classifiers, which are a decent alternative for lightweight real-time applications in a controlled setting with minimal noise. For the second process in an OPS, we describe that Geometric and Camera Calibration is unmatched for gaze-tracking in controlled and constrained environment, while Optical Flow provides an effective and adaptable way to track gaze.**

***Keywords*—Online Proctoring, Gaze Detection, Haar-Cascade Classi-fier, Geometric Calibration**

## I. Introduction

WITH the increased dependence on wireless video connectivity, learning has shifted online too. This has only increased post the COVID-19 pandemic with all institutions conducting activities like teaching, testing, interviewing and much more online. This has made the need for Online Proctoring Systems(OPS) a reality and an important concern for examiners and interviewers to ensure academic and ethical integrity. OPS have the capability of detecting unethical activities such as impersonation,cheating during both online exams and interviews enabling institutions to conduct assessments in a controlled and fair environment.Although manual proctoring systems have been present and employed by many, there has been no doubt that they are way less reliable and resource intensive as compared to their online counterparts. While a human-eye can be easily be deceived, the major drawback lies in the fact that manual systems require more human interference thus causing institutions a lot in human-resources. OPS on the other hand are more accessible and convenient to use to institutions. Apart from this, OPS offer a scalable, reliable and robust environment that can be tailored to the users choice to provide automated and unbiased insights of human movements. Further, OPS can be modeled to detect suspicious behavior by analyzing the difference in micro-patterns of a human by studying variations and deviations in movements without a third-human intervention. Thanks to the advancements made in Computer Vision, and rise of more sophisticated algorithm, creating an Online Proctoring System is less complex that it seems. A Computer Vision based OPS consists of majorly 4 core components. Firstly, Face Detection and Identity verification ensures that the out of the entire face, the region of interests are captured from the face of the candidate being proctored. [1] proposed the use of Haar-like features in cascade to detect faces in gray-scale images quickly a boosting classifier. Various Deep Learning based solutions have also been proposed to handle this aspect of an OPS. [2] integrates detection and alignment tasks, enabling efficient identity verification. Another approach , FaceNet, provides a robust matching technique against pre-registered profile. The second major process involves Gaze Tracking and Attention Monitoring. An integral process, This process monitors the located eye region including the pupil and iris for long durations to derive insights on whether the activity can be flagged as suspicious or not. [3] utilizes CNN models for gaze-detection using different web-cam captured images meanwhile [4] utilizes methods like pupil dilation and corneal reflection to enhance gaze tracking precision. The third process of an OPS involves Environment Monitoring and Anomaly Detection. Modern systems depend on scene understanding and real-time object detection to analyze the environment and detect any anomalous behavior. [5], [6] suggest some realtime models based on YOLO and Faster R-CNN to detect the presence of unauthorized objects in the form of additional monitors or multiple faces. Temporal models like LSTM and RNN architectures have also been utilized to analyze a sequence of frames to determine movements or changes in the backgrounds which might indicate the presence of anomalies. Finally, the last block/process of an OPS consists of an Inferential model that utilizes all the information gathered to classify the user as suspicious or not-suspicious. Further studies and research have been focused on accurately estimating the human behavior using pose analysis and estimation. Significant strides have been made in pose-estimation as well. [7] and [8] discuss var-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:54 UTC from IEEE Xplore. Restrictions apply.

979-8-3315-2169-1/25/$31.00 © 2025 IEEE

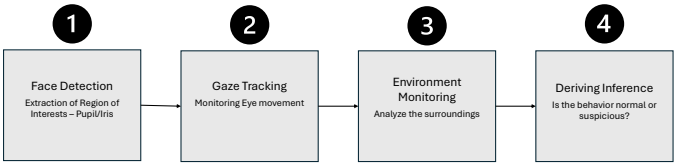



Fig. 1. Working of an Online Proctoring System

ious techniques to estimate pose-estimation using landmarks and CNN models respectively. Work has also been done in the pre-processing activities corresponding to OPS. [5] shed light on image enhancement and frame sampling methods to further enhance feature extraction and decision making. With a plethora of models and architectures available to model an OPS, it becomes important to benchmark the state-of-the-art methods with the ones drawn using first principle. Our work clearly lays out different alternatives to emulate the first two processes of Face/Feature Detection and Gaze Tracking. We further device a hardware model using an Arduino Mega board with a camera module.

## II. Methodologies

### A. Feature Extraction - Detecting the Pupil/Iris

*1) Thresholding Based Solutions and the Hough Transform:* Thresholding uses edge detection algorithms, such as Canny edge detection, to isolate the pupils' center by enhancing contrast between the pupil and surrounding areas. The method involves setting intensity thresholds to segment the image into regions of interest. The Hough Transform complements thresholding by detecting circular patterns, making it suitable for identifying the circular structure of the iris. This method involves transforming the image space into a parameter space to detect circles based on their geometric properties. However, while thresholding and Hough Transform are computationally efficient and suitable for resource-constrained setups like Arduino implementations, they are sensitive to noise and lighting variations. Their simplicity limits their adaptability to complex scenarios. This approach contrasts with more sophisticated methods as it focuses on basic intensity and geometric properties, making it less robust to environmental inconsistencies.

*2) Haar-based Cascaded Classifiers:* Haar-based classifiers operate by sliding a detection window across the image, applying multiple filters trained on thousands of positive (eye-containing) and negative (non-eye-containing) samples. The method identifies patterns resembling eyes based on Haar-like features and can process real-time data on low-resource systems. Its cascade structure progressively eliminates unlikely regions, enhancing computational efficiency. However, Haar classifiers struggle with high noise levels and fail to generalize well in dynamic environments, unlike deep learning methods. Compared to thresholding and the Hough Transform, Haar classifiers are more adaptive to varying shapes and textures, although they lack the flexibility of data-driven learning approaches like CNNs. Compared to thresholding and the Hough Transform, Haar classifiers are more adaptive to varying shapes and textures, although they lack the flexibility of data-driven learning approaches like CNNs.

*3) Deep Learning based methods:* CNNs are pre-trained models capable of classifying regions of live video feeds as eyes with unparalleled accuracy. They extract hierarchical features from input images, enabling them to handle occlusions, lighting variations, and complex shapes. CNNs rely on large annotated datasets like the BioID Eye Dataset to learn robust feature representations. However, their computational demands limit their utility in low-resource environments like Arduino systems.Deep learning methods significantly outperform traditional approaches by leveraging data-driven representations, making them ideal for dynamic and complex scenarios.

### B. Gaze Tracking - Detecting the movements of the eye

*1) Optical Flow:* Optical Flow calculates pixel-level motion between consecutive frames by analyzing brightness patterns. It assumes brightness constancy and smooth motion, making it effective for tracking gradual eye movements. The Lucas-Kanade algorithm is commonly used for this purpose, solving for motion vectors across the RoI. Optical Flow's reliance on frame-to-frame consistency differentiates it from calibration-based methods, making it suitable for real-time scenarios but less robust to abrupt motion or occlusions.

*2) Geometric and Camera Calibration:* This method maps eye movements to screen coordinates using geometric transformations and calibration algorithms. By analyzing the spatial relationship between the eyes' position and screen coordinates, this technique ensures precise gaze tracking. Compared to Optical Flow, calibration-based methods are more precise, but less adaptable to real-time changes or hardware-constrained environments. They are ideal for applications requiring high accuracy, such as assistive technologies.

*3) Template Matching:* Template Matching involves tracking the pupil by comparing it to a predefined template across frames. This method uses cross-correlation to measure the similarity between the template and regions within the RoI, allowing it to detect subtle movements effectively. Unlike Geometric Calibration, Template Matching is computationally simpler and more adaptable to low-resource environments like Arduino implementations. However, it is less precise in dynamic or noisy conditions.

## III. Hardware Model

Implementing an Online Proctoring System (OPS) on embedded hardware is combining hardware and software in a systematic method to observe and analyze an individual's engagement with his examination. The architecture blends input acquisition, data processing, and output management.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:54 UTC from IEEE Xplore. Restrictions apply.

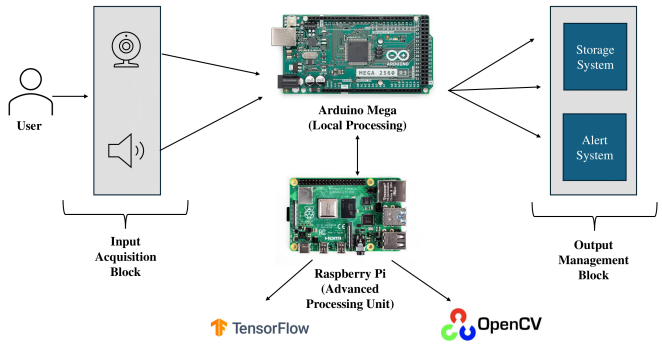



Fig. 2. Hardware Model: Online Proctoring System

Input acquisition acquires continuous video through the embedded camera and monitors the environment with motion and sound sensors. These three parameter inputs aid in the detection of user presence, movements, and disturbances in the environment. Studies have shown that motion sensors and environmental data can effectively provide evidence for video-based proctoring to identify possible irregularities. [6].

Data processing is the stage through which computational limitations faced by embedded systems are taken care of. Basic tasks like motion and noise detection are performed locally on the embedded hardware, while more complex operations, such as facial recognition and gaze tracking, have been offloaded to external processors. This hybrid approach allows the system to utilize efficient computer vision algorithms, including face detection and object recognition frameworks, to identify suspicious activities, such as multiple persons in the frame or use of unauthorized objects. [1], [2]. Output management allows instant detection and reporting of anomalies. The typical responses to such oddities include visual signals or auditory alerts for detectable deviations, unexpected movements, and environmental distractions. The logs are secured locally or sent to some distant site to be reviewed afterward. Thus, it provides a firm monitoring capability and offers a built-in scalability. Studies point out that modular processes incorporating varying data streams are appropriately able to flag cheating attacks with low false positives.

It should be understood that OPS was designed to be flexible and scalable, enabling the addition of an advanced feature set. For example, detection of motion also integrated with gaze tracking may improve in the determination of where examinees focus their attention [4]. Sound-level analysis can also be applied for detection of external disturbance, thereby increasing detection capability. So, because the acquisition of input is supported by embedded hardware with processing done on a less local machine, it is less expensive and supports deployment in low-resource settings. The hybrid manner also corresponds to recent research advocating distributed computation in real-time monitoring systems [7].

Overall, this approach provides a functional and adaptable solution for OPS, balancing the resource constraints of embedded hardware with the computational requirements of advanced computer vision techniques.

Building on this framework, implementing the Online Proctoring System (OPS) on an Arduino Mega Board exemplifies the integration of hardware and software for low-resource environments. The Arduino Mega serves as the central controller for input acquisition, utilizing components such as a camera module for continuous video capture, PIR motion sensors for detecting presence and movements, and KY-038 sound sensors for monitoring ambient noise levels. These inputs are processed in a hybrid manner to address the computational limitations of embedded systems. Basic tasks, such as motion detection and noise analysis, are handled locally on the Arduino Mega, while computationally intensive processes, like facial recognition and gaze tracking, are offloaded to external processors, such as Raspberry Pi or NVIDIA Jetson Nano. This setup leverages frameworks like OpenCV and TensorFlow to perform advanced computer vision operations, such as detecting multiple persons in the frame or unauthorized objects. Output management is facilitated through visual alerts

TABLE I
Necessary Hardware Accessories

| Accessory | Purpose |
|---|---|
| Arduino Mega Board | Central controller for basic tasks, sensor management, and local data processing. |
| PIR Motion Sensor | Detects examinee presence and movements. |
| KY-038 Sound Sensor | Monitors ambient noise levels for disturbance detection. |
| OV7670 Camera Module | Captures video streams for real-time monitoring (or use a USB camera with USB host shield). |
| External Processor (e.g., Raspberry Pi or NVIDIA Jetson Nano) | Handles computationally intensive tasks like facial recognition and gaze tracking. |
| Wi-Fi Module (ESP8266/ESP32) | Enables communication with external servers for data transmission. |
| SD Card Module | Stores logs and captured data locally. |
| LED Indicators | Provides visual alerts for anomalies. |
| Piezo Buzzer | Generates auditory alerts for environmental or behavioral disturbances. |
| USB Host Shield | Facilitates the integration of a USB camera with the Arduino Mega. |

using LEDs, auditory signals via a piezo buzzer, and secure logging of anomalies. Logs can be stored locally using an SD card module or transmitted to a remote server via a Wi-Fi module, ensuring scalability and real-time monitoring. The modular architecture of this implementation supports the integration of advanced features, such as combining gaze tracking with motion detection to analyze focus or utilizing sound-level analysis to detect external disturbances. By striking a balance between affordability and computational efficiency, this hybrid approach enhances monitoring capabilities while making OPS accessible in low-resource settings. Rigorous testing and calibration ensure the adaptability and reliability of the system, allowing it to handle real-time proctoring requirements effectively.

## IV. Performance Analysis

*1) Feature Extraction Methods:* Feature extraction is crucial for accurately identifying the Region of Interest (RoI),

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:54 UTC from IEEE Xplore. Restrictions apply.

such as eyes or pupils, in images or video feeds. CNNs, Haar Cascade Classifiers, Hough Transform, and Thresholding are commonly used techniques, and their performance has been evaluated in various studies. CNNs, using the BioID Eye Dataset, achieve the highest accuracy of 95%-98%, excelling in dynamic environments with variable lighting and occlusions. However, their performance on Arduino-emulated models is hindered by significant latency due to limited computational power, making them unsuitable for real-time applications on such hardware. Haar Cascade Classifiers, tested on the Caltech Faces Dataset [1], provide an accuracy of 85%-90% in controlled environments and are lightweight enough to process real-time data on standard CPUs and Arduino implementations. Although effective, they exhibit reduced accuracy in noisy environments, but their efficiency on low-resource hardware makes them a viable option for embedded systems. The Hough Transform, on the BioID Dataset [9], achieves 70%-80% accuracy, effectively detecting circular features like pupils in high-contrast scenarios. On Arduino implementations, its sensitivity to environmental inconsistencies and computational inefficiencies limits its applicability. Thresholding, used primarily for edge detection, achieves 65%-75% accuracy [10]. It performs efficiently on Arduino-emulated systems due to its simplicity and minimal computational requirements, making it suitable for low-resource applications requiring basic feature detection. From Arduino implementations, it was observed that Haar Cascade Classifiers and Thresholding provide the most practical solutions for low-resource environments, with Haar Cascade balancing accuracy and computational efficiency. In contrast, CNNs are limited to environments with high computational resources, and Hough Transform's sensitivity to noise reduces its reliability in hardware-constrained setups.

*2) Gaze Tracking Methods:* Gaze tracking techniques analyze pupil or iris movement to determine gaze direction and map it to screen coordinates. Optical Flow, Geometric and Camera Calibration, and Template Matching are the most widely used methods. Optical flow in the OpenEDS data set [11], achieves 80% - 90% accuracy to track smooth and continuous movements. On Arduino-emulated models, its computational demand leads to significant latency, making real-time tracking challenging. Geometric Calibration on the GazeCapture Dataset [12], provides a precision of 95%-98%, making it the most accurate technique for gaze-to-screen mapping. However, the calibration process becomes cumbersome on Arduino systems due to hardware limitations, leading to reduced usability in embedded applications. Template matching using the BioID dataset, achieves an accuracy of 85%-92%, performing reliably in stable environments with consistent lighting. Its low computational complexity makes it highly adaptable to Arduino setups, providing reasonable accuracy and efficiency.

Arduino-based implementations of these models revealed that Template Matching strikes the best balance of simplicity and accuracy for hardware-constrained systems. Optical Flow, while effective in robust systems, exhibited noticeable latency in Arduino setups, reducing its suitability for real-time applications. Geometric Calibration, although highly precise, was impractical for Arduino implementations due to the intensive calibration process and sensitivity to environmental changes.

TABLE II
Performance Analysis of Feature Extraction

| Method | Accuracy (%) | Hardware Feasibility and Robustness |
|---|---|---|
| Thresholding and Hough Transform | 65 | High feasibility, low robustness: Sensitive to noise and lighting |
| Haar Cascade Classifiers | 85 | High feasibility, medium robustness: Struggles in noisy environments |
| Deep Learning (CNNs) | 95 | Low feasibility, high robustness: Handles occlusions and lighting variations |

TABLE III
Performance Analysis of Gaze Tracking Methods

| Method | Accuracy (%) | Hardware Feasibility and Robustness |
|---|---|---|
| Optical Flow | 90 | Low feasibility, medium robustness: Struggles with abrupt motion |
| Geometric Calibration | 98 | Low feasibility, high robustness: Accurate but requires calibration |
| Template Matching | 92 | High feasibility, medium robustness: Reliable in stable conditions |

## V. Conclusion

CNNs have theoretically good performance in feature extraction under dynamic conditions, while not plausible in low-resource hardware like Arduino for latency problems. Although Haar Cascade Classifier and Thresholding are viable solutions for resource-constrained setups, the former is more effective in controlled conditions and the latter is better suited for less complex tasks. The most exact method for gaze tracking in a high-resource environment is Geometric Calibration but is not feasible in hardware-constrained platforms due to calibration overheads. Optical Flow provides an infrastructure for real-time tracking in robust systems but exhibits latency issues at Arduino implementations. Template Matching is the best choice for Arduino-based implementations in terms of simplicity but also accuracy since it strikes a balance between both where computational resources are limited.

## References

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1, pp. I–I, 2001.

[2] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[4] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2176–2184, 2016.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:54 UTC from IEEE Xplore. Restrictions apply.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.
[6] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.
[7] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 607–626, 2009.
[8] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1653–1660, 2014.
[9] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, 1972.
[10] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.
[11] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 1, 1981.
[12] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:54 UTC from IEEE Xplore. Restrictions apply.

Physics Letters A 551 (2025) 130625

Contents lists available at ScienceDirect

Physics Letters A

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

Letter

# Optimizing thermoelectric performance of halide perovskite $Ca_3AsBr_3$ through strain engineering

Sumedha Yadav, Sangeeta, Kulwinder Kumar, Mukhtiyar Singh *

*Computational Quantum Materials Design (CQMD) Laboratory, Department of Applied Physics, Delhi Technological University, Delhi 110042, India*

ARTICLE INFO

Communicated By: Lilia Woods

*Keywords:*
DFT
Halide perovskites
Thermoelectric properties
Lattice thermal conductivity
Figure of merit

ABSTRACT

We investigate structural, dynamical, elastic, electronic, and thermoelectric properties of cubic perovskite halide $Ca_3AsBr_3$ under the effect of strain using first-principles calculations. It is found to have a direct bandgap of 2.425 eV that further reduces under compressive strain making it an ideal candidate for thermoelectric application. The material satisfies various stability criteria, e.g., dynamical, thermodynamical, and mechanical. The transport calculations show the highest Seebeck coefficient of value -458.227 $\mu VK^{-1}$ at 700 K for *n*-type $Ca_3AsBr_3$ unstrained structure at carrier concentration $1 \times 10^{21}$ $cm^{-3}$, which is further enhanced to -482.366 $\mu VK^{-1}$ for -2 % strain. The lattice thermal conductivity of the material is reduced from 1.243 $Wm^{-1}K^{-1}$ to 0.627 $Wm^{-1}K^{-1}$ at 700 K under 3 % strain. This low thermal conductivity, coupled with positive power factor values, results in increased peak thermoelectric figure of merit from 0.36 (unstrained) to 0.56 (+3 % strain) at 700 K for an *n*-type doping concentration of $1 \times 10^{20}$ $cm^{-3}$.

## 1. Introduction

As global energy consumption continues to rise, the conventional energy sources becoming unprecedently more pressing. Moreover, the depletion of conventional fossil fuels and their environmental impact raises serious concerns about long-term sustainability. This challenge makes it essential to explore alternative, renewable energy solutions that can support future energy needs while minimizing ecological harm [1–5]. Thermoelectric (TE) materials represent one of the most promising avenues for sustainable energy conversion. These materials can harness useful electricity from waste heat and thus have been a research focus in recent years [6]. The efficiency of a TE material is given by a dimensionless quantity known as the Figure of Merit (FOM), which is defined by $ZT = \frac{S^2\sigma T}{k_e + k_l}$ where $S$ represents the Seebeck coefficient, $\sigma$ is the electrical conductivity, $\kappa_e$ is the electronic thermal conductivity, and $\kappa_l$ is the lattice thermal conductivity. In order to obtain a high $ZT$, high $S$, high $\sigma$, and a low thermal conductivity ($\kappa$) are required [7,8]. However, these parameters have very strong intercoupling, and it is very challenging to increase the $ZT$ of conventional TE materials [9]. To decouple these coefficients, researchers have opted for many strategies [8,10–12]. The inorganic materials like $Bi_2Te_3$ (effective at room temperature), chalcogenides and half Heusler compounds (suited for mid-to-high temperatures like 400 –800 K), skutterudites and clathrates (500–900 K), Zintl phases (covering various temperature ranges for different compositions ~300–900 K), and oxide perovskites (best for high temperatures ~700–1200 K) are widely used as efficient TE materials across different temperature regimes [13–16]. However, the scarcity of raw materials such as tellurium and lead, along with the complex and costly fabrication processes, significantly hinder the commercialization of such inorganic TE materials.

Inorganic perovskites are low-cost efficient materials emerging as promising materials in science and technology development [17]. Owing to their unique crystal structure $ABX_3$ (with A, and B as cations of different sizes and X as an anion that bonds with both cations) perovskite materials are widely used in chemistry, physics, and other fields with successful synthesization of hundreds of distinct materials that make up the large perovskite family including conductors, semiconductors, and insulators [18]. These perovskites reveal many intriguing properties such as high TE power, ferroelectricity, superconductivity, spin-dependent transport, etc. [19–23]. Further, exploring alternative perovskite frameworks $A_3BX_3$ such as $Sr_3NCl_3$, $Ca_3PX_3$, and $Sr_3PBr_3$, which naturally exhibit cubic symmetry like traditional $ABX_3$ perovskites, offers several advantages, including greater structural stability, improved tunability of electronic, and optical properties and can show great promise for TE applications [24–27]. For instance, $Ba_3AsX_3$ ($X = F$, Cl) has demonstrated $ZT$ values of 0.92 and 0.82, respectively, which

* Corresponding author.
*E-mail address:* mukhtiyarsingh@dtu.ac.in (M. Singh).

https://doi.org/10.1016/j.physleta.2025.130625
Received 18 March 2025; Received in revised form 1 May 2025; Accepted 4 May 2025
Available online 5 May 2025
0375-9601/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

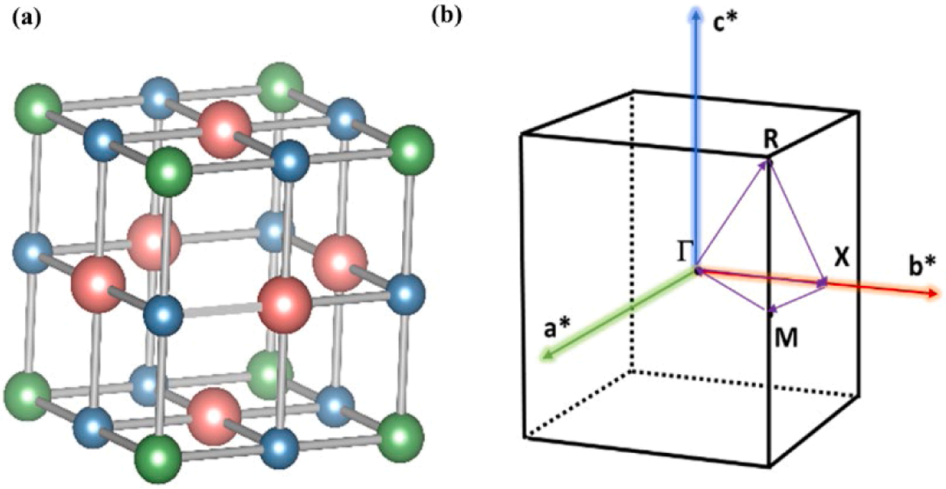



**Fig. 1.** **(a)** Unit crystal structure (Blue, Green, and orange colour represent Ca, As, and Br atoms, respectively) **(b)** Brillouin zone of $Ca_3AsBr_3$.

result from its favorable $\sigma$, optimized $S$, and a stable perovskite framework that supports efficient charge transport [24].

Advancing TE technology requires materials with inherently low $\kappa_l$. Recently, Tranås et al. investigated a wide range of materials using a machine-learning model and first-principles calculations and identified some materials that have intrinsically low $\kappa_l$ [28]. Among the identified candidates, Calcium Arsenide Bromide ($Ca_3AsBr_3$), a cubic halide perovskite, might stand out as a promising TE material, which motivated us to select it for further investigation. This material was first studied experimentally in 1984 by Hadenfeldt et al. [29], where its thermal behavior, was also analyzed. Recently, theoretical work has explored some properties of $Ca_3AsBr_3$, but its TE performance remains unexplored [30]. Many studies have demonstrated how strain and spin-orbit coupling (SOC) reduce the bandgap in halide perovskites, potentially enhancing their TE efficiency [31,32]. Strain influences $S$ and $\sigma$ by modifying the bandgap or enhancing the density of states (DOS) near the Fermi level [33].

In this study, we have investigated the electronic and thermoelectric properties of halide perovskite $Ca_3AsBr_3$ under the effect of compressive and tensile strain using first-principles calculations followed by semi-classical Boltzmann transport theory. The relaxation time and carriers mobilities have been calculated considering various scattering processes. We have shown, by analyzing properties at different carrier concentrations for both $p$-type and $n$-type material, that the carrier doping is an effective way to enhance TE performance. The deep insights into the lattice dynamics of the material by observing group velocity, Grüneisen parameter, and phonon lifetime have helped in better analysis of the thermal properties of the investigated structures. An enhanced $ZT$ (0.55) has obtained for +3 % strain at 700 K compared to the unstrained structure (0.36), but it is still not that high and thus leaves scope for further material optimization.

## 2. Computational details

The full potential linearized augmented plane wave (FP-LAPW) method, which is implemented in WIEN2k software, was used to perform the structural, elastic, electrical, and TE characteristics of $Ca_3AsBr_3$ within the framework of DFT [34,35]. The generalized gradient approximation (GGA) of Perdew-Burke-Ernzerhof (PBE) was used to optimize the structure [36]. We employed an optimized 10 × 10 × 10 $k$-point mesh (Supplementary information Fig. S1) that produced 1000 $k$-points in the Brillouin zone, for the self-consistent calculations. The cut-off energies of the core and valence states were set to −6 Ry, and the plane wave cut-off was set as RMT × $k$max = 7. With an energy convergence of 0.000001 Ry, the Kohn–Sham equations were solved self-consistently. We used the Tran-Blaha modified Becke-Johnson (TB-mBJ) approach to determine the electronic band structure more accurately as most of the time GGA underestimates the band gaps of semiconductors [37]. It is a semi-local approximation to an atomic exact-exchange potential and a screening term. We used a mixing factor of 0.2 in the self-consistent calculations to ensure efficient convergence. For the density of states (DOS) calculations, we used a dense $k$-mesh of 25 × 25 × 25. The Boltzmann Transport Equation (BTE) was solved under the rigid band approximation (RBA) and constant relaxation time approximation (CRTA), as implemented in the BoltzTraP2 code, in order to determine the TE characteristics [38]. For Vienna *Ab-initio* Simulation Package (VASP) calculations, we set the cutoff energy for plane wave basis functions at 520 eV, and a width of 0.01 eV for Gaussian smearing was used. To ensure accuracy, a strict convergence criterion for energy was kept as $1 \times 10^{-7}$ eV for the self-consistent calculations. Using the AMSET code interfaced with the VASP, we computed scattering rates and mobilities within the Born approximation by solving the electronic BTE in the momentum relaxation-time approximation [39,40]. Instead of using CRTA, this code estimates the relaxation time ($\tau$) based on the number of distinct bands and $k$-point dependent scattering approaches. We computed scattering rates in this work by taking into account the following: (a) acoustic deformation potential (ADP), which is in charge of the phonon-electron coupling; (b) ionized impurity (IMP) scattering that denotes the scattering of charge carriers by ionization of the lattice; and (c) polar optical phonon (POP) scattering, which takes into account the interaction among polar optical phonons and electrons. We used Matthiessen's rule to determine the resulting carrier $\tau$, which is the inverse of the scattering rate: $\frac{1}{\tau} = \frac{1}{\tau^{ADP}} + \frac{1}{\tau^{IMP}} + \frac{1}{\tau^{POP}}$, where $\tau^{ADP}$, $\tau^{IMP}$, and $\tau^{POP}$ are the relaxation times received from the ADP, IMP, and POP scattering, respectively [41]. Contrary to POP scatterings, which show inelastic scattering, ADP and IMP scatterings show elastic scattering, where electrons do not gain or lose energy. DFT and density functional perturbation theory (DFPT) were used to calculate the parameters to produce the scattering rates, including polar-phonon frequency, elastic constants, wave-function coefficients, static and high-frequency dielectric constants, potentials, and dense and uniform band structures. An interpolation factor of 50 is used to predict $\tau$ and carrier mobility. The finite displacement approach was used to calculate the phonon dispersion of material to verify the dynamical stability, by building a 2 × 2 × 2 supercell and using a 4 × 4 × 4 $k$-point mesh using the Phonopy code [42]. Using the same supercell, we calculated anharmonic third-order interatomic force constants to obtain the lattice thermal conductivity using Phono3py [42] interfaced with VASP [39]. To obtain the lattice thermal conductivity and other related phonon properties such as

**Table 1**
The optimized lattice parameters (in Å), Volume of the unit cell (in Å$^3$), and bond lengths (in Å) of $Ca_3AsBr_3$, along with available experimental data.

| System | $a = b = c$ | Volume | Bond lengths | |
|---|---|---|---|---|
| | | | Ca-As | Ca-Br |
| $Ca_3AsBr_3$ | 5.963[a], 5.92[b], 5.926[c], 5.95[d] | 212.07[a], 210.64[d] | 2.97[a], 2.96[c], 2.98[d] | 2.97[a], 2.96[c], 2.98[d] |

[a] Present work,.
[b] [29],.
[c] [45],.
[d] [30].

phonon density of states, phonon group velocity, Grüneisen parameter, and phonon lifetime, we used a converged $q$-mesh of 25 × 25 × 25 to sample the Brillouin zone.

## 3. Results and discussion

### 3.1. Structural properties

$Ca_3AsBr_3$ crystallizes in the cubic $Pm\bar{3}m$ space group (221) and the Wyckoff positions occupied by Ca atoms, As atoms, and Br atoms are 3c (½, 0, 0), 1a (0, 0, 0), and 3d (0, ½, ½), respectively, as shown in Fig. 1a. Each As atom is surrounded by six Ca atoms, forming a corner-sharing

**Table 2**
The elastic constants $C_{11}$, $C_{12}$, and $C_{33}$ are in GPa, $G_v$, $G_r$, $G_h$, $E$, and $B$ are in GPa, Poisson's ratio, Pugh's ratio, and $A$ are dimensionless and melting temperature (K) corresponding to all compressive and tensile strains (in %).

| Strain (%) → Elastic Properties ↓ | −3.00 | −2.00 | −1.00 | 0.00 | 1.00 | 2.00 | 3.00 |
|---|---|---|---|---|---|---|---|
| $C_{11}$ | 116.53 | 115.17 | 106.05 | 71.58 | 62.49 | 53.47 | 36.36 |
| $C_{12}$ | 13.70 | 7.31 | 5.42 | 15.20 | 13.75 | 14.24 | 17.28 |
| $C_{44}$ | 21.26 | 30.69 | 28.59 | 15.37 | 13.49 | 10.84 | 9.16 |
| Voigt Shear Modulus ($G_v$) | 33.32 | 39.99 | 37.28 | 20.50 | 17.84 | 14.35 | 9.31 |
| Reuss Shear Modulus ($G_r$) | 27.77 | 37.08 | 34.56 | 18.79 | 16.43 | 13.21 | 9.31 |
| Hill Shear Modulus ($G_h$) | 30.55 | 38.54 | 35.92 | 19.64 | 17.13 | 13.78 | 9.31 |
| Young's Modulus ($E$) | 75.59 | 89.14 | 82.43 | 49.41 | 43.18 | 35.39 | 24.69 |
| Poisson's Ratio ($\nu$) | 0.24 | 0.16 | 0.15 | 0.26 | 0.26 | 0.28 | 0.33 |
| Bulk modulus ($B$) | 47.98 | 43.26 | 38.96 | 33.99 | 30.00 | 27.31 | 23.64 |
| Anisotropy Factor ($A$) | 0.41 | 0.57 | 0.57 | 0.55 | 0.55 | 0.55 | 0.96 |
| Pugh's Ratio ($B/G$) | 1.57 | 1.12 | 1.08 | 1.73 | 1.75 | 1.98 | 2.54 |
| Melting Temperature | 1241.72 | 1233.68 | 1179.74 | 976.02 | 922.33 | 868.99 | 767.87 |

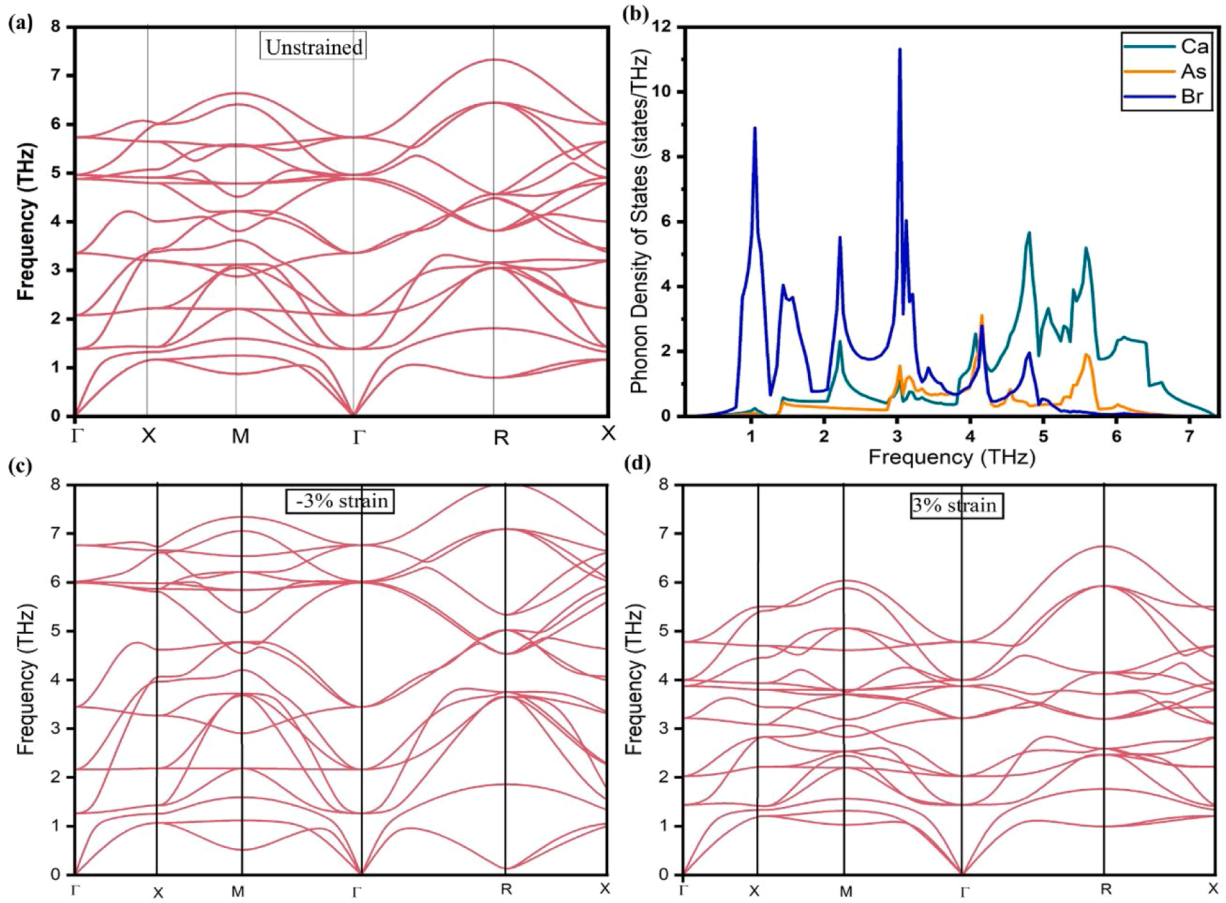



**Fig. 2.** **(a)** The phonon dispersion curve of unstrained $Ca_3AsBr_3$ and **(b)** Phonon Density of states **(c)** phonon dispersion curve of $Ca_3AsBr_3$ under strain –3 % **(d)** phonon dispersion curve of $Ca_3AsBr_3$ under strains +3 %.

octahedral with no tilting in these shared octahedra showing that they are perfectly aligned whereas each Br atom is bonded to four Ca atoms and these four Ca atoms are arranged in a square, flat (co-planar) [43]. We have performed structural optimization of this compound and fitted the Birch-Murnaghan equation of state to compute its equilibrium lattice parameters [44]. Table 1 lists the bond length and optimal structural characteristics. In the absence of strain, the lattice constant of $Ca_3AsBr_3$ is 5.96 Å, this finding confirms the correctness and dependability of our computational methods by showing agreement with both experimental data and previously published theoretical predictions [29,30,45].

### 3.2. Stability analysis

#### 3.2.1. Thermodynamic stability

The thermodynamical stability of the compound is assessed through the calculation of the formation energy $E_{for}$ using the following relation [46]:

$$E_{for} = \frac{\left[E_{Ca_3AsBr_3} - (3E_{Ca} - E_{As} - 3E_{Br})\right]}{N} \quad (1)$$

where $N$ is the total number of atoms in the unit cell and $E_{Ca_3AsBr_3}$ is the ground state energy of $Ca_3AsBr_3$, $E_{Ca}$, $E_{As}$, and $E_{Br}$ are the total energy per atom for Ca, As, and Br in bulk form, respectively. The calculated $E_{for}$ is determined to be −0.14 eV/atom, showing that the compound is thermodynamically stable.

#### 3.2.2. Mechanical stability

For a cubic system, the mechanical stability is usually analyzed by elastic constants of the material through Born stability criteria: $C_{11}>0$, $C_{44}>0$, $C_{11}-C_{12}>0$, $C_{11}+2C_{12}>0$ [47]. The elastic constants calculated for considered unstrained and strained structures along with other mechanical parameters are mentioned in Table 2. Both unstained and strained $Ca_3AsBr_3$ are found to satisfy the Born stability criteria and hence are mechanically stable.

The bulk modulus ($B$) is obtained by $B=(C_{11}+2C_{12})/3$ and it decreases from 33.99 GPa for the unstrained structure to 23.64 GPa for 3 % strain, whereas for 3 % compressive strain it increases to 47.98 GPa. The shear modulus ($G_h$) is calculated based on the Voigt, Reuss, and Hill approximations, and is given by the relation $G_h=(G_v+G_r)/2$, where $G_v=(C_{11}-C_{12}+3C_{44})/5$ and $Gr=(5(C_{11}-C_{12})\,C_{44})/(4C_{44}+3(C_{11}-C_{12}))$ [48]. The Young's modulus ($E$) determined by $E = 9BG_h/(3B+G_h)$, decreases sharply with strain, i.e., from 75.59 GPa at −3 % to 24.69 GPa at +3 %. The Poisson's ratio ($v$) is represented as $v=(3B-2G_h)/2(3B+G_h)$ its value rises from 0.26 for unstrained material to 0.33 for +3 % strain indicating more lateral expansion upon axial compression with strain. Pugh's ratio ($B/G$) categorizes materials as brittle ($B/G < 1.75$) or ductile ($B/G > 1.75$) and this ratio increased with strain from 1.57 at −3 % to 2.54 at +3 %, indicating a variation from a slightly brittle nature of compressive strains to highly ductile behavior under tensile strains. This ductility is advantageous for TE applications, as it enhances the ability of the material to withstand thermal and mechanical pressures during practical use. The anisotropy factor ($A$) is the degree of mechanical anisotropy in the material and it is found to be 0.41 at −3 % strain and 0.96 at +3 %. We have also calculated the melting temperature for different strains as mentioned in Table 2. It is estimated using the empirical relation: $T_m = 553\ K+(5.91/\text{GPa})\ C_{11} \pm 300$ K, $T_m$ decreases with increasing strain, indicating reduced thermal stability. The experimental value shows that deformation of the $Ca_3AsBr_3$ compound does not happen till 920 K [45], which verifies its stability. The value of calculated mechanical parameters are comparable to previously reported data [30].

#### 3.2.3. Dynamical stability

We have also analyzed the dynamical stability of the $Ca_3AsBr_3$ compound at ambient (Fig. 2a) and for 3 % compressive and tensile strain (Fig. 2c and 2d) by obtaining phonon dispersion along with $\Gamma$-$X$-$M$-$\Gamma$-$R$-$X$ direction in the Brillouin zone (shown in Fig. 1b), calculated using finite displacement approach. The phonon dispersion curves for ±1 % and ±2 % strains are shown in Fig. S2 (see supporting information). The absence of the negative frequencies in the phonon dispersion validates the dynamical stability of the cubic phase of $Ca_3AsBr_3$ at ambient and under applied strain. The unit cell has 7 atoms, so there are 21 total phonon branches, i.e., 3 acoustic and 18 optical phonon modes in the phonon dispersion curve. Higher branches of the dispersion spectra indicate the optical branches that correlate to the frequency range of 2–8 THz, while the lower branches correspond to the acoustic branches with the corresponding frequency range of 0–2 THz. It is observed that the phonon dispersion becomes narrower with the increase in tensile strain, which might lead to enhanced phonon scattering and low phonon group velocity, consequently, the reduced lattice thermal conductivity, compared to the unstrained structure. On the other hand, the opposite trend is observed with increasing compressive strain. Additionally, the phonon density of states (PDOS) as shown in Fig. 2(b), is determined to examine the atomic contributions to the acoustic and optical modes. The acoustic and low-frequency lying optical modes are mainly contributed by the vibrations of Br atoms with a negligible contribution of Ca and As atoms. The high-frequency optical modes are dominated by the vibrations of Ca atoms and the substantial contribution of As and Br atoms. There is no band gap between acoustic and low-frequency optical modes, which might enhance phonon scattering and lead to low lattice thermal conductivity [49].

**Table 3**
Bandgaps for different strains.

| Strain (%) | Band Gap in (eV) | | | |
|---|---|---|---|---|
| | GGA (without SOC) | GGA (with SOC) | TB-mBJ (without SOC) | TB-mBJ (with SOC) |
| −3 | 1.559 | 1.480 | 2.275 | 2.205 |
| −2 | 1.642 | 1.564 | 2.356 | 2.287 |
| −1 | 1.664 | 1.587 | 2.364 | 2.296 |
| 0 | 1.729 | 1.654 | 2.425 | 2.385 |
| 1 | 1.744 | 1.670 | 2.440 | 2.374 |
| 2 | 1.796 | 1.796 | 2.493 | 2.428 |
| 3 | 1.842 | 1.842 | 2.541 | 2.447 |

### 3.3. Electronic properties

The electronic band structures of both unstrained and strained material with and without SOC effect were calculated using both GGA and TB-mBJ to understand the electronic properties of $Ca_3AsBr_3$. The band gap values obtained using these functionals with and without SOC are presented in Table 3, corresponding to all considered structures. The band gap value of unstrained $Ca_3AsBr_3$ calculated using GGA-PBE is 1.65 eV, which is in good agreement with the previously reported value of 1.63 eV by Islam et al. [30]. The electronic band structure of $Ca_3AsBr_3$ with 0 % strain along the high symmetry path ($R$-$\Gamma$-$X$-$M$-$\Gamma$) in the Brillouin zone is shown in Fig. 3. Also, the band structures for both compressive and tensile strains with and without SOC are shown in Fig. S3 and Fig. S4 (see supporting information), respectively. The Fermi energy level was set to 0 eV to make it simpler to determine the bandgap. The band structure for this system revealed that it is a semiconductor with a direct bandgap and remains direct under the impact of all strains. Also, from the electronic structure, it is observed that the bandgap decreases and increases with the increase in the compressive and tensile strain, respectively, which might result in enhanced $ZT$ for the strained structures. On comparing the dispersion curves of +3 % and −3 %, we obtained a more dispersive valence and conduction bands in the vicinity of the Fermi level for the −3 % strain, while relatively flattened bands are obtained for the +3 %. It means a large Seebeck coefficient and enhanced electronic density of states can be observed for tensile strain, and the high carrier mobility and high electrical conductivity for the compressive strain for both types of charge carriers. Further, analysis of

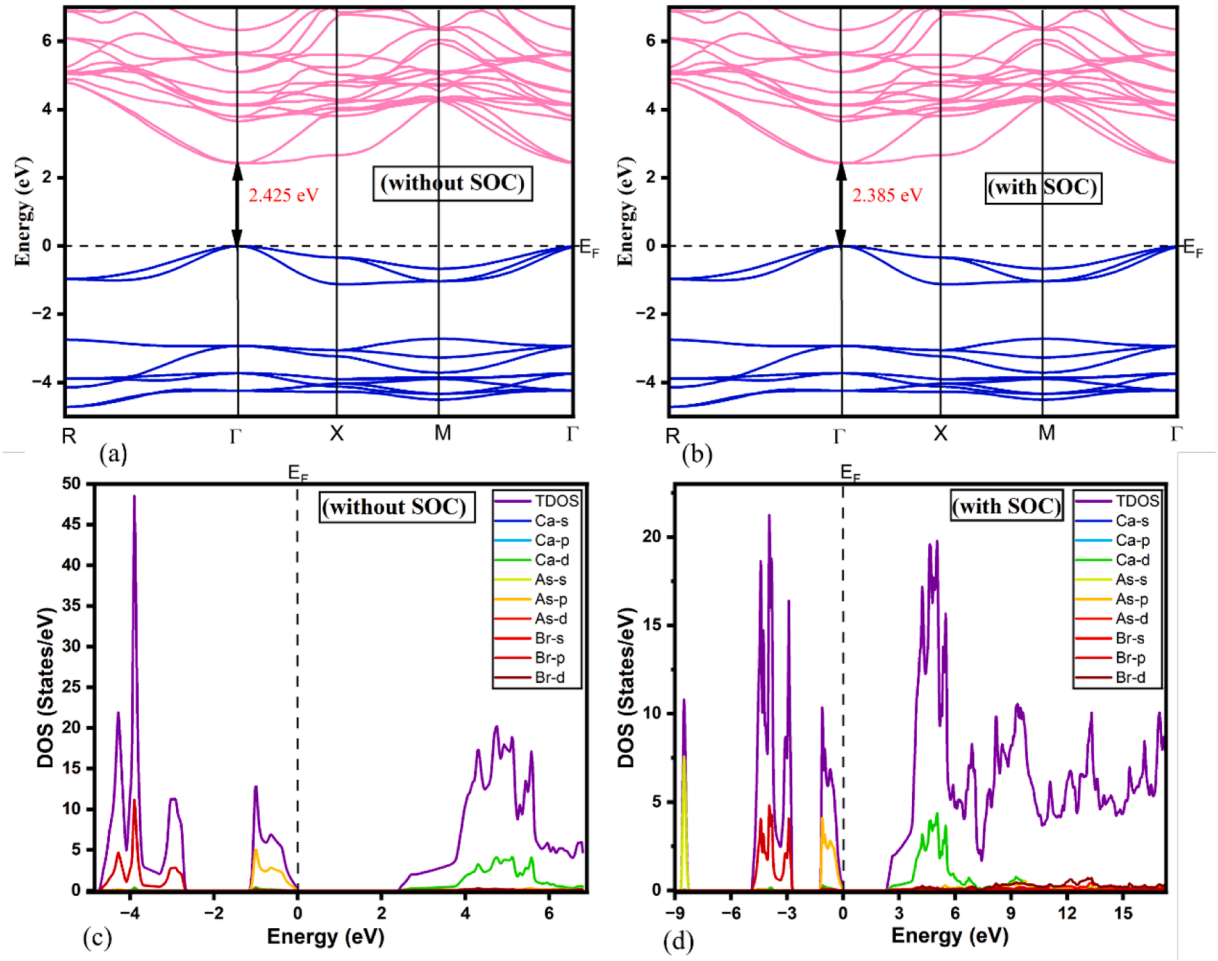



**Fig. 3.** Electronic band structures and density of states of unstrained $Ca_3AsBr_3$ with and without SOC.

the band structure shows that the material has triply degenerate states at Γ-point. This degeneracy affects the TE performance by impacting important parameters such as *S, σ,* etc. The density of states (DOS) was obtained to get deeper insights about the electronic characteristics of the material, as shown bottom panel in Fig. 3 (c and d). The high effective DOS near the Fermi level, contributed by the triply degenerate bands, increases the number of accessible states for charge carriers also leading to higher *S*. The degeneracy also provides multiple paths for carrier transport, reducing scattering and thereby improving the carrier mobility ($\mu$), which directly increases the $\sigma$ of the material. The balance between carrier concentration and *S* is also maintained by this triply degenerate bands, which enhances power factor, which affects *ZT* in a good way. The *p* and *d* orbitals of As atoms contribute most to the valence band (VB), whereas the *d*-orbital of Ca atoms contribute most to the conduction band (CB). The VB maxima show sharp peaks, suggesting the possibility of a high *S*. The 2 eV bandgap in DOS proves that it is ideal case for TE as it allows for selective carrier transport without excessive carrier excitation, maintaining a balance between $\sigma$ and *S*.

### 3.4. Thermoelectric properties

The TE properties of $Ca_3AsBr_3$ were analysed using the semi-classical BTE as implemented in the BoltzTraP2 code [38]. The semiclassical Boltzmann model under the CRTA and RBA for an anisotropic system was used to calculate the transport coefficients (*S, σ*, $\kappa_e$), and the relations used were:

$$S_{\alpha\beta}(\mu,T)=\frac{1}{eT\sigma_{\alpha\beta}(\mu,T)}\int\sigma_{\alpha\beta}(\varepsilon)(\varepsilon-\mu)\left(-\frac{\partial f(T,\varepsilon,\mu)}{\partial\varepsilon}\right)d\varepsilon \tag{2}$$

$$\kappa^{e}_{\alpha\beta}(\mu,T)=\frac{1}{e^2T\Omega}\int\sigma_{\alpha\beta}(\varepsilon)(\varepsilon-\mu)^2\left(-\frac{\partial f(T,\varepsilon,\mu)}{\partial\varepsilon}\right)d\varepsilon \tag{3}$$

$$\sigma_{\alpha\beta}(\mu,T)=\frac{1}{\Omega}\int\sigma_{\alpha\beta}(\varepsilon)\left(-\frac{\partial f(T,\varepsilon,\mu)}{\partial\varepsilon}\right)d\varepsilon \tag{4}$$

where $f(T,\varepsilon,\mu)$ is the Fermi-Dirac function, *e* is the electron charge, Ω is the reciprocal space volume, $\sigma_{\alpha\beta}(\varepsilon)$ is the conductivity tensor, and $\varepsilon$ is the carrier energy, which is expressed as:

$$\sigma_{\alpha\beta}(\varepsilon)=\frac{1}{N}\sum_{i,k}\sigma_{\alpha\beta}(i,k)\frac{\delta(\varepsilon-\varepsilon_{i,k})}{d\varepsilon} \tag{5}$$

From Eqs. (2–4) it can be inferred that the transport coefficients have strong dependence on the carrier concentration and temperature. A positive $\mu$ shows a rise in the Fermi level during *n*-type doping. On the other hand, the Fermi level falls with *p*-type doping, resulting in a negative $\mu$ [50]. Therefore, for both *p*-type and *n*-type doping, we calculated *S, σ*, $\kappa_e$ and associated *PF* and *ZT* as a function of temperature for charge carrier concentrations $\pm\,10^{19}\mathrm{cm}^{-3}$, $\pm\,10^{20}\,\mathrm{cm}^{-3}$, and $\pm\,10^{21}\,\mathrm{cm}^{-3}$, and strains 0 %, ±1 %, ±2 %, and ±3 %.

For degenerate semiconductors, *S* is expressed as [51]:

$$S=\frac{8\pi^2k_B^2}{3eh^2}m^*T\left(\frac{\pi}{3n}\right)^{\frac{2}{3}} \tag{6}$$

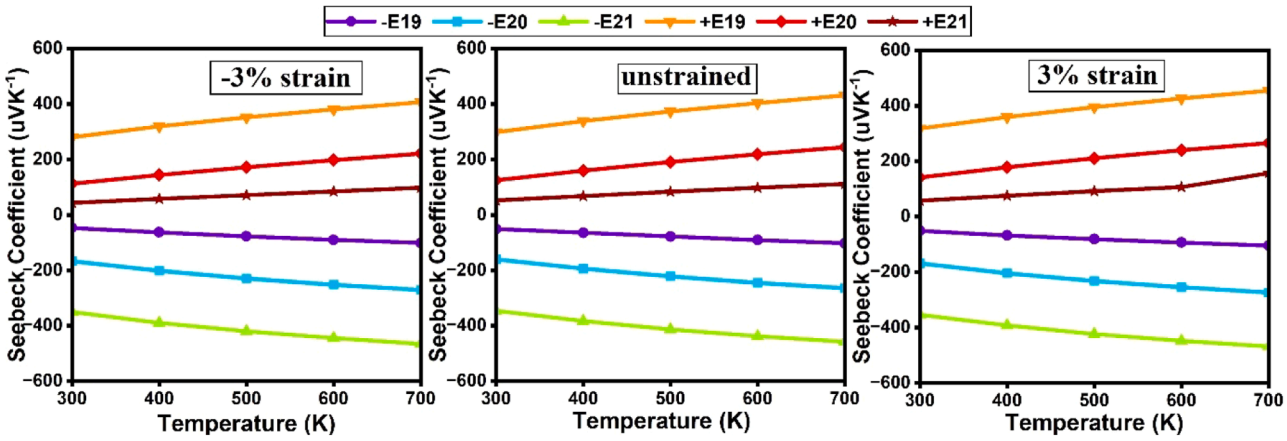



**Fig. 4.** Seebeck coefficient vs temperature for different carrier concentrations for different strains.

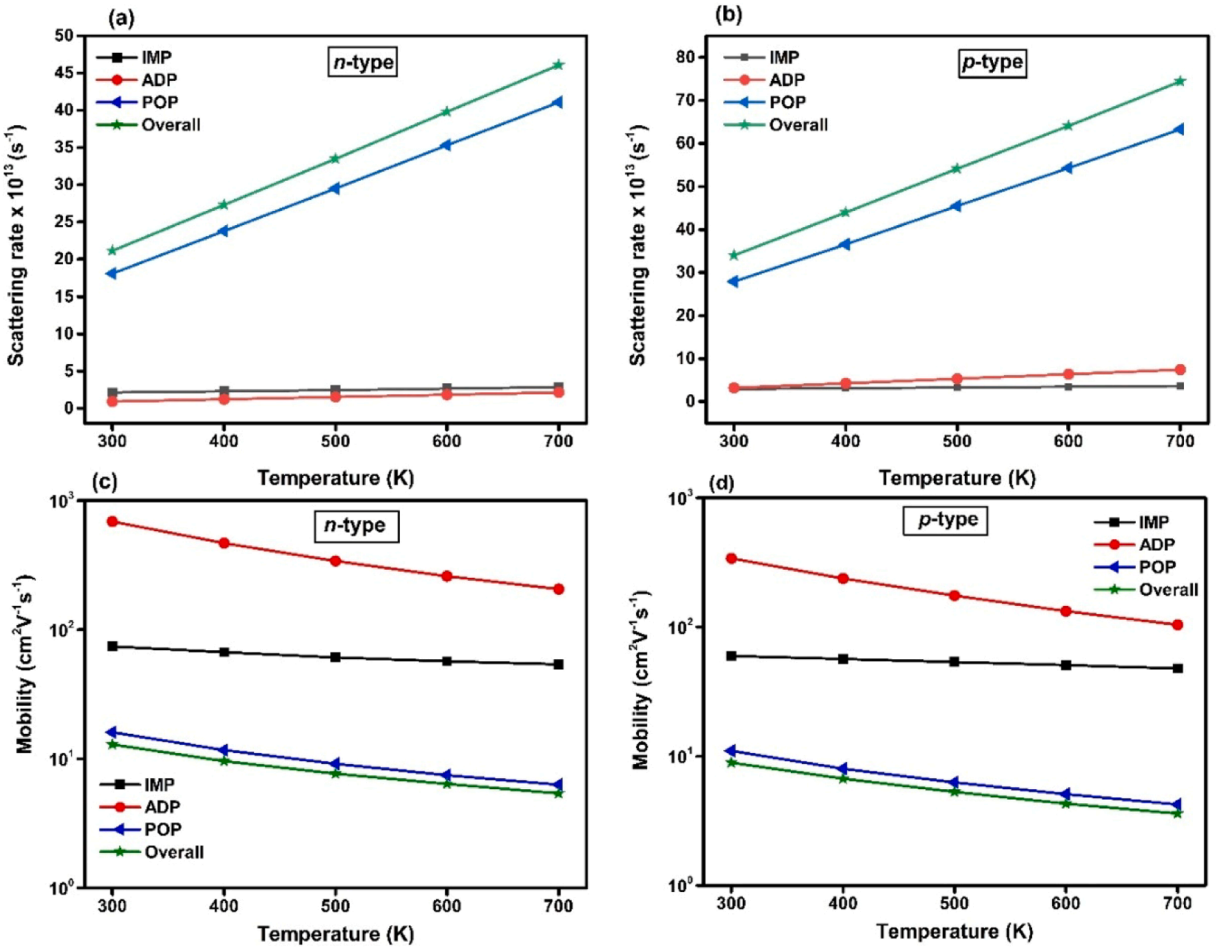



**Fig. 5.** Scattering rates and carrier mobility due to different scatterings by varying the carrier concentration of $Ca_3AsBr_3$ for both *p*-type and *n*-type.

where *n* is the charge carrier concentration, $m^*$ is effective mass, and *T* is the absolute temperature. As shown in Fig. 4 for unstrained and ±3 % strain, at fixed carrier concentration, for both *p*- and *n*-type *S* increased with an increase in temperature. *S* increases with temperature in semiconductors due to the increasing influence of thermally excited carriers. However, at fixed temperature *S,* it decreased with an increase in carrier concentration from $1 \times 10^{19}$ $cm^{-3}$ to $1 \times 10^{21}$ $cm^{-3}$ for both *p*- and *n*-type. *S* behaved in a similar way for ±1, ±2 % strains as shown in Fig. S5 (see supporting information). The maximum value observed of *S* for the unstrained structure is −458.23 $\mu VK^{-1}$ at 700 K, also it is −434.283, −482.366, −464.998, −451.827, −458.449, and −468.289 $\mu VK^{-1}$ for −1, −2, −3, +1, +2, and +3 % strains, respectively, at 700 K for optimal carrier concentrations. A negative sign represents *n*-type semiconducting behaviour. The *S* goes as high as 482 $\mu VK^{-1}$ at compressive strain, comparable to $KInC_3$ with *S* of 500 $\mu VK^{-1}$ [52].

As Boltztrap2 provides $\sigma/\tau$, $\kappa_e/\tau$ where $\tau$ is taken as constant value, so we used AMSET code to find exact carrier (hole/electron) relaxation time and overall mobility by assessing different scattering phenomena in

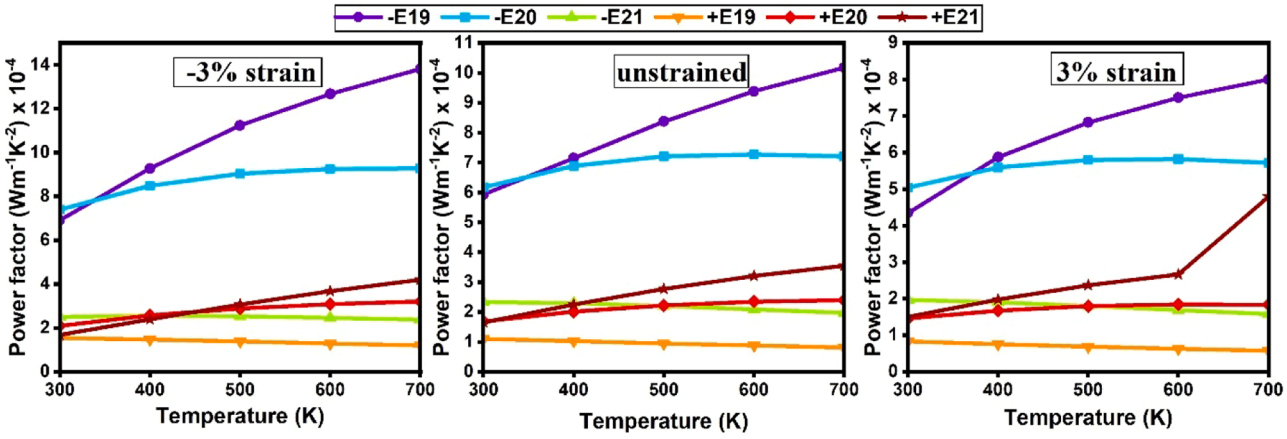



**Fig. 6.** Power factor vs temperature for different carrier concentrations for different strains.

system. We have taken into account the three fundamental carrier scattering rates (ADP, IMP, and POP). The upper panel of Fig. 5 displays the scattering rate across the temperature range (300–700 K), for unstrained material for $1 \times 10^{20}$ $cm^{-3}$ charge carrier concentration for both *n*- and *p*-type. It shows that the POP lifetimes are an order of magnitude larger than ADP and IMP scattering-based lifetimes which shows that POP scattering increased the overall lifespan value. The order of dominance for both conduction- and valence-band edges was POP > ADP > IMP for all strains. Additionally, the lower panel in Fig. 5 shows the mobility of charge carriers at concentration $1 \times 10^{20}$ $cm^{-3}$ for 0 % strained structure, it shows at 300 and 700 K, the values for the total mobility of electrons (holes) are approximately 12.97 (9.05) $cm^2V^{-1}s^{-1}$ and 5.51 (3.74) $cm^2V^{-1}s^{-1}$, respectively. The $\mu$ of the electrons is comparatively greater than that of the holes, confirming that the material could effectively raise the *ZT* for *n*-doping. Furthermore, we see a drop in ADP, POP, and overall mobilities as the temperature rises, this can be explained by the increased quantity of phonons contributing to high scattering. Since temperature often has less of an effect on IMP scattering, their mobility is rather constant over the temperature ranges.

Fig. S6 (see supporting information) illustrates that for 0 %, ±1, ±2 %, and ±3 % strains, $\sigma$ reduces for almost all doping concentrations with temperature. The small electrical conductivity is obtained for the *p*-type in comparison to the *n*-type as a consequence of the flatter valence bands around the Fermi level. The minimum $\sigma$ for each strain was shown by carrier concentration $1 \times 10^{19}$ $cm^{-3}$ for *n*-type doping with values 14.64, 16.20, 17.89, 13.28, 12.20, 11.32, and $9.42 \times 10^4$ $\Omega^{-1}m^{-1}$ for −1, −2, −3, 0, +1, +2, +3 % strain, respectively. The minimum value of $\sigma$ for *p*-type doping with hole concentration of $1 \times 10^{19}$ $cm^{-3}$ is 0.05, 0.06, 0.07, 0.04, 0.03, 0.03, and $0.02 \times 10^4$ $\Omega^{-1}m^{-1}$ for −1, −2, −3, 0, +1, +2, +3 % strain, respectively. The reduced electrical conductivity for tensile strain can be attributed to strain-induced band structure modifications. i.e., less dispersive conduction and valence bands (as depicted in Fig. 3) that

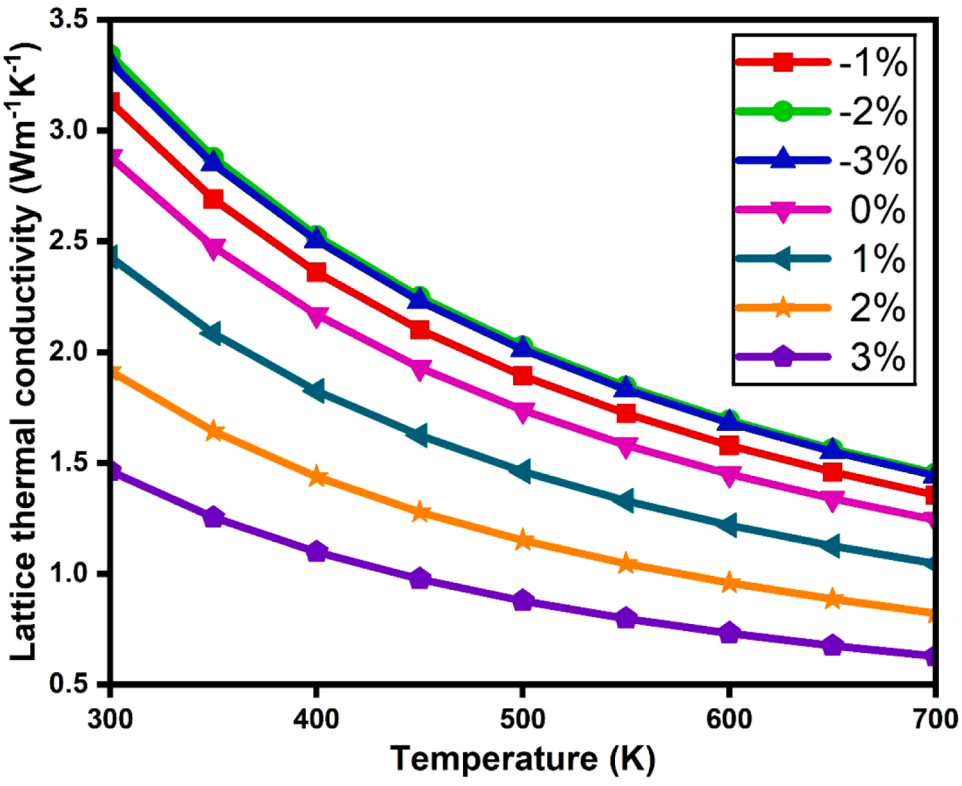



**Fig. 7.** Lattice thermal conductivity for various strains at temperatures 300 to 700 K.

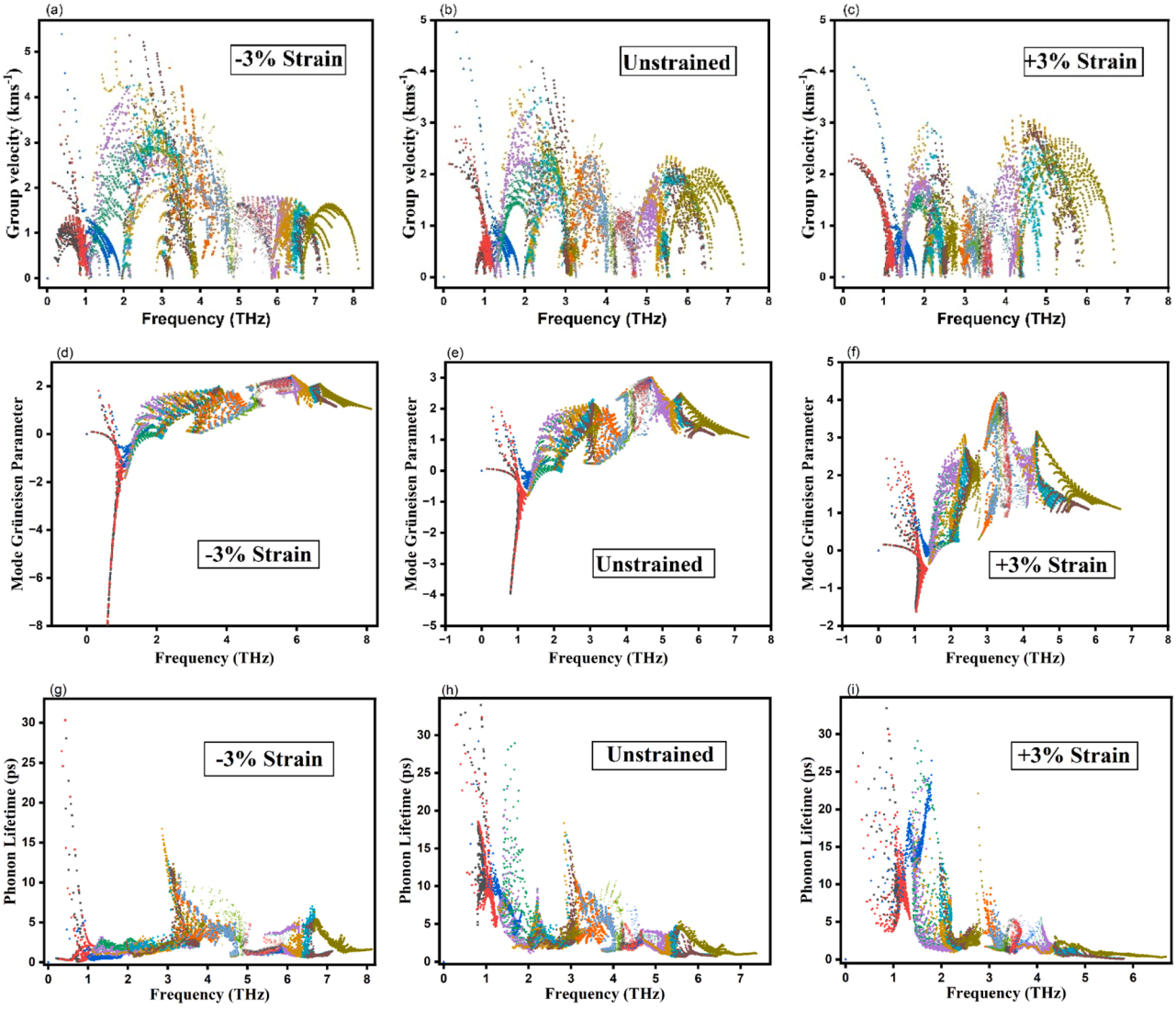



**Fig. 8. (a-c)** Phonon group velocity, **(d-f)** Mode Grüneisen parameter **(g-i)** phonon lifetime corresponding to all phonon modes as a function of frequency for the unstrained and ±3 % strained $Ca_3AsBr_3$.

increase the effective mass of charge carriers and hence decrease their carrier mobilities. The behaviour is still temperature-dependent, and phonon scattering is the leading factor at high temperatures. The $\kappa_e$ for 0 %, ±1, ±2 %, and ±3 % strains are displayed in Fig. S7 (see supporting information). The highest values are shown by carrier concentration of $1 \times 10^{19}$ $cm^{-3}$ for *n*-type doping for each strain. This is an immediate consequence of their higher electrical conductivity as prescribed by the Wiedemann-Franz law which states that $\kappa_e$ and σ are related by formula $\kappa_e = L\ \sigma\ T$ (where *L* is Lorentz number) [53]. The weak temperature dependence implies that electron-phonon scattering is of lesser importance for heat transport in this regime. Furthermore, increased carrier concentrations map to greater $\kappa_e$, reinforcing the importance of charge carriers for electronic thermal conduction. This is important for TE performance because high $\kappa_e$ tends to decrease the *ZT*.

The variation of *PF* for different temperatures at various doping concentrations is shown in Fig. 6 for 0 and ±3 % strains, the rest of the strains are displayed in Fig. S8 (see supporting information). As charge carrier concentration increased from $1 \times 10^{19}$ $cm^{-3}$ to $1 \times 10^{21}$ $cm^{-3}$, *PF* decreased for *n*-type doping. It increased for all temperatures for all carrier concentrations and this behaviour can be analysed on basis of their *S* and σ as $PF = S^2\ \sigma$. The maximum values for −1, −2, −3, 1, 0, 1, 2, and 3 % strains, respectively, are $11.66 \times 10^{-4}$, $13.85 \times 10^{-4}$, $13.81 \times 10^{-4}$, $10.17 \times 10^{-4}$, $9.36 \times 10^{-4}$, $8.95 \times 10^{-4}$, and $8 \times 10^{-4}$ $Wm^{-1}K^{-2}$ at 700 K each. We obtained higher *PF* values for the *n*-type because of high carrier mobility and high electrical conductivity of the electron carriers, whilst the lower *PF* for the *p*-type, corresponding to all considered structures, and this behaviour is consistent with the electronic band structure (as shown in Fig. 3).

Using the single-mode relaxation-time approximation, $\kappa_l$ was calculated as a summation of individual phonon modes $\lambda$ contributions following the equation [54]:

$$\kappa_l = \frac{1}{NV}\sum_{\lambda} C_\lambda v_\lambda \otimes v_\lambda \tau_\lambda \tag{7}$$

where *V* is the volume of the unit cell, and *N* is the number of phonons wavevectors, $v_\lambda \otimes v_\lambda$ are the tensor products of the group velocities, $C_\lambda$are the heat capacities and$\tau_\lambda$ are the lifetimes.

Fig. 7 shows the variation of $\kappa_l$ for each strain with temperature. The value of $\kappa_l$ estimated at room temperature is 2.88 $Wm^{-1}K^{-1}$ for unstrained structure, and it decreased to 1.24 $Wm^{-1}K^{-1}$ at 700 K temperature. For −1 %, −2 %, −3 %, +1 %, +2 %, and +3 % strains, the minimum value of $\kappa_l$ is at 700 K i.e. 1.45, 1.44, 1.24, 1.04, 0.82, and 0.63 $Wm^{-1}K^{-1}$, respectively. This reduction in $\kappa_l$ is advantageous for enhancing the high TE conversion efficiency of the material. Finally, $\kappa$ is obtained by adding

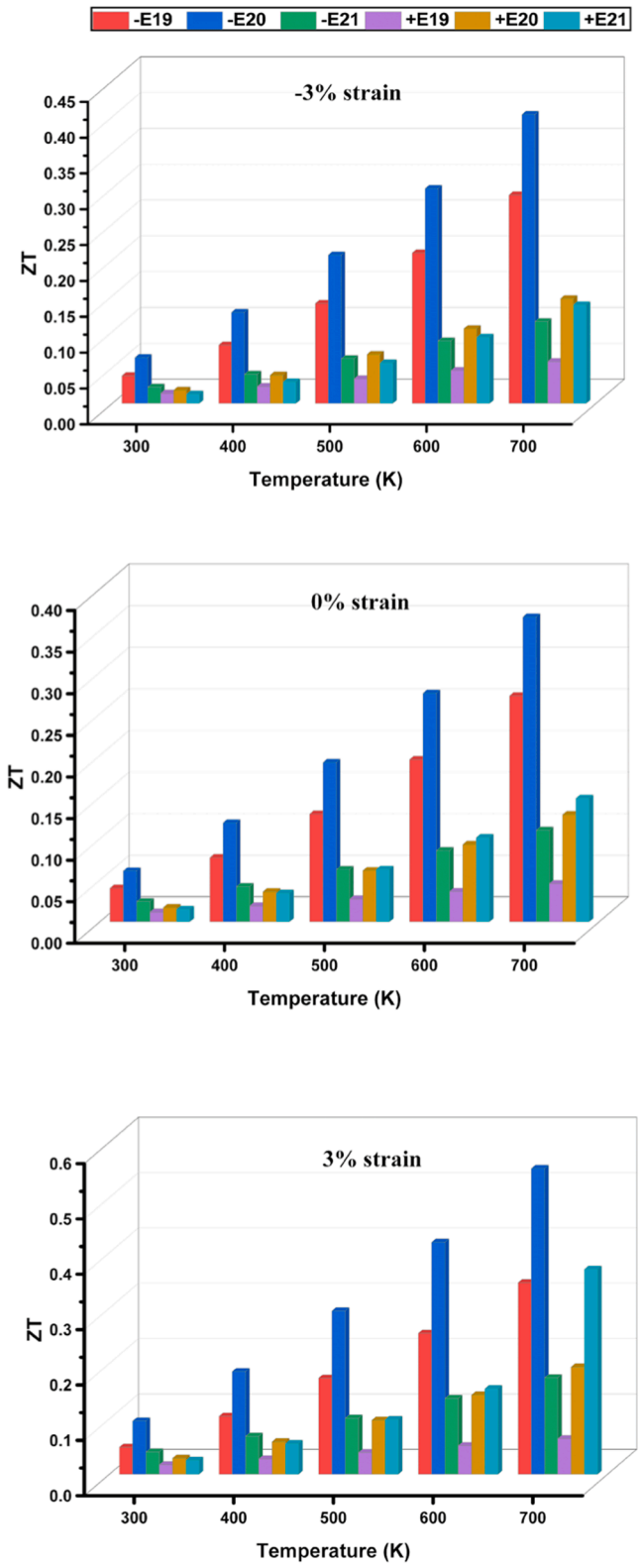



**Fig. 9.** *ZT* value for various strains for 300–700 K temperature for fixed carrier concentrations.

$\kappa_e$ and $\kappa_l$. The calculated value of κ is 0.72 $Wm^{-1}K^{-1}$ at 700 K temperature for *n*-type doping concentration of $10^{20}$ $cm^{-3}$ for 3 % tensile strain.

To gain a deeper understanding of the lattice dynamics that causes a significant strain-induced change in $\kappa_l$, we analyzed the group velocity, phonon lifetime, and Grüneisen parameter (γ) for $Ca_3AsBr_3$ as a function of frequency, as displayed in Fig. 8. The phonon $v_g$ is obtained from the phonon dispersion relations, which are calculated using force constants within harmonic approximation. As expected, and illustrated in Fig. 8 (a)–(c) for −3 %, 0 %, and 3 % strains and for other strains in Fig S9 (see supplementary information), the strain in these systems has a substantial influence on both the $v_g$ of phonons and hence the thermal transport of phonons. Consequently, when the tensile/compressive strain increases, the $v_g$ of phonons are noticeably reduced/increased. Increasing tensile strain results in narrower dispersions for phonon modes, which is depicted in Fig. 2, and weaker bonds. To characterize the values of $v_g$ quantitatively for each strain, the average $v_g$ over the phonon modes is evaluated. The average $v_g$ of $Ca_3AsBr_3$ for −3 %, −2 %, −1 %, 0 %, 1 %, 2 %, and 3 %, respectively, comes out to 1.31, 1.21, 1.12, 1.04, 0.97, 0.92, and 0.89 ($kms^{-1}$). Fig 8(d)-(f) shows the mode γ for unstrained and ±3 %, whereas γ for other strains are represented by Fig. S10 (see supplementary information). The average value of γ for unstrained $Ca_3AsBr_3$ is 1.25 while on increasing the strain, the average γ rises from 0.57 to 1.68 for −3 % to +3 % strained structure, indicating enhanced lattice anharmonicity in the case of tensile strain. A higher γ implies stronger phonon–phonon scattering, which directly contributes to the observed reduction in phonon lifetime and thermal conductivity. Next, we considered the phonon lifetime to analyse the phonon-phonon scattering, as shown in Fig. 8(g)–(i) and in Fig. S11 (see supplementary information) at 300 K. It exhibits a consistent trend, a significant rise in phonon scattering is observed as the strain increases. Also, based on the distribution of phonon lifetimes, it is evident that the primary factor influencing $\kappa_l$ is the acoustic branches, whereas the optical mode has a lesser impact. For all structures at a low frequency, we can see phonon lifetime reaches up to 30 ps. At mid to high frequencies, phonon lifetimes are shorter and clustered to a few picoseconds, indicating higher scattering. Therefore, the combined effects of reduced group velocity, shortened phonon lifetime, and increased anharmonicity with tensile strain comprehensively explain the decreased lattice thermal conductivity in strained $Ca_3AsBr_3$.

The TE figure of merit (*ZT*) is a significant parameter used to evaluate the efficiency of a material for power generation from heat. The higher the value of *ZT*, better the TE efficiency, and thus by determining it, the influence of strain, doping and temperature on the TE performance of material were analysed.

Fig. 9 shows that for 0 and ±3 % strain for each temperature *ZT* is highest at $1 \times 10^{20}$ $cm^{-3}$ for *n*-type doping, a similar trend is followed by ±1 and ±2 % strains as shown in Fig. S12 (see supporting information). The maximum values for −1, −2, −3, 0, 1, 2, and 3 % strains are 0.34, 0.4, 0.4, 0.36, 0.38, 0.45, and 0.55, respectively at 700 K. Temperature and charge carrier concentration have an impact on the *ZT* value. The relationship between the *S* and carrier concentration is inverse for both electrical and thermal conductivity. As a result, it is not simply an assumption that raising the carrier concentration will result in a higher *ZT* value. This is due to the fact that while the $\sigma$ and $\kappa_e$ both rise with increasing carrier concentration, the *S* falls. Overall, these coupled parameters are optimized in such a way that *ZT* is maximized. Our research revealed strained $Ca_3AsBr_3$ with a peak ZT of 0.55 at 700 K, being as high as of double perovskite $Cs_2AuYCl_6$ [55]. These findings indicate that strain engineering in $Ca_3AsBr_3$ can efficiently improve TE performance.

## 4. Conclusion

In this study, we have examined the thermoelectric properties of halide perovskite $Ca_3AsBr_3$ for various strains at different temperatures for different carrier concentrations using density functional theory and the Boltzmann transport equation. The obtained structural parameters aligned well with the available experimental and theoretical values. We have verified mechanical, thermodynamic, and dynamical stability by calculating elastic constants, formation energy, and analysing the phonon dispersion relations, respectively. The electronic band

structures showed that it is a semiconductor with a direct bandgap. The strains modified the electronic band structure and improved the mobility of the charge carrier and hence the PF. We have found that the isotropic tensile strain significantly reduces lattice thermal conductivity ($\kappa_l$) by decreasing phonon group velocities, shortening phonon lifetimes, and enhancing anharmonicity, as reflected in the Grüneisen parameter. For 3 % strain, a $\kappa_l$ is found as 0.63 $Wm^{-1}K^{-1}$ and a peak value of *ZT*, i.e., 0.55, is obtained at 700 K for the *n*-type doping concentration of $10^{20}$ $cm^{-3}$. The *ZT* can be further enhanced by decreasing the bandgap by introducing vacancies or interstitials that can create defects, or by doping of some alloy or other element that can introduce more energy states or by further varying carrier concentration. Also, material can be nanostructured, which not only reduce the bandgap but also reduce lattice thermal conductivity. The electron and hole carrier concentrations explored in the present work are limited to a few discrete values. A further fine tuning of carrier concentration may result in greater improvement of *PF* and *ZT*. Moreover, the specific strain and carrier concentration is challenging to realize experimentally. Therefore, a further research on the proposed halide perovskite may be conducted to investigate the combined effects of defects, explore anisotropic strain conditions (e.g., uniaxial or biaxial strain), and consider additional band structure engineering strategies like alloying or doping. We believe that the present study will pave the way for future experimental thermoelectric energy harvesting studies on this material.

## CRediT authorship contribution statement

**Sumedha Yadav:** Writing – original draft, Investigation, Data curation, Conceptualization. **Sangeeta:** Writing – review & editing, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Kulwinder Kumar:** Software, Formal analysis, Data curation. **Mukhtiyar Singh:** Writing – review & editing, Validation, Supervision, Software, Formal analysis, Conceptualization.

## 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.

## Acknowledgments

One of the authors, Sangeeta would like to acknowledge the financial support provided by Delhi Technological University, Delhi. All the authors acknowledge the National Supercomputing Mission (NSM) for providing computing resources of 'PARAM SMRITI' at NABI, Mohali, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and the Department of Science and Technology (DST), Government of India.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.physleta.2025.130625.

## Data availability

Data will be made available on request.

## References

[1] Md.Ferdous Rahman, P. Barman, Md.Azizur Rahman, M. Mushtaq, Md. Rasidul Islam, Md.Atikur Rahman, Md.Z. Sultan, A.R. Chaudhry, A. Irfan, Electronic, optical, thermophysical, and mechanical properties of lead-free Ba3SbBr3 perovskite, Polyhedron. 254 (2024) 116937, https://doi.org/10.1016/j.poly.2024.116937.

[2] T. Mori, S. Priya, Materials for energy harvesting: at the forefront of a new wave, MRS Bull. 43 (2018) 176–180, https://doi.org/10.1557/mrs.2018.32.

[3] C. Tong, Emerging materials for energy harvesting. Introduction to Materials for Advanced Energy Systems, Springer International Publishing, Cham, 2019, pp. 719–817, https://doi.org/10.1007/978-3-319-98002-7_11.

[4] M. Massetti, F. Jiao, A.J. Ferguson, D. Zhao, K. Wijeratne, A. Würger, J. L. Blackburn, X. Crispin, S. Fabiano, Unconventional thermoelectric materials for energy harvesting and sensing applications, Chem. Rev. 121 (2021) 12465–12547, https://doi.org/10.1021/acs.chemrev.1c00218.

[5] X. Zhu, Y. Yu, F. Li, A review on thermoelectric energy harvesting from asphalt pavement: configuration, performance and future, Constr. Build. Mater. 228 (2019) 116818, https://doi.org/10.1016/j.conbuildmat.2019.116818.

[6] T.M. Tritt, Thermoelectric phenomena, materials, and applications, Annu Rev. Mater. Res. 41 (2011) 433–448, https://doi.org/10.1146/annurev-matsci-062910-100453.

[7] G. Tan, M. Ohta, M.G. Kanatzidis, Thermoelectric power generation: from new materials to devices, Phil. Trans. R. Soc. A. 377 (2019) 20180450, https://doi.org/10.1098/rsta.2018.0450.

[8] G.J. Snyder, E.S. Toberer, Complex thermoelectric materials, Nature Mater. 7 (2008) 105–114, https://doi.org/10.1038/nmat2090.

[9] T. Zhu, Y. Liu, C. Fu, J.P. Heremans, J.G. Snyder, X. Zhao, Compromise and synergy in high-efficiency thermoelectric materials, Adv. Mater. 29 (2017) 1605884, https://doi.org/10.1002/adma.201605884.

[10] Sangeeta, M. Singh, Computational study of the thermoelectric properties and lattice dynamics of Li2MN2 (M = Zr or Hf), Mater. Res. Bull. 172 (2024) 112650, https://doi.org/10.1016/j.materresbull.2023.112650.

[11] R. Kumar, R. Kumar, A. Vij, M. Singh, A first-principle study of electronic, thermoelectric, and optical properties of sulfur doped c-HfO $_2$,, Phys. Scr. 97 (2022) 075813, https://doi.org/10.1088/1402-4896/ac7678.

[12] Sangeeta, R. Kumar, M. Singh, In-silico realization of YX (X = N, P, As) pnictide monolayers as highly efficient thermoelectric materials, Surf. Interf. 55 (2024) 105442, https://doi.org/10.1016/j.surfin.2024.105442.

[13] Md.N. Hasan, H. Wahid, N. Nayan, M.S. Mohamed Ali, Inorganic thermoelectric materials: a review, Int. J. Energy Res. 44 (2020) 6170–6222, https://doi.org/10.1002/er.5313.

[14] H. Wang, W. Su, J. Liu, C. Wang, Recent development of n-type perovskite thermoelectrics, J. Materiomics 2 (2016) 225–236, https://doi.org/10.1016/j.jmat.2016.06.005.

[15] Sangeeta, R. Kumar, M. Singh, Realizing high thermoelectric performance in p-type RbZn4P3 zintl compound: a first-principles investigation, J. Mater. Sci. 57 (2022) 10691–10701, https://doi.org/10.1007/s10853-022-06953-y.

[16] Sangeeta, M. Singh, Augmented thermoelectric performance of LiCaX (X = As, Sb) half Heusler compounds via carrier concentration optimization, J. Phys. Chem. Solids 174 (2023) 111182, https://doi.org/10.1016/j.jpcs.2022.111182.

[17] A. Ghosh, F. Ahmed, Md.J. Ferdous, Mst.M.J. Juhi, M.F.I. Buian, A.A. Miazee, M. Sajid, Md. Maniruzzaman, A.M. Tighezza, M.F. Ahmmed, Md.S. Islam, Strain-induced changes in the electronic, optical and mechanical properties of the inorganic cubic halide perovskite Sr3PBr3 with FP-DFT, J. Phys. Chem. Solids 191 (2024) 112053, https://doi.org/10.1016/j.jpcs.2024.112053.

[18] D. Behera, D.R. Lawati, M. Agouri, A. Abbassi, S. Taj, B. Manaut, S.K. Mukherjee, A DFT insight into the physical features of alkaline based perovskite compounds AInBr3 (A = K, Rb), Solid State Ionics 409 (2024) 116513, https://doi.org/10.1016/j.ssi.2024.116513.

[19] B. Akenoun, S. Dahbi, N. Tahiri, O. El Bounagui, H. Ez-Zahraouy, Engineering the optoelectronic and thermoelectric properties of Cs2BiAgY6 (Y= Br or Cl) double perovskites through doping with iodine: a DFT study, J. Phys. Chem. Solids 194 (2024) 112229, https://doi.org/10.1016/j.jpcs.2024.112229.

[20] G. Murtaza, I. Ahmad, B. Amin, A. Afaq, M. Maqbool, J. Maqssod, I. Khan, M. Zahid, Investigation of structural and optoelectronic properties of BaThO3, Opt. Mater. 33 (2011) 553–557, https://doi.org/10.1016/j.optmat.2010.10.052.

[21] M.L. Scullin, C. Yu, M. Huijben, S. Mukerjee, J. Seidel, Q. Zhan, J. Moore, A. Majumdar, R. Ramesh, Anomalously large measured thermoelectric power factor in Sr1−xLaxTiO3 thin films due to SrTiO3 substrate reduction, Appl. Phys. Lett. 92 (2008) 202113, https://doi.org/10.1063/1.2916690.

[22] J.M.D. Coey, M. Viret, S. Von Molnár, Mixed-valence manganites, Adv. Phys. 48 (1999) 167–293, https://doi.org/10.1080/000187399243455.

[23] A. Ali, A.U. Rahman, G. Rahman, Thermoelectric properties of KCaF3, physica B, Condens. Matter. 565 (2019) 18–24, https://doi.org/10.1016/j.physb.2019.04.019.

[24] S.M. AL-Shomar, M. Liaqat, I. Anmol, A.M. Quraishi, I. Khan, Amina, K. Arshad, U. Turdialiyev, A. Almahri, V. Tirth, A. Algahtani, A.M. Alsuhaibani, Abdullah, M. S. Refat, A. Zaman, Study of structural, electronic, mechanical, optical and thermoelectric properties of As based halide-perovskites Ba3AsX3 (X= F, Cl): a first-principles insights, J. Mater. Res. Technol. 31 (2024) 2450–2460, https://doi.org/10.1016/j.jmrt.2024.06.218.

[25] Md.F. Rahman, Md.H. Rahman, Md.R. Islam, M.K. Hossain, A. Ghosh, Md.S. Islam, Md.M. Islam, Md. Harun-Or-Rashid, H. Albalawi, Q. Mahmood, The optical and electronic properties of inorganic halide perovskite Sr3NCl3 under applied biaxial strain, J. Mater. Sci. 58 (2023) 13100–13117, https://doi.org/10.1007/s10853-023-08825-5.

[26] A. Algahtani, Amina, F. Rehman, M. Liaqat, N. Juraev, I. Khan, A.M. Alsuhaibani, Abdullah, V. Tirth, M.S. Refat, A. Zaman, Probing the physical properties of Sr3AsX3 (X = F and Br) perovskite compounds for prospective solar cell applications employing the DFT framework, Inorg. Chem. Commun. 162 (2024) 112186, https://doi.org/10.1016/j.inoche.2024.112186.

[27] I.K.G.G. Apurba, Md.R. Islam, Md.S. Rahman, Md.F. Rahman, J. Park, Tuning the physical properties of inorganic novel perovskite materials Ca3PX3 (X=I, Br and Cl): density function theory, Heliyon. 10 (2024) e29144, https://doi.org/10.1016/j.heliyon.2024.e29144.
[28] R. Tranås, O.M. Løvvik, K. Berland, Lattice thermal conductivity from first principles and active learning with gaussian process regression, (2023). https://doi.org/10.48550/ARXIV.2309.06786.
[29] C. Hadenfeldt, P. Schulz, Darstellung, struktur und temperaturabhängigkeit der phasenbreite der phase $Ca_{2-x}$ $As_{1-x}$ $Br_{1+x}$ und thermisches verhalten der verbindung $Ca_3$ $AsBr_3$, Zeitschrift Anorg Allge Chemie 518 (1984) 77–86, https://doi.org/10.1002/zaac.19845181108.
[30] Md.A. Ul Islam, O. Das, D.B. Khadka, Md.R. Islam, Md.F. Rahman, S. Kato, T. Soga, Effect of low to high pressure on the structural, mechanical, electrical, and optical properties of inorganic material $Ca_3$ $AsBr_3$ : an ab initio investigation, ACS Omega (2024), https://doi.org/10.1021/acsomega.3c08131 acsomega.3c08131.
[31] M.F.I. Buian, Most.S.Islam Ria, A. Ghosh, Md.A. Rahman, M.S. Hossain, M. F. Ahmmed, A.A. Miazee, M.M.R. Sonic, H.A. Alrafai, A novel investigation into strain-induced changes in the physical properties and solar cell performances of lead-free Ca3NCl3 perovskite, Mater. Mater. Sci. Semicond. Process. 180 (2024) 108580, https://doi.org/10.1016/j.mssp.2024.108580.
[32] Md.F. Rahman, Md.A. Rahman, Md.R. Islam, A. Ghosh, Md.A.Bashar Shanto, M. Chowdhury, Md.Alljajul Islam, Md.H. Rahman, M.K. Hossain, M.A. Islam, Unraveling the strain-induced and spin–orbit coupling effect of novel inorganic halide perovskites of Ca3AsI3 using DFT, AIP. Adv. 13 (2023) 085329, https://doi.org/10.1063/5.0156961.
[33] A. Yadav, S. Kumar, M. Muruganathan, R. Kumar, Thermoelectric properties of half heusler topological semi-metal LiAuTe, EPL 132 (2020) 67003, https://doi.org/10.1209/0295-5075/132/67003.
[34] P. Blaha, K. Schwarz, F. Tran, R. Laskowski, G.K.H. Madsen, L.D. Marks, WIEN2k: an APW+lo program for calculating the properties of solids, J. Chem. Phys. 152 (2020) 074101, https://doi.org/10.1063/1.5143061.
[35] K. Schwarz, P. Blaha, G.K.H. Madsen, Electronic structure calculations of solids using the WIEN2k package for material sciences, Comput. Phys. Comput. Phys. Commun. 147 (2002) 71–76, https://doi.org/10.1016/S0010-4655(02)00206-0.
[36] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865–3868, https://doi.org/10.1103/PhysRevLett.77.3865.
[37] F. Tran, P. Blaha, Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential, Phys. Rev. Lett. 102 (2009) 226401, https://doi.org/10.1103/PhysRevLett.102.226401.
[38] G.K.H. Madsen, J. Carrete, M.J. Verstraete, BoltzTraP2, a program for interpolating band structures and calculating semi-classical transport coefficients, Comput. Phys. Comput. Phys. Commun. 231 (2018) 140–145, https://doi.org/10.1016/j.cpc.2018.05.010.
[39] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6 (1996) 15–50, https://doi.org/10.1016/0927-0256(96)00008-0.
[40] A.M. Ganose, J. Park, A. Faghaninia, R. Woods-Robinson, K.A. Persson, A. Jain, Efficient calculation of carrier scattering rates from first principles, Nat. Commun. 12 (2021) 2222, https://doi.org/10.1038/s41467-021-22440-5.
[41] S. Ramawat, S. Kukreti, A. Dixit, β − SrZrS 3 : a superior intermediate temperature thermoelectric through complex band geometry and ultralow lattice thermal conductivity, Phys. Rev. Mater. 7 (2023) 085403, https://doi.org/10.1103/PhysRevMaterials.7.085403.
[42] A. Togo, First-principles phonon calculations with phonopy and Phono3py, J. Phys. Soc. Jpn. 92 (2023) 012001, https://doi.org/10.7566/JPSJ.92.012001.
[43] A. Gilani, Nasarullah, S.A. Aldaghfag, M. Yaseen, Magnetic, optical, electronic, structural and thermoelectric features of A2MgS4 (A=Sm, Tb): a DFT study, Journal of Physics (College Park Md) and Chemistry (Easton) of Solids 175 (2023) 111208, https://doi.org/10.1016/j.jpcs.2022.111208.
[44] T. Katsura, Y. Tange, A simple derivation of the Birch–Murnaghan equations of State (EOSs) and comparison with EOSs derived from other definitions of finite strain, Minerals 9 (2019) 745, https://doi.org/10.3390/min9120745.
[45] C. Hadenfeldt, Darstellung und eigenschaften von $Ca_2$ AsBr, $Sr_2$ AsCl und $Ca_3$ $AsBr_3$ .: preparation and properties of $Ca_2$ AsBr, $Sr_2$ AsCl, and $Ca_3$ $AsBr_3$, Zeitschrift Für Naturforschung B 31 (1976) 408–410, https://doi.org/10.1515/znb-1976-0403.
[46] M. Mushtaq, M.A. Sattar, S.A. Dar, Phonon phase stability, structural, mechanical, electronic, and thermoelectric properties of two new semiconducting quaternary Heusler alloys CoCuZnZ (Z = Ge and Sn), Int. J. Energy Res. 44 (2020) 5936–5946, https://doi.org/10.1002/er.5373.
[47] I. Waller, *Dynamical theory of crystal lattices* by M. Born and K. Huang, Acta Cryst. 9 (1956) 837–838, https://doi.org/10.1107/S0365110X56002370.
[48] R. Hill, The elastic behaviour of a crystalline aggregate, Proc. Phys. Soc. A 65 (1952) 349–354, https://doi.org/10.1088/0370-1298/65/5/307.
[49] K.S. Rana, R. Gupta, D. Sarkar, N.K. Singh, S. Acharya, S. Vitta, C. Bera, K. Biswas, A. Soni, Interaction of acoustic and optical phonons in a soft-bonded Cu-Se framework of large unit cell minerals with anionic disorders, Phys. Rev. B 108 (2023) 045202, https://doi.org/10.1103/PhysRevB.108.045202.
[50] H.S. Patel, V.A. Dabhi, A.M. Vora, Advancing thermoelectric potential: strontium telluride under compression strain analyzed with HSE hybrid functional and Wannier interpolation, Sci. Rep. 14 (2024) 20700, https://doi.org/10.1038/s41598-024-55519-2.
[51] M. Dutta, T. Ghosh, K. Biswas, Electronic structure modulation strategies in high-performance thermoelectrics, APL. Mater. 8 (2020) 040910, https://doi.org/10.1063/5.0002129.
[52] D. Behera, T.A. Geleta, I. Allaoui, M. Khuili, S.K. Mukherjee, B. Akila, S. Al-Qaisi, First-principle analysis of optical and thermoelectric properties in alkaline-based perovskite compounds AInCl3 (A = K, Rb), Eur. Phys. J. Plus. 139 (2024) 127, https://doi.org/10.1140/epjp/s13360-024-04921-w.
[53] M. Jonson, G.D. Mahan, Mott's formula for the thermopower and the Wiedemann-Franz law, Phys. Rev. B 21 (1980) 4223–4229, https://doi.org/10.1103/PhysRevB.21.4223.
[54] A. Togo, L. Chaput, I. Tanaka, Distributions of phonon lifetimes in Brillouin zones, Phys. Rev. B 91 (2015) 094306, https://doi.org/10.1103/PhysRevB.91.094306.
[55] S. Mahmud, M.A. Ali, M.M. Hossain, M.M. Uddin, DFT mediated X2AuYZ6 (X= Cs, Rb; Z= Cl, Br, I) double perovskites for photovoltaic and wasted heat management device applications, (2024). https://doi.org/10.48550/ARXIV.2404.15693.

# Performance Analysis of a Dual FSO/RF Hybrid Link in the Presence of Malicious Intruder

Himani Mehta
*Delhi Technological University, India*
himanimehta_2k21phdec16@dtu.ac.in

Vishal Kumar
*Delhi Technological University, India*
vishalkumar_2k21phdec07@dtu.ac.in

Rohit Kumar
*Delhi Technological University, India*
rohit.kumar@dtu.ac.in

***Abstract*—One of the fundamental issues in the field of wireless communication is security. To overcome this problem, hybrid radio frequency (RF)/free space optical (FSO) link with the presence of an eavesdropper present are modelled and an algorithm is proposed to select the most secure link based on the derived security metrics. The FSO links are modelled using exponential distribution considering different parameter values, thus exhibiting a strong atmospheric turbulent nature of the environment. Divergent environments and channel conditions tend to make the RF and FSO channels behave uniquely and complementary to each other. To attain high data rates along with higher reliability, RF and FSO links are incorporated together to form a hybrid FSO/RF communication system. The proposed algorithm is based on a security metric which is determined by computing the probability of strictly positive secrecy capacity (SPSC). Analytical closed-form expressions of SPSC have been obtained for the proposed system. Further, monte carlo simulations are performed to validate and confer the analytical results.**

***Index Terms*—Hybrid FSO/RF, Free Space Optical, Secrecy Capacity, Strictly Positive Secrecy Capacity**

## I. INTRODUCTION

WITH the growing demands for high-speed data, a lot of efforts have been made in the field of wireless communication to meet user's data requests. Online activities such as online teaching, work from home, massive multiplayer online games, development in the field of web of things, 5G and 6G networks have resulted in laying a huge amount of stress on the radio frequency (RF) spectrum. Along with high data rates, the other major challenge of RF is to maintain proper security, as it is immensely susceptible to wiretapping. The best alternative is to take advantage of the FSO link to subdue the burden of spectrum scarcity and enhance security [1]. FSO is considered secure, but it is highly affected by atmospheric turbulence and misalignment in order to maintain a line of sight (LOS) link. To utilize the best possible benefits of both FSO and RF systems, a hybrid FSO/RF system is proposed. A hybrid FSO/RF system uses the advantages of both FSO and RF technologies to provide a reliable, effective, and adaptable communication system. For applications needing great dependability, scalability, and adaptability, this makes them the perfect option.

The outflow of an enormous amount of critical data takes place through the RF channel when compared to the highly secure FSO channel [2]. Wyner first proposed the physical layer security (PLS) concept to effectively exploit the characteristics of the physical layer to enhance the dependability and safety of a transmission system [3]. The concept of a memoryless discrete wiretap channel was introduced and its achievable inherent secrecy capacity (SC) rate was analyzed. In [4], secrecy outage probability (SOP) and SC of a hybrid FSO/RF model is proposed. Here, a FSO/RF hybrid link is considered and it's performance is evaluated in the presence of a malicious intruder modeled using Nakagami distribution.

In order to enhance the security of a communication system, joint cooperative beamforming and other techniques in the field of jamming have been proposed [5]. The effect of an eavesdropper (Eve) on the RF link in a hybrid FSO/RF system is analyzed in [6]. Authors in [7] have studied the concept of PLS and derived the exact expressions for average secrecy capacity (ASC) and SOP of a selection combining-based hybrid FSO/RF system.

Due to the highly focused nature of laser beams, FSO is regarded as a communication link less susceptible to eavesdropping compared to RF links [8]. FSO communication system is secure for small divergence angles as a smaller divergence angle makes the divergence diameter smaller. The narrow diameter of divergence signifies a narrow range to be eavesdropped at the receiving side [9]. Moreover, for the FSO link to be wiretapped, the eavesdropper must be placed closely in such a way that it blocks the laser beam [10]. In this work, the RF eavesdropper is static and placed at a distance such that it does not obstruct the laser beam. Several works have been done on link selection in wireless communication to enhance the performance of the system. In [11], the best link among the FSO subsystems is used for transmission in which the FSO link's aperture having the highest gain is treated as the best link. When the FSO subsystem is in an outage, the RF link is used for communication. Similarly, authors in [12] have proposed a selection scheme for the best link based on the signal-to-noise ratio (SNR) and further derived the expressions for symbol error rate and outage probability.

Security is one of the most important aspect in the field of wireless communication. Thus, it is crucial that communication should take place through the most secure and reliable link. In

2025 National Conference on Communications (NCC) | 979-8-3315-3601-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/NCC63735.2025.10983482

979-8-3315-3601-5/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 06:33:43 UTC from IEEE Xplore. Restrictions apply.

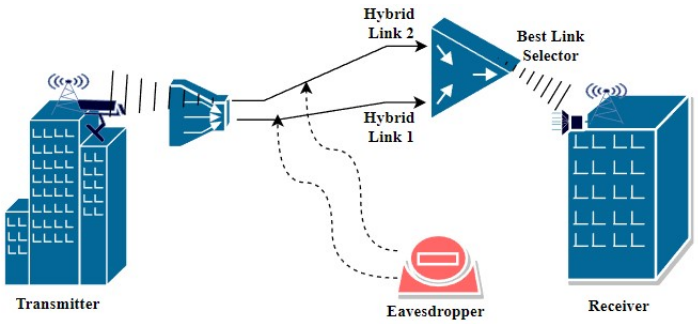



Fig. 1: Proposed system model

military access and critical confidential research fields, it is of utmost importance to choose the best secured path for data transfer. Thus, in this work, we have proposed dual hybrid links and an algorithm that selects the best secured link on the basis of the probability of strictly positive secrecy capacity (SPSC). Our proposed work can be practically deployed in areas demanding higher security. To the best of our apprehension, none of the previous works exists in the literature that selects the best hybrid link based on the security metric. Hence, this work is the first of its own kind. The major contributions of our work are listed below:

- To improve the security of the communication system, a dual hybrid FSO/RF communication system in the whereabouts of a malicious intruder is proposed.
- The cumulative distribution function (CDF) and probability density function (PDF) for both hybrid links have been derived in terms of SNR.
- The probability of SPSC for both hybrid links is expressed in closed form.
- Further, an algorithm is proposed based on the probability of SPSC, which selects the most secure link for transmitting data and results in an enhanced security of the overall system.
- Along with the mathematical expressions, monte carlo simulations are performed to validate and confer the analytical results.

Remaining paper is sort out as follows. System and channel model is proposed in Section II. In Section III, performance analysis and algorithm is discussed. Results were considered in the Section IV and at last Section V concludes the paper.

## II. SYSTEM AND CHANNEL MODEL

A network model is considered in which a transmitter sends confidential information that is intercepted by a malicious intruder before reaching its intended destination. Fig. 1 depicts the basic schematic representation of the proposed system model. It consists of two hybrid links which are affected by a static RF eavesdropper and placed at a distance such that it does not obstruct the laser beam.

### A. Hybrid Link 1

*1) RF Channel:* The RF channel is modelled as a Rayleigh fading channel where $h_1$ is a random variable (RV) with $\sigma_f^2 = 1$ as the scale parameter [8]. The PDF is expressed as

$$f_{h_1}(h_1) = \frac{h_1}{\sigma_f^2} \exp\left(\frac{-h_1^2}{2\sigma_f^2}\right), h_1 > 0 \tag{1}$$

The received signal is

$$r(t) = h_1 s(t) + n_f(t) \tag{2}$$

where s(t) is the signal being sent using the RF channel and $n_f(t)$ is additive white Gaussian noise (AWGN) having mean zero and $N_0$ as deviation. The received instantaneous SNR for RF channel is

$$\gamma_f = \frac{P_{tf} G_f}{N_0} h_1^2 = \bar{\gamma}_f h_1^2 \tag{3}$$

where $P_{tf}$ is transmitted signal's power, $G_f$ is the average power gain and $\gamma_f > 0$. By using random variable transformation, PDF and CDF in terms of SNR can be obtained respectively as

$$f_{\gamma_f}(\gamma_f) = \frac{1}{2\bar{\gamma}_f} \exp\left(\frac{-\gamma_f}{2\bar{\gamma}_f}\right) \tag{4}$$

$$F_{\gamma_f}(\gamma_f) = 1 - \exp\left(\frac{-\gamma_f}{2\bar{\gamma}_f}\right) \tag{5}$$

*2) Fiberless optical Channel:* The FSO link is modeled with the assumption that it operates under conditions of strong turbulence. Similar to the RF channel, $I_1$ is a RV which denotes the instantaneous irradiance of the fiberless optical path which is modeled using exponential distribution with a scale parameter $\alpha$ and represented as $I_1 \sim \exp(\alpha)$ [2].

$$f_{I_1}(I_1) = \frac{1}{\alpha} \exp\left(\frac{-I_1}{\alpha}\right), I_1 > 0 \tag{6}$$

The fiberless optical channel photo-current at the photo-detector (PD) is

$$i(t) = P_{oi} R_{od}[1 + \delta_1 m(t)] I_1 + n_{fso1}(t) \tag{7}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 06:33:43 UTC from IEEE Xplore. Restrictions apply.

where $P_{oi}$ is the optical incident power on PD, $R_{od}$ is responsivity, $\delta_1$ is modulation depth, $m(t)$ is electrical subcarrier which is pre-modulated, $n_{fso1}(t)$ is AWGN having mean zero and $\sigma_{no}^2$ as variance. The instantaneous received SNR for the FSO channel is

$$\gamma_{fso1} = \frac{(P_{oi}R_{od}\delta_1)^2}{\sigma_{no}^2} I_1^2 = \bar{\gamma}_{fso1} I_1^2 \tag{8}$$

where $\bar{\gamma}_{fso1}$ is the mean electrical SNR and $\gamma_{fso1} > 0$. By using RV transformation, the PDF and CDF in terms of SNR of the fiberless optical channel can be derived respectively as

$$f_{\gamma_{fso1}}(\gamma_{fso1}) = \frac{1}{2\alpha\sqrt{\bar{\gamma}_{fso1}\gamma_{fso1}}} \exp\left(\frac{-1}{\alpha}\sqrt{\frac{\gamma_{fso1}}{\bar{\gamma}_{fso1}}}\right) \tag{9}$$

$$F_{\gamma_{fso1}}(\gamma_{fso1}) = 1 - \exp\left(\frac{-1}{\alpha}\sqrt{\frac{\gamma_{fso1}}{\bar{\gamma}_{fso1}}}\right) \tag{10}$$

*3) Hybrid Channel:* Maximal ratio combining (MRC) requires the familiarity of fading coefficients of each branch, which makes it a complex combining technique [13]. Selection combining is less computationally demanding and resource-intensive, making it particularly well-suited for implementation in proposed system with limited power and budget constraints [13]. Thus, in this work, the selection combining technique is applied at the receiver. Selection combining is the least complicated combining technique using which the CDF of the hybrid link is mathematically obtained

$$F_{\gamma_1}(\gamma) = \left(1 - \exp\left(-\frac{\sqrt{\gamma}}{\alpha\sqrt{\bar{\gamma}_{fso1}}}\right)\right)\left(1 - \exp\left(-\frac{\gamma}{2\bar{\gamma}_f}\right)\right) \tag{11}$$

### B. Hybrid Link 2

*1) RF Channel:* Considering more severe fading in hybrid link 2, the RF channel is modelled using Weibull distribution with parameters $\zeta$ and $\lambda$ and denoted as $h_2 \sim wbl(\zeta, \lambda)$. Its PDF [14] is

$$f_{h_2}(h_2) = \frac{\lambda}{\zeta} h_2^{\lambda-1} \exp\left(\frac{-h_2^{\lambda}}{\zeta}\right), h_2 \geq 0 \tag{12}$$

where $\lambda > 0$ is the shape metric that measures the severity of fading. Similar to hybrid link 1 analysis, the CDF of the RF channel in terms of SNR with $\gamma_{2f} > 0$ is derived

$$F_{\gamma_{2f}}(\gamma_{2f}) = 1 - \exp\left(\frac{-1}{\zeta}\left(\sqrt{\frac{\gamma_{2f}}{\bar{\gamma}_{2f}}}\right)^{\lambda}\right) \tag{13}$$

*2) Fiberless optical Channel:* The FSO channel for hybrid link 2 is modelled using an exponential distribution with a scale parameter $\beta$ which depends on atmospheric turbulence and is denoted as $I_{fso2} \sim \exp(\beta)$ and its PDF is

$$f_{I_2}(I_2) = \frac{1}{\beta} \exp\left(\frac{-I_2}{\beta}\right), I_2 > 0 \tag{14}$$

Similar to hybird link 1, the CDF in terms of SNR with $\gamma_{fso2} > 0$ is obtained as

$$F_{\gamma_{fso2}}(\gamma_{fso2}) = 1 - \exp\left(\frac{-1}{\beta}\sqrt{\frac{\gamma_{fso2}}{\bar{\gamma}_{fso2}}}\right) \tag{15}$$

*3) Hybrid Channel:* By using (13) and (15), the expression for hybrid link 2 is obtained as

$$F_{\gamma_2}(\gamma) = \left(1 - \exp\left(-\frac{\sqrt{\gamma}}{\zeta}\left(\sqrt{\frac{1}{\bar{\gamma}_{2f}}}\right)^{\lambda}\right)\right) \left(1 - \exp\left(-\frac{\sqrt{\gamma}}{\beta}\sqrt{\frac{1}{\bar{\gamma}_{fso2}}}\right)\right) \tag{16}$$

### C. Eavesdropper Channel Model

We model the RF-Eve using a Rayleigh fading channel. In order to deceit the hybrid link the intruder characteristics need to be similar to RF link as the RF link is more susceptible to wiretapping. Similar to hybrid link 1, the RF channel PDF of the eavesdropper is expressed as

$$f_{\gamma_e}(\gamma_e) = d_e \exp\left(-d_e\gamma_e\right) \tag{17}$$

where $\gamma_e > 0$ and $d_e$ is a parameter depending on the received electrical average SNR of the Eve.

## III. PERFORMANCE ANALYSIS

In this section, secrecy capacity (SC), probability of SPSC and an algorithm for the best link selection are discussed. Further, closed-form expressions for probability of SPSC are derived for both the links.

### A. Secrecy Capacity

Secrecy capacity is the maximal rate of transmission at which a malicious intruder is not able to fetch any data. Capacity is defined with a metric that is positive valued. Therefore, secrecy capacity is treated to be zero in case Eve channel gain is greater than that of the channel gain of a genuine receiver. The secrecy capacity can be expressed as [9]

$$SC = \begin{cases} C_{Link} - C_{Eve}, \text{if } C_{Link} > C_{Eve} \\ 0, \text{otherwise} \end{cases} \tag{18}$$

where $C_{Link}$ and $C_{Eve}$ are the instantaneous capacities of hybrid links and Eves respectively. The channel capacity of link is expressed as [9]

$$C_k = R_b \log_2(1 + \gamma_k) \tag{19}$$

where $k \in \{Link1, Link2, Eve\}$. and $R_b$ denotes the bandwidth. By using (18) and (19), the secrecy capacity of hybrid links 1 and 2 are expressed as (20) and (21) respectively

$$SC_1 = \begin{cases} \log_2(1+\gamma_1) - \log_2(1+\gamma_e), \gamma_1 > \gamma_e \\ 0, \text{otherwise} \end{cases} \tag{20}$$

$$SC_2 = \begin{cases} \log_2(1+\gamma_2) - \log_2(1+\gamma_e), \gamma_2 > \gamma_e \\ 0, \text{otherwise} \end{cases} \tag{21}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 06:33:43 UTC from IEEE Xplore. Restrictions apply.

### B. Probability of SPSC

The probability of SPSC is a significant probabilistic metric to characterize the security of communication between a pair of legitimate transmitter and receiver. The probability of presence of secure transmission is defined using the probability of SPSC. The probability of SPSC is defined as

$$P^{+}_{SPSC} = Prob(SC > 0) \tag{22}$$

Using (20), (21) and (22), the expression for the probability of SPSC of $n$-th hybrid link is written as

$$P^{+}_{SPSC_n} = Prob\left(R_b \log_2\left(\frac{1+\gamma_n}{1+\gamma_e}\right) > 0\right) \tag{23}$$

where $n \in \{1, 2\}$. In terms of secrecy outage probability (SOP), the probability of SPSC is defined [15]

$$P^{+}_{SPSC_n} = 1 - SOP_n, \text{ for } \varphi = 1 \tag{24}$$

where $\varphi = \exp(R_s)$, $R_s$ is the predetermined information transmission rate and SOP of $n$-th hybrid link is expressed as

$$SOP_n = \int_0^{\infty} f_{\gamma_e}(\gamma) F_{\gamma_n}(\gamma) d\gamma \tag{25}$$

**Lemma 1.** *The probability of SPSC for hybrid link 1 is obtained as*

$$\begin{aligned} &P^{+}_{SPSC1} \\ &= 1 - \frac{\sqrt{\pi}}{2\alpha\sqrt{d_e\bar{\gamma}_{fso1}}} \exp\left(\frac{1}{4\alpha^2 d_e\bar{\gamma}_{fso1}}\right) \\ &\times \operatorname{erfc}\left(\frac{1}{2\alpha\sqrt{d_e\bar{\gamma}_{fso1}}}\right) + \frac{d_e\bar{\gamma}_f\sqrt{2\pi\bar{\gamma}_f}}{(1+2d_e\bar{\gamma}_f)^{3/2}\sqrt{\bar{\gamma}_{fso1}}} \\ &\times \exp\left(\frac{-0.5\alpha^{-2}\bar{\gamma}_f}{\bar{\gamma}_{fso1}(1+2d_e\bar{\gamma}_f)}\right) \operatorname{erfc}\left(\sqrt{\frac{2\bar{\gamma}_f}{\bar{\gamma}_{fso1}(1+2d_e\bar{\gamma}_f)}}\right) \end{aligned} \tag{26}$$

*where* erfc *is the complementary error function defined as* $\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}}\int_z^{\infty} e^{-t^2} dt$.

*Proof.* Substituting (11) and (17) in (25) the SOP for hybrid link 1 is written as

$$\begin{aligned} SOP_1 = \int_0^{\infty} d_e e^{-d_e\gamma} \left(1 - \exp\left(-\frac{\sqrt{\gamma}}{\alpha\sqrt{\bar{\gamma}_{fso1}}}\right)\right) \\ \times \left(1 - \exp\left(-\frac{\gamma}{2\bar{\gamma}_f}\right)\right) d\gamma \end{aligned} \tag{27}$$

Further, by substituting (27) in (24) and by solving, the final expression is obtained as (26) and hence proof of Lemma 1 is completed. ■

**Lemma 2.** *The probability of SPSC for hybrid link 2 is obtained as*

$$\begin{aligned} &P^{+}_{SPSC2} \\ &= 1 - \frac{\sqrt{\pi}}{2\beta\sqrt{d_e}} \exp\left(\frac{1}{4\beta^2 d_e\bar{\gamma}_{fso2}}\right) \operatorname{erfc}\left(\frac{1}{2\beta\sqrt{d_e\bar{\gamma}_{fso2}}}\right) \\ &- \frac{\sqrt{\pi}}{2\zeta\sqrt{d_e}} \exp\left(\frac{1}{4\zeta^2(\bar{\gamma}_{2f})^{\lambda-1} d_e\bar{\gamma}_{2f}}\right) \operatorname{erfc}\left(\frac{(\sqrt{\bar{\gamma}_{2f}})^{1-\lambda}}{2\zeta\sqrt{d_e\bar{\gamma}_{2f}}}\right) \\ &+ \frac{\sqrt{\pi}\sqrt{\bar{\gamma}_{fso2}\zeta(\bar{\gamma}_{2f})^{\lambda}}}{2\zeta\sqrt{\bar{\gamma}_{2f}d_e\bar{\gamma}_{fso2}}} \exp\left(\frac{\bar{\gamma}_{fso2}\zeta\bar{\gamma}_{2f}^{\lambda}}{4\zeta^2 d_e\bar{\gamma}_{fso2}\bar{\gamma}_{2f}}\right) \\ &\times \operatorname{erfc}\left(\frac{\sqrt{\bar{\gamma}_{fso2}\zeta(\bar{\gamma}_{2f})^{\lambda}}}{2\zeta\sqrt{d_e\bar{\gamma}_{fso2}\bar{\gamma}_{2f}}}\right) \end{aligned} \tag{28}$$

*Proof.* Substituting (16) and (17) in (25) the SOP for hybrid link 2 is written as

$$\begin{aligned} SOP_2 = \int_0^{\infty} d_e e^{-d_e\gamma} \left(1 - \exp\left(-\frac{\sqrt{\gamma}}{\zeta}\left(\sqrt{\frac{1}{\bar{\gamma}_{2f}}}\right)^{\lambda}\right)\right) \\ \times \left(1 - \exp\left(-\frac{\sqrt{\gamma}}{\beta}\sqrt{\frac{1}{\bar{\gamma}_{fso2}}}\right)\right) d\gamma \end{aligned} \tag{29}$$

Moreover, by substituting (29) in (24) and by solving, the final expression is obtained as (28) and hence concluding the proof of Lemma 2. ■

### C. Algorithm for Selection of Best Secured Hybrid Link

The most secured link for communication between legitimate transmitter and receiver pairs is selected by our proposed algorithm, which compares the probability of SPSC for both hybrid links. Here $\ell_i$ signifies a selection metric where $i$ denotes the corresponding link. If $\ell_1 = 1$ then it indicates that link 1 is selected for data transmission, else if $\ell_2 = 1$ it indicates that link 2 is selected.

The parameters such as $P^{+}_{SPSC1}(j)$, $P^{+}_{SPSC2}(j)$, $\ell_1$ and $\ell_2$ are initialized and set to zero. Thereafter, $P^{+}_{SPSC1}$ and $P^{+}_{SPSC2}$ are obtained based on randomly generated channel conditions and characteristics of eavesdropper. Further, a comparison is carried out between the probability of SPSC for both hybrid links. If $P^{+}_{SPSC1} > P^{+}_{SPSC2}$, then it shows that link 1 is more secure and the probability to be eavesdropped is less. Hence, $\ell_1$ becomes 1 indicating that link 1 is selected for communication. Therefore, the proposed algorithm helps in enhancing the security of our system.

## IV. RESULTS AND DISCUSSIONS

In this section, to validate the performance of the proposed hybrid system, simulation results are presented and discussed. Since the noise and behavior of an eavesdropper are random in nature, the results are obtained for two different conditions

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 06:33:43 UTC from IEEE Xplore. Restrictions apply.

**Algorithm 1** Best secure link selection

**procedure** $\chi_\mu(P^+_{SPSC1}, P^+_{SPSC2})$
  Initialization: j=0
  $P^+_{SPSC1}(j) = 0$, $P^+_{SPSC2}(j) = 0$
  $\ell_1 = 0$, $\ell_2 = 0$
  Iteration: $j + 1 \leftarrow j$
  $\chi_\mu = \arg\max[P^+_{SPSC1}(j-1), P^+_{SPSC2}(j-1)]$
  **if** $P^+_{SPSC1} > P^+_{SPSC2}$ **then**
    $\ell_1 \rightarrow 1$
  **else**
    $\ell_2 \rightarrow 1$
  **end if**
**end procedure**

of atmospheric channel. Thus, the performance of the hybrid links at different instants of time is distinct. In order to show the effectiveness of our proposed algorithm, we consider two different random channel conditions referred to as condition 1 and condition 2. Further Theo. and Si. in the subsequent figures refer to theoretical and simulated results respectively.

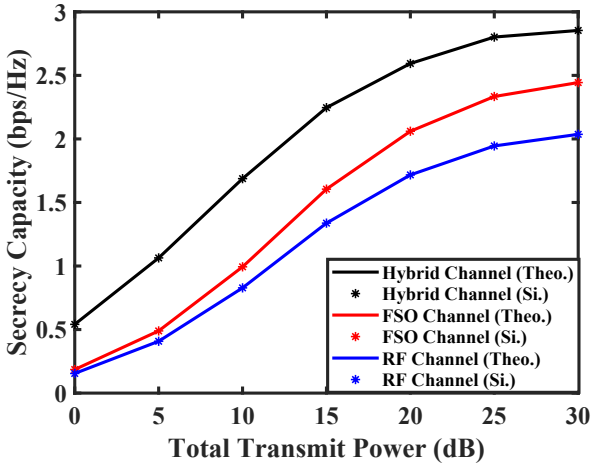



Fig. 2: Comparison of secrecy capacity for FSO, RF and hybrid links for Link 1

The path loss exponent value is 3.5 as its typical value ranges between 1.5 and 5. The FSO links typically range between 300 m to 5 km depending on speed and required availability. In this work, the separation between the source and destination node is 2 km. For the FSO links, the wavelength is 1550 nm, photosensitivity is 0.8 A/W, modulation index is 0.8 and optical power is 45 mW. For the RF links, the transmitted power is 5 mW and the wavelength is 10 GHz.

In Fig. 2 and Fig. 3, the secrecy capacity of FSO, RF and hybrid channels for links 1 and 2 are presented. From Fig. 2 and Fig. 3, it is evident that the secrecy capacity of the hybrid link is higher as compared to independent optical and RF links. This shows the advantage of hybrid link over independent FSO

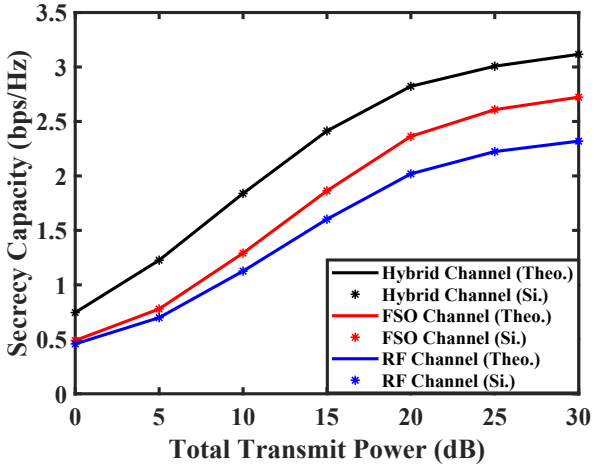



Fig. 3: Secrecy capacity versus total transmit power for hybrid, FSO and RF links for Link 2

and RF links due to the complementary nature of FSO and RF subsystems.

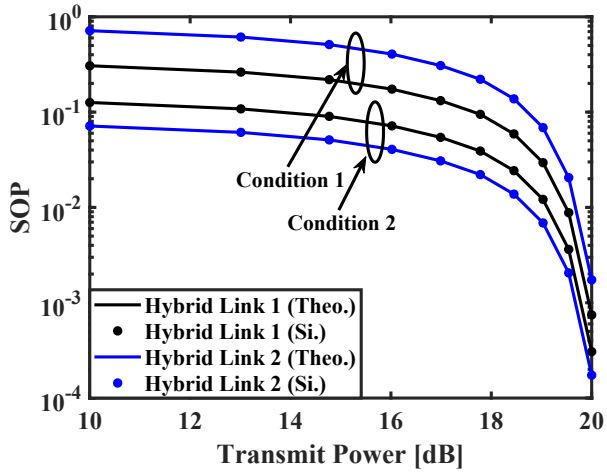



Fig. 4: Comparison of SOP for link 1 and link 2 under different channel conditions

The SOP characteristics against the transmit power are shown in Fig. 4. Fig. 4 depicts that an increase in transmit power improves the legitimate channel's capacity, which can reduce the SOP. Clearly in condition 1, hybrid link 1 performs better when compared with hybrid link 2. The SOP for hybrid links decreases more steeply with increasing transmit power since it provides a secure path. In condition 2, the SOP for hybrid link 2 decreases faster with increasing transmit power due to a better balance between the legitimate and eavesdropping channel capacities.

Fig. 5 shows the impact of SNR on the probability of SPSC. In Fig. 5, with the increase in SNR, the probability of communi-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 06:33:43 UTC from IEEE Xplore. Restrictions apply.

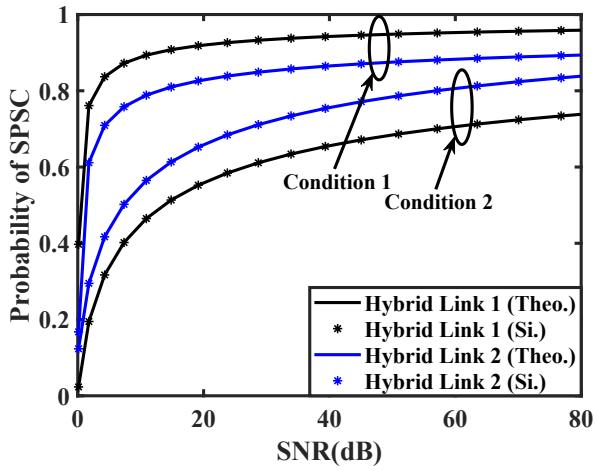



Fig. 5: Comparison of probability of SPSC for link 1 and link 2 under different channel conditions

cation links being more secure increases. Higher the probability of SPSC better is the link performance under a particular channel condition. For instance, considering condition 1 shown in Fig. 5, at an SNR of 10 dB, the probability of SPSC for link 1 and link 2 is found to be 0.89 and 0.7656 respectively which is obtained using (26) and (28). These values clearly matches with our simulation results as well. The probability of SPSC of hybrid link 1 in condition 1 is 1.67 times higher than that of condition 2 at SNR of 20 dB. On the other hand, the probability of SPSC of hybrid link 2 in condition 1 is 1.27 times higher than that of condition 2 for the same SNR. This variation of SPSC is due to randomness in the behaviour of Eves and channel conditions. Thus, in a certain condition it will be important to use that link for communication which is more secure.

Based on the proposed Algorithm 1, a secure hybrid link is selected on the basis of the security metric for the transmission of confidential data. It can be observed that in Fig. 5, since the probability of SPSC of link 1 is better for condition 1, therefore link 1 is needed to be selected for communication. Additionally, Fig. 5 displays the likelihood of SPSC for condition 2, which is a different set of variables from condition 1. For condition 2, link 2 has more probability of being secure than link 1, and hence, link 2 needs to be selected for transmission. This selection can be effectively done using Algorithm 1. Thus, this makes the overall communication system highly secure for the transmission of information.

## V. CONCLUSION

In this paper, an algorithm is proposed to select the most secured communication channel between two hybrid FSO/RF links. Thus improving the overall security of the whole communication system. Moreover, the mathematical closed-form expressions for the probability of SPSC for both the hybrid links are derived in the presence of a static RF-Eve. Further, to validate the analytical results, Monte-Carlo simulations are performed. This work can be extended further by estimating channel state information (CSI) using machine learning techniques to enhance the reliability of a hybrid communication system.

## VI. ACKNOWLEDGEMENT

Himani Mehta is supported by DST, Government of India vide reference no. DST/WISE-PhD/ET/2023/6 under 'WISE Fellowship for Ph.D.' programme to carry out this work

## REFERENCES

[1] S. C. Tokgoz, S. Althunibat, S. L. Miller, and K. A. Qaraqe, "On the secrecy capacity of hybrid fso-mmwave wiretap channels," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 4, pp. 4073–4086, 2022.

[2] G. Narang, M. Aggarwal, H. Kaushal, A. Kumar, S. Ahuja, and N. K. Shukla, "Performance evaluation of dual hop mixed fso rf system using differential chaos shift keying with secrecy analysis," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 11, pp. 17347–17358, 2024.

[3] A. D. Wyner, "The wire-tap channel," *Bell system technical journal*, vol. 54, no. 8, pp. 1355–1387, 1975.

[4] H. Mehta and S. Sengar, "Computation of secrecy and outages for hybrid fso/rf system with modified switching," in *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pp. 1–6, IEEE, 2021.

[5] T. M. Hoang, T. Q. Duong, N.-S. Vo, and C. Kundu, "Physical layer security in cooperative energy harvesting networks with a friendly jammer," *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 174–177, 2017.

[6] W. M. R. Shakir, "Physical layer security performance analysis of hybrid fso/rf communication system," *IEEE access*, vol. 9, pp. 18948–18961, 2020.

[7] Y. Ai, A. Mathur, H. Lei, M. Cheffena, and I. S. Ansari, "Secrecy enhancement of rf backhaul system with parallel fso communication link," *Optics communications*, vol. 475, p. 126193, 2020.

[8] Q. Sun, D. Yan, and Y. Wang, "Research on physical layer security performance of hybrid rf/fso system based on ris assistance under fso eavesdropping environment," *IEEE Photonics Journal*, vol. 16, no. 5, pp. 1–9, 2024.

[9] Y. Ai, A. Mathur, L. Kong, and M. Cheffena, "Secure outage analysis of fso communications over arbitrarily correlated málaga turbulence channels," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3961–3965, 2021.

[10] M. J. Saber and S. M. S. Sadough, "On secure free-space optical communications over málaga turbulence channels," *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 274–277, 2017.

[11] S. Sharma, A. Madhukumar, and R. Swaminathan, "Performance analysis of hybrid fso/rf system with transmit aperture selection," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, IEEE, 2020.

[12] S. Sharma, J. Tan, A. Madhukumar, and R. Swaminathan, "Switching-based transmit antenna/aperture selection in a miso hybrid fso/rf system," in *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2018.

[13] M. D. Selvaraj and S. R, "Performance of hybrid selection and switch-and-stay combining with decode-and-forward relaying," *IEEE Communications Letters*, vol. 18, no. 12, pp. 2233–2236, 2014.

[14] E. Balti, M. Guizani, and B. Hamdaoui, "Hybrid rayleigh and double-weibull over impaired rf/fso system with outdated csi," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2017.

[15] P. V. Trinh, A. Carrasco-Casado, A. T. Pham, and M. Toyoshima, "Secrecy analysis of fso systems considering misalignments and eavesdropper's location," *IEEE Transactions on Communications*, vol. 68, no. 12, pp. 7810–7823, 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 06:33:43 UTC from IEEE Xplore. Restrictions apply.

Check for updates

*Luminescence*

WILEY

LUMINESCENCE
The Journal of Biological and Chemical Luminescence

RESEARCH ARTICLE

# Photoluminescence Behavior of $Sm^{3+}$-Doped $YPO_4$ Phosphors for Solid-State Lighting

Vijay Singh[1] | Sheetal Kumari[2] | M. M. Haidari[1] | Aadil Ahmad Bhat[1] | A. S. Rao[2] | Ji Bong Joo[1]

[1]Department of Chemical Engineering, Konkuk University, Seoul, Republic of Korea | [2]Department of Applied Physics, Delhi Technological University, New Delhi, India

**Correspondence:** Vijay Singh (vijayjiin2006@yahoo.com)

**Received:** 16 March 2025 | **Revised:** 21 April 2025 | **Accepted:** 26 April 2025

**Funding:** This work was supported by the National Research Foundation of Korea (2021R1A2C1092509).

**Keywords:** coprecipitation | photoluminescence | $Sm^{3+}$ ions | XRD | $YPO_4$

## ABSTRACT

Phosphors emitting orange light under ultraviolet (UV) excitation are crucial for enhancing the performance of white light–emitting diodes. In this study, we investigate the photoluminescence properties of $Sm^{3+}$-doped yttrium orthophosphate ($YPO_4$) phosphors synthesized using the coprecipitation method. X-ray diffraction and Fourier-transform infrared spectroscopy were used to examine the structure and vibrational properties of the synthesized phosphors. By analyzing diffuse reflectance spectra using the Kubelka–Munk theory, the optical bandgap was determined to be 4.03 eV, corresponding to the UV region. Photoluminescence spectra, obtained under 406-nm excitation, showed orange emission at 602 nm, attributable to the $^4G_{5/2} \rightarrow {}^6H_{7/2}$ transition of $Sm^{3+}$ ions. Dipole–dipole interactions were responsible for the concentration-quenching effect, investigated using Dexter's hypothesis. Photometric and colorimetric investigations were performed to explore the emission behavior of the phosphors. The calculated CIE coordinates confirmed emission in the orange region, and photometric investigations revealed that the phosphor exhibited orange emissions under UV illumination. Overall, this study provides valuable insights into the emission behavior of $YPO_4$ doped with $Sm^{3+}$ ions, which are essential for assessing their suitability and potential applications for advanced photonic devices.

## 1 | Introduction

Developing efficient and eco-friendly phosphors for solid-state lighting (SSL) and white light–emitting diodes (w-LEDs) is a key challenge in optoelectronics. Traditional rare-earth–doped materials are expensive, have limited emission tuning, and pose environmental concerns. SSL has drawn significant interest because of its excellent luminescence, low operating voltage, and chemical and thermal durability [1–3]. w-LEDs, regarded as the fourth generation of SSL, have replaced conventional incandescent lamps. Fluorescent lamps have been extensively studied because of their unique benefits, such as environmental friendliness, short response times, high luminous efficiency, long lifespan, reliability, and lack of mercury [4–6]. Recent research has focused on developing phosphors with strong luminescence, which can be activated in the UV region to enhance the performance of w-LEDs. At present, phosphors are widely used in scintillators, w-LEDs, display devices, phototherapy lamps, upconversion lasers, and fluorescent lamps [5, 7].

White light can be generated using two approaches: one technique involves combining a blue LED with a yellow phosphor, such as $Y_3Al_5O_{12}:Ce^{3+}$ [6]. However, this technique results in a low color rendering index (CRI) and high color-correlated temperature. These drawbacks can be addressed using an alternative technique wherein RGB phosphors are excited using a near-ultraviolet (n-UV) semiconductor chip to produce white light. However, this method is not widely used because commercially available red phosphors, such as $Y_2O_2S:Eu^{3+}$, are ineffective at absorbing light from the n-UV semiconductor chip

© 2025 John Wiley & Sons Ltd.

[3]. Therefore, active research is being conducted to develop effective red phosphors with enhanced n-UV light absorption characteristics.

Luminescent materials are used in various applications, such as information displays, lighting, labeling, and sensor technologies [8, 9]. Despite the potential of lanthanide-doped inorganic nanophosphors in applications such as LEDs, lasers, optical amplifiers, biological fluorescence labels, and drug-release schemes [10, 11], their applicability is limited owing to their lower luminescent intensity compared with those of bulk materials. This limitation, attributable to quenching effects caused by surface defects and capping ligands, must be addressed to enhance the performance of lanthanide-doped nanostructured materials and promote their widespread use [12–15].

Recently, nanomaterials in various forms have been developed, such as nanofilms, nanoparticles, nanorods, and nanopores [16, 17]. Lanthanide phosphates demonstrate versatility in various applications such as lasers, ceramics, sensors, phosphors, heat-resistant materials, field-effect transistors, optoelectronics, medical and biological labeling, solar cells, and light sources [18]. Compositions based on lanthanide phosphates serve as host matrices in fluorescent lamps and color television monitors, benefiting from their strong vacuum UV absorption, attributable to the ($PO_4$) tetrahedral structure [19]. The flexible $PO_4$ tetrahedron can adopt various coordination geometries by varying the P–O bond lengths within the 1.45–1.65 Å range [17]. Furthermore, $MO_6$ (M = transition or rare-earth (RE) metal) groups can be combined with $PO_4$ groups to form diverse structures where the $MO_6$ and $PO_4$ groups are joined by a common O atom [17–19]. Among RE orthophosphates, yttrium orthophosphate ($YPO_4$) has garnered significant research interest as an ideal host structure for rare-earth activator ions due to its remarkable chemical and optical properties, reduced toxicity, and good resistance to photochemical degradation [20]. The luminescent properties of $YPO_4$ arise from intra 4f and 4f–5d transitions of RE ions, resulting in enhanced stability and reliability.

The trivalent samarium ion ($Sm^{3+}$), with its $4f^5$ configuration, is an RE ion that has emerged as an excellent activator for generating strong orange light through $^4G_{5/2} \rightarrow {}^6H_J$ (J = 5/2, 7/2, 9/2, 11/2) transitions [8, 9]. Luminous materials doped with $Sm^{3+}$ ions exhibit substantial absorption and are advantageous for w-LEDs that use n-UV LED chips. Additionally, $Sm^{3+}$ ions can be codoped with $Tb^{3+}$ ions to produce white light (orange + cyan = white), leading to a higher CRI and enhanced color tenability [7]. Consequently, numerous $Sm^{3+}$-doped luminous materials, such as $Ba_2La_2ZnW_2O_{12}$:$Sm^{3+}$ [21], $Ba_{3-2x}Sm_xNa_xBPO_7$ [22], $LiCa_3MgV_3O_{12}$:$Sm^{3+}$ [23], $BaLaAlO_4$:$Sm^{3+}$ [24], $Mg_3Y_2Ge_3O_{12}$:$Sm^{3+}$ [25], $La_6Ba_4Si_6O_{24}F_2$:$Sm^{3+}$ [26], and $Ba_2SiO_4$:$Sm^{3+}$ [27], have been explored. Given these advantages, further investigation of $Sm^{3+}$-doped orange-emitting phosphors is crucial for their potential use in SSL.

Phosphates are preferred inorganic hosts for RE ions owing to their low synthesis temperature, ease of preparation, effective incorporation of RE ions, high luminescence efficiency, excellent chemical and thermal stability, effective charge stabilization, and strong absorption in the n-UV region [28–30]. Owing to its high dielectric constant, large indirect bandgap, moderate phonon energy, and high refractive index, $YPO_4$ is a suitable host [31–33]. The crystal structure of $YPO_4$ is a zircon-type tetragonal system with space group I41/amd, characterized by two deformed tetrahedral units formed by eight oxygen atoms and four oxygen atoms surrounding each phosphorus atom. This arrangement provides a suitable environment for doped RE ions. Owing to its excellent luminous efficiency, chemical stability, and low biological toxicity, $YPO_4$ has been extensively used. Moreover, $YPO_4$:$Sm^{3+}$, an orange-emitting phosphor, has been successfully used to fabricate w-LEDs.

Despite extensive studies on the doping of $YPO_4$ with $Eu^{3+}$, $Nd^{3+}$, $Pr^{3+}$, $Er^{3+}$, $Ce^{3+}$, and $Tm^{3+}$, a comprehensive study focusing on the photoluminescence (PL) behavior of $Sm^{3+}$ ions within the $YPO_4$ matrix, particularly synthesized via a cost-effective coprecipitation route and rigorously evaluated for its potential in advanced photonic devices, remains comparatively underexplored [34, 35]. $Sm^{3+}$ in $YPO_4$ exhibits the lowest excited level of $^4G_{5/2}$, leading to three main emitting levels ($^6H_j$, where j = (5/2, 7/2, and 9/2)) at 564, 601, and 645 nm, resulting in orange emission [36]. However, when preparing inorganic luminescent materials using wet-chemistry synthesis techniques, surface-adsorbed quenchers such as OH may coordinate with the RE ions [37], thereby reducing their emission intensity. Additionally, controlling the presence of water molecules associated with different phase structures poses a challenge in wet-chemistry synthesis. These factors highlight the need for optimizing synthesis methods to achieve the desired luminescent properties in $Sm^{3+}$-doped $YPO_4$ materials.

Among the various methods available for nanomaterial fabrication, such as sol–gel, hydrothermal, microwave, solid-state, and chemical methods [38–42], the coprecipitation method is considered suitable for synthesizing $YPO_4$. First, this technique enables precise control of stoichiometry and composition, facilitating the formation of well-defined crystal structures. Second, coprecipitation typically results in fine and homogeneous particles. Additionally, this method is simple and cost-effective.

Considering these aspects, in this study, $YPO_4$:$Sm^{3+}$ phosphors were synthesized via systematic optimization of coprecipitation synthesis method for enhanced orange emission. A thorough analysis of the material's suitability for next-generation SSL applications is provided based on its unique spectral characteristics and color purity (CP). Various analytical techniques, including x-ray diffraction (XRD), Fourier-transform infrared (FT-IR) spectroscopy, diffuse reflectance spectroscopy (DRS), and PL spectroscopy, were used to thoroughly examine the structural, optical, and luminescent properties of the prepared phosphors. Emission data obtained from PL spectroscopy were analyzed to determine the Commission internationale de l'éclairage (CIE) coordinates of the $YPO_4$:$Sm^{3+}$ phosphors. The findings provide valuable insights for developing phosphors with enhanced luminosity for potential applications in photonic devices.

## 2 | Experimental

A series of $Y_{1-x}PO_4$:$xSm^{3+}$ ($0.01 \leq x \leq 0.11$) phosphors were synthesized using the coprecipitation method, as illustrated in Figure 1. The base materials included $Y(NO_3)_3 \cdot 6H_2O$,



15227243, 2025, 5, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/bio.70190 by Delhi Technological University, Wiley Online Library on [12/05/2025]. 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

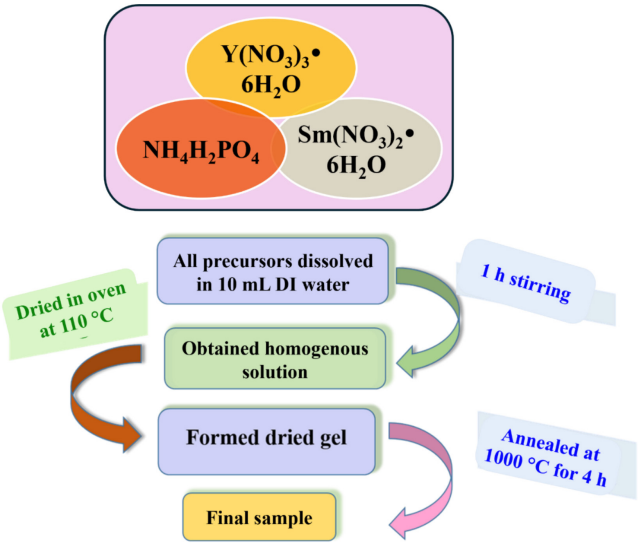



**FIGURE 1** | Synthesis of $Y_{1-x}PO_4$:$xSm^{3+}$ phosphors.

**TABLE 1** | Commission internationale de l'éclairage (CIE) color coordinates for $Y_{1-x}PO_4$:$xSm^{3+}$ phosphors using PL emission results (at $\lambda_{exc}=406$ nm).

| | $\lambda_{exc}=406$ nm | | |
|---|---|---|---|
| $Y_{1-x}PO_4$:$xSm^{3+}$ | $x$ | $y$ | Color purity (%) |
| $x=0.01$ | 0.579 | 0.419 | 86.95 |
| $x=0.03$ | 0.575 | 0.424 | 86.83 |
| $x=0.05$ | 0.573 | 0.425 | 85.71 |
| $x=0.07$ | 0.574 | 0.424 | 86.02 |
| $x=0.09$ | 0.579 | 0.421 | 86.50 |
| $x=0.11$ | 0.579 | 0.420 | 87.55 |

$NH_4H_2PO_4$, and Sm $(NO_3)_3{\cdot}6H_2O$. Table 1 summarizes the compositions of the synthesized phosphors and base materials. All base materials were weighed in stoichiometric ratios and dissolved in 10 mL of distilled water in a 100-mL beaker. The solution was stirred for 1 h using a magnetic stirrer to form a homogeneous precipitate. The resulting precipitate was dried in an oven at 110°C, followed by calcination in a muffle furnace at 1000°C for 4 h.

The phase identification of the synthesized powders was performed using a Rigaku x-ray diffractometer. Sample patterns were recorded over a 2θ range of 10°–80° using Cu–Kα radiation ($\lambda=1.5406$ Å). FT-IR spectra were derived using a Thermo Fisher Nicolet 6700 spectrometer. Thin films—prepared by mixing the samples with KBr—were scanned in the 4000–400 $cm^{-1}$ wavelength range. PL excitation (PLE) and emission spectra were collected at room temperature (RT) using a Shimadzu RF-5301PC spectrofluorophotometer equipped with a xenon flash lamp. UV–visible–near-infrared (UV–Vis–NIR) spectra were obtained at RT using a Cary 5000 UV–Vis–NIR spectrophotometer.

## 3 | Results and Discussion

### 3.1 | Crystal Structure and Particle Size Analysis

Figure 2a shows the crystal structure of the $YPO_4$ lattice in the standard orientation. $YPO_4$ has a tetragonal crystal structure with lattice parameters $a=b=6.882$ Å, $c=6.018$ Å, $Z=4$, unit cell volume $V=286.53$ $Å^3$, and space group I41/amd (No. 141) [43]. As shown in Figure 2a, eight oxygen atoms with two distinct bond lengths surround the trivalent crystallographic $Y^{3+}$ site, forming the $YPO_8$ dodecahedron. A $PO_4$ tetrahedron is formed by P atoms surrounded by four oxygen atoms with similar bond lengths (P–O = 1.5504 Å). Two oxygen atoms are shared by the $YO_8$ and $PO_4$ units in a corner configuration.

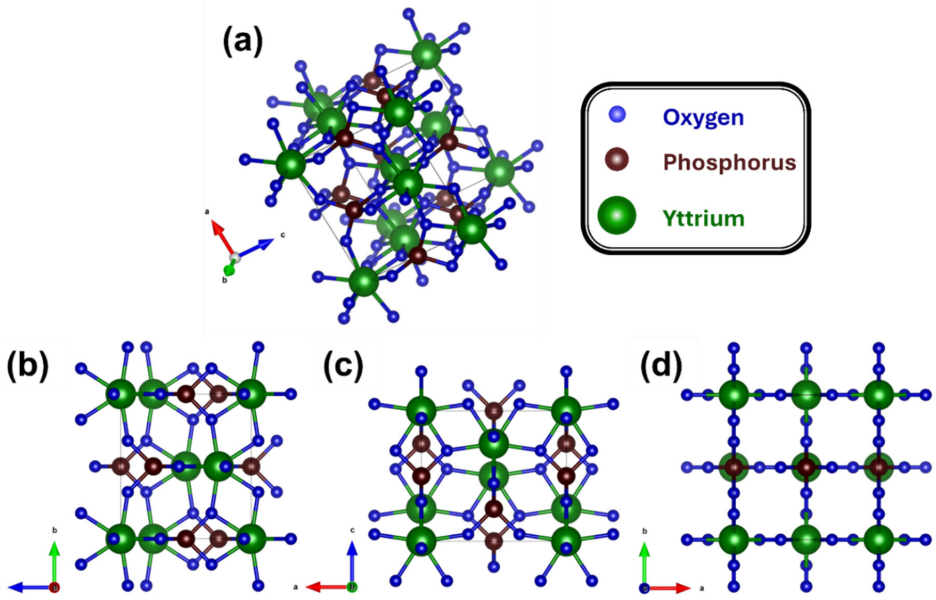



**FIGURE 2** | Crystal structure of $YPO_4$: (a) standard, (b) 100, (c) 010, and (d) 001 orientations. Blue, brown, and green spheres represent oxygen (O), phosphor (P), and yttrium (Y) atoms, respectively.



15227243, 2025, 5, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/bio.70190 by Delhi Technological University, Wiley Online Library on [12/05/2025]. 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

Different crystal orientations of the $YPO_4$ lattice are shown in Figure 2b–d.

Figure 3 shows the XRD patterns of $YPO_4$:x$Sm^{3+}$ phosphors, consistent with the standard JCPDS data of $YPO_4$ (JCPDS No. 084-0335) [44, 45]. The absence of impurity peaks indicates that the $Sm^{3+}$ ions are thoroughly integrated into the $YPO_4$ lattice. Owing to the similar ionic radii of $Sm^{3+}$ (1.09 nm) and $Y^{3+}$ (1.04 nm), $Sm^{3+}$ ions effectively substitute for $Y^{3+}$. The average crystallite size, $D$, is computed using the Debye–Scherrer relation [46]:

$$cos\theta = \frac{k\lambda}{\beta D} \quad (1)$$

where $\lambda$ is the x-ray wavelength in the diffractometer, $\theta$ is the Bragg diffraction angle, and $\beta$ is the full-width at half-maximum. The average crystallite size calculated for $YPO_4$ phosphors doped with $Sm^{3+}$ ions in concentrations of 0.01, 0.03, 0.05, 0.07, 0.09, and 0.11 mol is 26.3, 22.7, 19.4, 17.6, 16.3, and 15.2 nm, respectively. The decrease in average crystallite size with increasing $Sm^{3+}$ concentration is attributable to various factors associated with the doping process [47, 48]. For example, introducing $Sm^{3+}$ ions into the $YPO_4$ lattice may induce lattice strain and introduce defects into the crystal structure. As the concentration of $Sm^{3+}$ ions increases, the likelihood of lattice distortion and defects also increases. These defects can act as nucleation sites for crystal growth, reducing the crystallite size. Moreover, $Sm^{3+}$ ions may inhibit the growth of $YPO_4$ crystals during the synthesis process, resulting in the formation of smaller crystallites as the concentration of $Sm^{3+}$ ions increases. Substituting $Sm^{3+}$ ions for $Y^{3+}$ ions can result in a size mismatch, leading to lattice distortion and strain. The ionic radius of $Sm^{3+}$ may differ from that of the host lattice, causing local perturbations in the crystal structure, thereby increasing the lattice strain of $Sm^{3+}$-doped $YPO_4$. The lattice strain, determined using the Williamson–Hall equation and XRD data, is approximately 0.0003, indicating minimal distortion.

## 3.2 | FT-IR Analysis

In FT-IR spectroscopy, the locations of peaks in the spectrum correspond to specific vibrational modes of chemical bonds in the material. The FT-IR spectrum of $YPO_4$ shows characteristic peaks associated with phosphate groups, as well as the potential presence of hydroxyl groups or water. Figure 4 shows the FT-IR spectra recorded in the range of 4000–400 $cm^{-1}$ to identify the types of bonds present in $YPO_4$:0.01$Sm^{3+}$ and $YPO_4$:0.11$Sm^{3+}$ phosphors. The peak at 528 $cm^{-1}$ is attributable to the vibration of the Y–O bond [49]. The peaks at 638 $cm^{-1}$ and near 1025 $cm^{-1}$ are associated with antisymmetric bending and antisymmetric stretching vibrations of the $(PO_4)^{3-}$ groups, respectively. The peak at 2030 $cm^{-1}$ is associated with molecular vibrations or bond stretching in $Sm^{3+}$-doped $YPO_4$. Specifically, this peak may be related to vibrational modes influenced by the presence of $Sm^{3+}$ ions in the crystal lattice, potentially associated with their coordination environment in the $YPO_4$ lattice. Peaks at around 2345 $cm^{-1}$ are associated with the bending vibrations of the O–H group, arising from water molecules adsorbed on the material surface. The peak at approximately 2871 $cm^{-1}$ is ascribed to the asymmetric stretching of C–H group from organic residue [50]. Other peaks at 1510, 1710, and 3646 $cm^{-1}$ correspond to vibrational modes of specific molecular bonds or groups in the $YPO_4$ structure. The peak at 1510 $cm^{-1}$ pertains to the asymmetric stretching vibrations of phosphate groups ($\nu_3PO_4$), whose exact position may vary depending on the crystal structure and local environment. The peak at 1710 $cm^{-1}$ is typically associated with the stretching vibration of C=O bonds. However, in the case of $YPO_4$, a phosphate compound, the presence of C=O bonds is atypical, suggesting possible contamination or impurities. Ensuring sample purity and the absence of organic or carbonate contaminants is thus crucial. The peak at 3646 $cm^{-1}$ is indicative of O–H stretching vibrations, suggesting the presence of hydroxyl groups or water molecules in the $YPO_4$

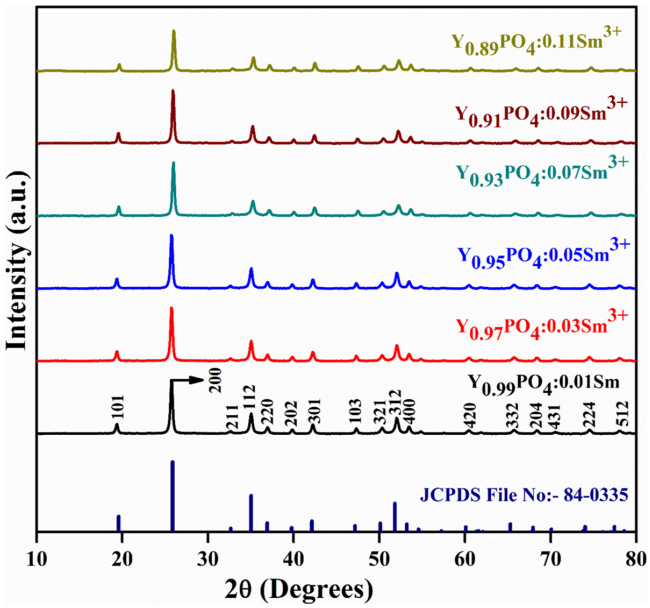



**FIGURE 3** | XRD patterns of $Y_{1-x}PO_4$:x$Sm^{3+}$ phosphors.

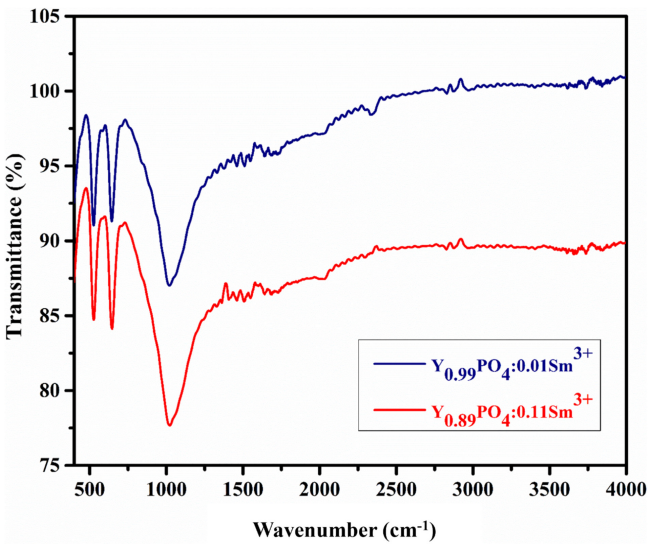



**FIGURE 4** | FT-IR spectra of $Y_{0.99}PO_4$:0.01$Sm^{3+}$ and $Y_{0.89}PO_4$:0.11$Sm^{3+}$ phosphors.



15227243, 2025, 5, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/bio.70190 by Delhi Technological University, Wiley Online Library on [12/05/2025]. 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

sample, specifically water molecules either adsorbed on the surface or incorporated within the crystal lattice.

### 3.3 | UV–Vis–NIR Spectroscopy

Figure 5 presents the DRS and estimated extinction absorption coefficient spectrum of $YPO_4$:0.03$Sm^{3+}$. Various absorption peaks appear in the UV–Vis–NIR spectral region, attributable to the complex energy level structure of the $Sm^{3+}$-doped $YPO_4$ crystal. The inhomogeneous crystal structure and undefined field lines contribute to line broadening, resulting in only a few well-defined absorption bands corresponding to $^{2S+1}L_J$ manifolds. These bands are centered at 1596, 1535, 1423, 1271, 1109, 965, 482, 411, 380, 366, and 349 nm and pertain to transitions from $^6H_{5/2}$ to $^6H_{15/2}$, $^6F_{3/2}$, $^6F_{5/2}$, $^6F_{7/2}$, $^6F_{9/2}$, $^6F_{11/2}$, $^4I_{11/2}$, $^4I_{11/2}+{}^4I_{13/2}+{}^4M_{15/2}$, $^6P_{5/2}$, $^6P_{7/2}$, $^4D_{1/2}$, and $^4D_{3/2}$, respectively, corresponding to the 4f–4f transitions for the $Sm^{3+}$ ion [49]. Notably, assigning absorption bands in the UV–Vis region is challenging owing to the overlap of different $^{2S+1}L_J$ levels [51]. Thus, the energy transitions of all samples are examined based on their induced electric dipole (ED) nature, adhering to the selection rule of $\Delta J \leq 6$, and magnetic dipole (MD) nature, following the selection rule of $\Delta J = (0, \pm 1)$. Visible absorption bands pertaining to spin-forbidden transitions are observed, indicating weak intensities. The NIR absorption band $^6H_{5/2} \rightarrow {}^6F_{5/2}$ is identified as the hypersensitive transition (HST) of $Sm^{3+}$, which is highly sensitive to the local environment. The selection rules governing the HST are $\Delta J \leq 2$, $\Delta L \leq 2$, and $\Delta S = 0$ [52]. Furthermore, the HST is associated with the nephelauxetic ratio, $\beta$, which provides insights into the degree of covalent character in Sm–O bonds, shedding light on the nature of the chemical interactions in the crystal structure involving $Sm^{3+}$ ions and oxygen atoms [53].

The results obtained through DRS are transformed into equivalent absorption coefficients using the Kubelka–Munk equation [54]. This transformation is crucial for characterizing absorption properties, as it establishes the relationship between the absorption and scattering characteristics of a material. Thus, the Kubelka–Munk theory enables the determination of the potential effects of $Sm^{3+}$ doping on the phosphor bandgap:

$$F(R) = \frac{(1-R^2)}{2R} = \frac{\alpha}{S} \tag{2}$$

where $\alpha$ is the absorption coefficient, $s$ is the scattering coefficient, and $R$ represents the reflectance.

The following expression is used to calculate the optical bandgap ($E_g$) [55]:

$$\alpha h\nu = C(h\nu - E_g)^n \tag{3}$$

where $n$ denotes the type of allowed transition, that is, 2 and 1/2 for indirect and direct transitions, respectively; $h\nu$ represents the photon energy; $C$ is an energy-independent constant. To determine the $E_g$ value of $Sm^{3+}$-doped $YPO_4$, the measured absorption coefficient is fitted to the Tauc equation, which models the absorption behavior of materials near their band edges as $(\alpha h\nu)^n = A(h\nu - E_g)$, with $A$ being a constant. In this study, an exponent value of $n = 1/2$ is used, considering the direct bandgap of the phosphor. The reverse tangent is drawn on the curve, as it intersects the $x$-axis, with the corresponding intercept representing the bandgap. The plot is drawn between $[F(R\infty)h\nu]^2$ and $h\nu$, as illustrated in Figure 5b, indicating the experimental bandgap of $Sm^{3+}$-doped $YPO_4$. Lower-energy transitions are observed in the optical absorption spectra. The determined bandgap of 4.03 eV indicates the energy difference between the valence and conduction bands in the material. This information is crucial for understanding the optical properties, light absorption behavior, and electronic transitions of $Sm^{3+}$-doped $YPO_4$. Wide-bandgap materials, such as $Sm^{3+}$-doped $YPO_4$, hold significant potential for optoelectronic applications and the development of energy-efficient devices, given their ability to manipulate and respond to light in distinct ways. The 4.03 eV bandgap in $Sm^{3+}$-doped $YPO_4$ is ideal for SSL applications, enabling efficient light absorption and emission in the UV to visible range. This property is crucial for creating phosphors in white LEDs. The wide bandgap minimizes overlap between absorption and emission spectra, enhancing luminescence efficiency. The direct bandgap nature promotes strong light–matter interactions, enabling precise control over emission properties

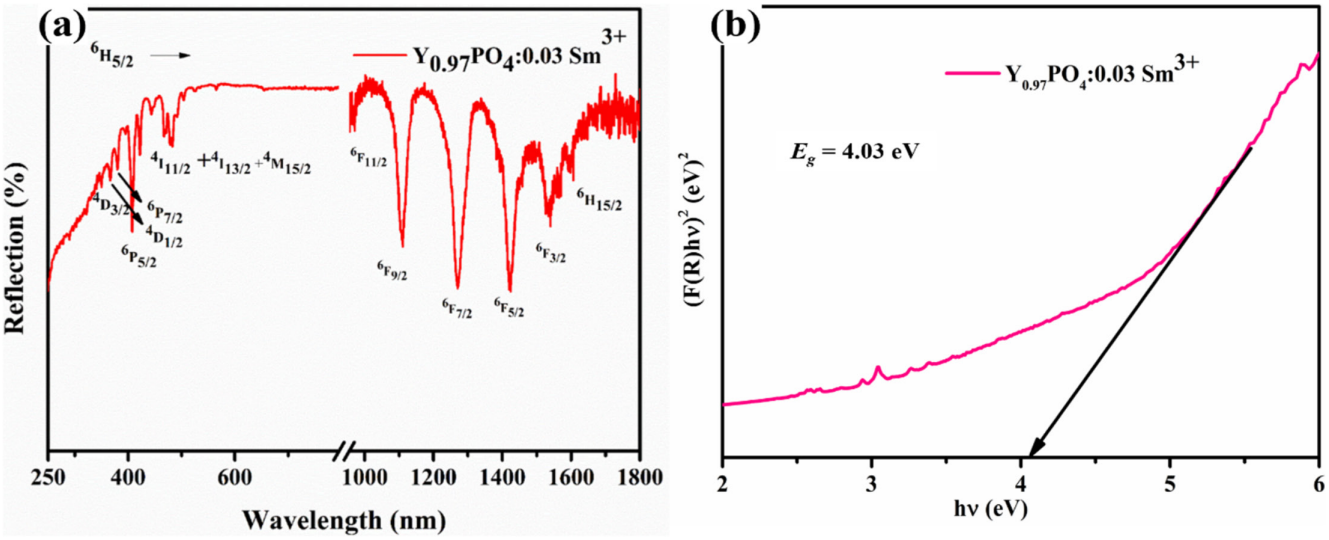



**FIGURE 5** | (a) UV–Vis–NIR spectrum and (b) Tauc plot of $Y_{0.97}PO_4$:0.03$Sm^{3+}$ to determine optical bandgap.



15227243, 2025, 5, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/bio.70190 by Delhi Technological University, Wiley Online Library on [12/05/2025]. 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

for customized color rendering and high-performance SSL devices. The material's ability to accommodate rare-earth ions like $Sm^{3+}$ allows for fine-tuning of emission wavelengths, making it suitable for energy-efficient lighting solutions with improved color quality and stability.

### 3.4 | PL Analysis

RE elements such as Sm are extensively studied owing to their potential applications in phosphors, lasers, and other optoelectronic devices. The excitation wavelength depends on the chemical environment and host material in which Sm is incorporated. PLE spectra of as-synthesized $YPO_4$:x$Sm^{3+}$ are obtained at RT. Figure 6a displays these PLE spectra, captured at an emission wavelength of 602 nm in the spectral range of 250–500 nm. The peaks at 306, 318, 333, 345, 363, 377, 391, 406, 419, 441, 463, 481, 501, and 532 nm are attributable to transitions from $^6H_{5/2}$ to excited states $^2F_{5/2}$, $^4L_{15/2}$, $^4G_{9//2,}$ $^4H_{9/2,}$ $^4D_{3/2}$, $^6P_{7/2,}$ $^4L_{15/2}$, $^4F_{7/2,}$ $^4L_{13/2,}$ $^4G_{9/2}$, $^4I_{13/2,}$ $^4M_{15/2}$, $^4F_{3/2}$, and $^4G_{7/2,}$ respectively [56]. Among these, $^6H_{5/2} \rightarrow {}^4F_{7/2}$ (406 nm) transitions exhibit the highest intensity; thus, this wavelength is considered when recording the emission spectra. The electronic transitions responsible for the absorption and emission of light are associated with f–orbitals. The absorption of light by electrons in these orbitals leads to excitation, and the emitted light corresponds to the return of electrons to lower energy states. The 406-nm excitation peak indicates that Sm absorbs light most efficiently at this wavelength. This peak is attributable to the arrangement of electrons within the atom and energy differences between different electronic states. The intensity of the PLE spectra depends on $Sm^{3+}$ concentration, with the maximum intensity observed at $x = 0.003$ mol. Notably, these predominant excitations are correlated with emissions suitable for n-UV or visible light–emitting device chips. The prepared phosphors demonstrate strong absorption in both the UV and visible light bands, and their emission spectra prominently lie within the orange spectral range. Consequently, the optimized phosphor shows promise for application in orange LEDs. Overall, the $Sm^{3+}$ concentration considerably affects the intensity of the PL spectra, and the absorption and emission characteristics of the selected material make it well suited for use in visible LEDs.

The emission spectra of $YPO_4$:x$Sm^{3+}$ phosphors are recorded under an excitation wavelength of $\lambda_{exc} = 406$ nm within the 500–750 nm range, as illustrated in Figure 6b. Four distinct emission peaks, corresponding to f–f transitions in the $Sm^{3+}$ $4f^5$ electron configuration, are observed. Specifically, the transitions of the $Sm^{3+}$ ion from $^4G_{5/2} \rightarrow (^6H_{5/2}, {}^6H_{7/2}, {}^6H_{9/2}$, and $^6H_{11/2})$ are attributable to peaks at 563, 602, 645, and 704 nm, respectively [57, 58]. The highest intensity is observed at 602 nm, corresponding to the $^4G_{5/2} \rightarrow {}^6H_{7/2}$ transition, suggesting that the $YPO_4$:$Sm^{3+}$ phosphor can be effectively used as an orange-emitting material.

Among the observed peaks, the peak at 563 nm is attributed to the MD transition $^4G_{5/2} \rightarrow {}^6H_{5/2}$, whereas the peak at 645 nm corresponds to the ED transition $^4G_{5/2} \rightarrow {}^6H_{9/2}$. The MD transition involves a change in the magnetic quantum number, typically resulting in stronger intensities compared with the ED transition, which involves a change in the ED moment. The most prominent peak at 602 nm results from a combination of both ED and MD transitions, influenced by the specific electronic configuration and local symmetry of $Sm^{3+}$ ions in the $YPO_4$ lattice. In 4f–4f transitions, ED transitions are typically forbidden. The analysis of symmetry in the local environment around $Sm^{3+}$ ions is reflected in the intensity ratio of the ED and MD transitions. In this study, the significantly higher intensity of the MD transition compared with that of the ED transition suggests a more symmetric environment for $Sm^{3+}$ ions in the lattice. This symmetry indicates the prevalence of the MD transition in the emission spectra.

Figure 7 shows the variation in emission intensity as a function of $Sm^{3+}$ concentration. The emission intensity increases as the concentration increases to $x = 0.03$ mol, beyond which it decreases owing to concentration quenching. Higher concentrations are associated with an increased probability of ion–ion interactions, including the energy transfer processes between adjacent $Sm^{3+}$ ions. These interactions can lead to nonradiative decay pathways, such as cross-relaxation, where energy is transferred from one ion to another without photon emission, reducing the overall luminescence efficiency. Moreover, as the dopant concentration increases, the available crystallographic sites for $Sm^{3+}$ ions become saturated. When these sites are fully occupied, additional dopant ions may not have adequate sites to

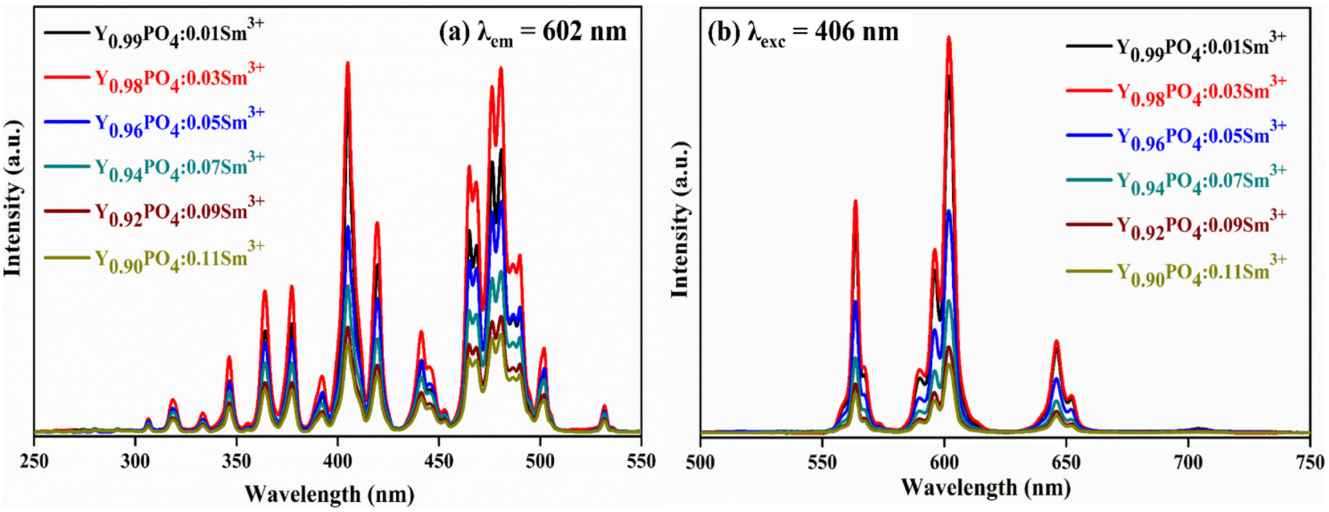



**FIGURE 6** | Photoluminescence spectra of $Y_{1-x}PO_4$:x$Sm^{3+}$ phosphors. (a) Emission spectra ($\lambda_{em} = 602$ nm). (b) Excitation spectra ($\lambda_{exc} = 406$ nm).



15227243, 2025, 5, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/bio.70190 by Delhi Technological University, Wiley Online Library on [12/05/2025]. 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

be incorporated, resulting in distortions in the crystal lattice and nonradiative processes. Furthermore, at higher concentrations, the excitation and emission bands of neighboring $Sm^{3+}$ ions may overlap, leading to energy transfer between ions, thereby quenching the luminescence [59, 60]. Figure 8 shows the energy level diagram, an intuitive representation of the excitation and emission mechanisms. Upon excitation from the ground state ($^6H_{5/2}$) to the excited state, the electrons in the excited states undergo nonradiative relaxation owing to the energy differences between distinct excited states, enabling electrons to reach the $^4G_{5/2}$ state. After radiative relaxation, the electrons in the $^4G_{5/2}$ state relax to lower levels, specifically $^6H_{5/2}$, $^6H_{7/2}$, $^6H_{9/2}$, and $^6H_{11/2}$, resulting in the distinctive emission transitions of the $Sm^{3+}$ ions. The cross-relaxation channels for the nonradiative energy transfer of $Sm^{3+}$ ions are

$$^4G_{5/2} + {}^6H_{5/2} \rightarrow {}^6F_{9/2} + {}^6F_{7/2}, {}^4G_{5/2} + {}^6H_{5/2} \rightarrow {}^6F_{7/2} + {}^6F_{9/2},$$

$$^4G_{5/2} + {}^6H_{5/2} \rightarrow {}^6F_{5/2} + {}^6F_{11/2}, \text{and} {}^4G_{5/2} + {}^6H_{5/2} \rightarrow {}^6F_{11/2} + {}^6F_{5/2}.$$

transfer distance ($R_c$) is computed using the Blasse formula to clarify the concentration quenching mechanism [61]:

$$R_c \approx 2\left[\frac{3V}{4\pi X_c N}\right]^{1/3} \quad (4)$$

where $X_c$, $V$, and $N$ denote the $Sm^{3+}$ concentration, unit cell volume, and number of accessible crystallographic sites occupied by activators, respectively. In this study, $V = 286.53 Å^3$, $X_c = 0.03$ mol, and $N = 4$, resulting in an $R_c$ value of 16.58 Å. The $R_c$ value for exchange interactions is typically approximately 5 Å. Consequently, as $R_c$ exceeds 5 Å for the studied system, multipolar–multipolar interactions between $Sm^{3+}$ ions are primarily responsible for the concentration quenching in $YPO_4$:$Sm^{3+}$. The relationship between the activator concentration and emission intensity clarifies the concentration quenching mechanism [62]:

$$\log\frac{I}{x} = A - \frac{Q\log x}{3} \quad (5)$$

where $I$ is the emission intensity and $A$ is a coefficient. The multipolar interaction constant $Q$ is equivalent to 6, 8, or 10 for dipole–dipole, dipole–quadrupole, and quadrupole–quadrupole interactions, respectively. Figure 9 displays the $\log(I/x)$ versus $\log(x)$ relationship, which fits linearly with a slope of 1.71. Thus, $Q$ is 5.13, suggesting that phosphor quenching is controlled by the dipole–dipole interaction.

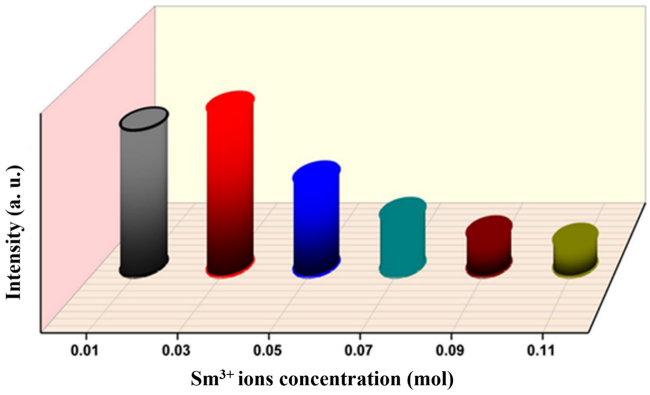



**FIGURE 7** | Variation in strong emission intensity ($I_{602\ nm}$) of $Y_{1-x}PO_4$:$xSm^{3+}$ phosphors as a function of $Sm^{3+}$ concentration.

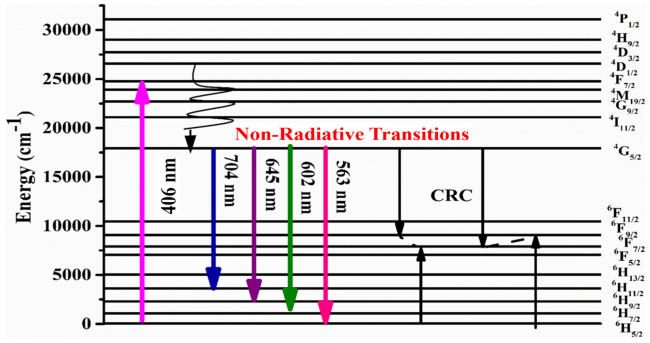



**FIGURE 8** | Schematic energy level diagram of $Sm^{3+}$ ions.

## 3.5 | Colorimetric and Photometric Analyses

Based on the emission spectra, the chromaticity coordinates of the $Sm^{3+}$-doped $YPO_4$ phosphors are determined. Figure 10a,b shows the CIE diagram for the $YPO_4$:$Sm^{3+}$ phosphors, and Table 1 summarizes the computed CIE color coordinates. The CIE coordinates of the $Sm^{3+}$-activated $YPO_4$ phosphors (Figure 10) are situated within the orange region. Under UV excitation at 365 nm, the phosphor emits orange light (Figure 11). The consistent placement of all samples within the orange region in the CIE diagram, as well as their compatibility with UV excitation, suggests that the prepared phosphors are suitable for various applications, particularly in the domain of LEDs requiring orange light emission. Thus, the findings provide valuable insights for optimizing the phosphor properties to meet the requirements of specific lighting or display applications.

As determined by a reported equation [63], the CP factor (Table 1) enables the assessment of emission color quality. For the optimized phosphor, the calculated CP ranges between 85% and 87%, indicating a high degree of CP in the orange emission range. This high purity suggests that the $YPO_4$:$Sm^{3+}$ phosphor can serve as an effective and high-quality orange-emitting component, especially in scenarios where accurate and vibrant

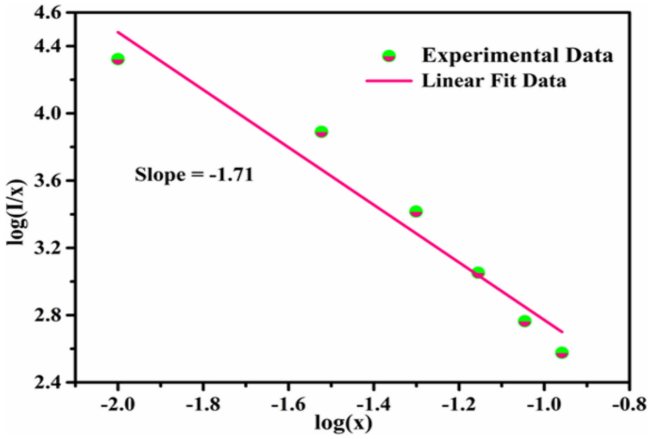



**FIGURE 9** | Dexter plot of $\log(I/x)$ versus $\log(x)$.



15227243, 2025, 5, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/bio.70190 by Delhi Technological University, Wiley Online Library on [12/05/2025]. 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

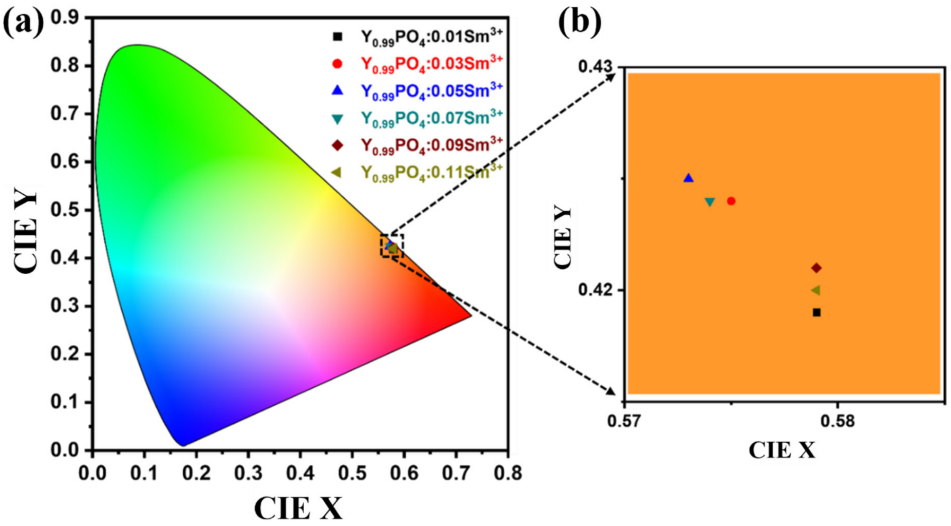



**FIGURE 10** | (a) CIE chromaticity diagram of $Y_{1-x}PO_4$:x$Sm^{3+}$ phosphors. (b) Enlarged view.

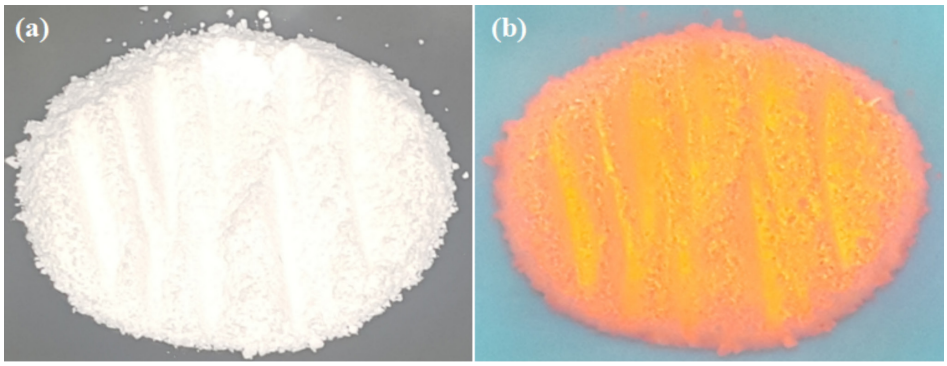



**FIGURE 11** | Typical photographs of $Y_{0.97}PO_4$:0.03$Sm^{3+}$ (a) under ambient light (the phosphor appears as a light brown powder) and (b) under UV-365 nm (the phosphor appears as an orange powder).

color representation is essential. These characteristics make it well suited for application in w-LEDs, where it can contribute to achieving the desired color temperature and quality of white light output. Overall, owing to the optimized composition, minimum energy transfer, and optical quality, the emission from $Sm^{3+}$-doped $YPO_4$ is concentrated in a specific wavelength range with high CP, rendering them valuable for LED applications requiring precise color representation.

## 4 | Conclusions

This study presents a cost-effective method for producing $YPO_4$ phosphors doped with $Sm^{3+}$ for improved orange emission. The synthesis involved precise control of the dopant concentration, resulting in $Y_{1-x}PO_4$:x$Sm^{3+}$ phosphors with a tetragonal crystal structure. XRD analysis confirmed the space group as $I4_1/amd$. Average crystallite size was determined as 19.47 nm using the Debye–Scherrer technique. FT-IR spectra revealed vibrational bonds in the phosphors, clarifying their composition and bonding. The bandgap of $YPO_4$:0.03$Sm^{3+}$ was determined to be 4.03 eV, providing insights into its electronic properties for optoelectronic applications. PL analysis showed that the synthesized phosphors emitted strong orange light at 602 nm ($^4G_{5/2} \rightarrow {}^6H_{7/2}$) when excited at 406 nm. The optimal $Sm^{3+}$ doping concentration was found to be 0.03 mol, with concentration quenching observed at higher levels. Photometric analysis revealed orange color emission characteristics of the synthesized phosphors. This preliminary investigation shows that overall, the comprehensive structural, electronic, and optical characterizations in this study establish $Sm^{3+}$-doped $YPO_4$ phosphors as promising candidates for various photonic applications.


### Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1092509).


### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

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

### References


1. D. R. Golja, F. B. Dejene, and J. Y. Kim, "Trivalent Rare-Earth-Codoped Silicate Phosphor Materials ($Ba_{1.3}Ca_{0.7-x-y}SiO_4$:x$Dy^{3+}$/y$Eu^{3+}$)





15227243, 2025, 5, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/bio.70190 by Delhi Technological University, Wiley Online Library on [12/05/2025]. 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

for Solid-State Lighting," *Journal of Physics: Condensed Matter* 34 (2022): 294006.

2. D. N. Game, C. B. Palan, N. B. Ingale, and S. K. Omanwar, "Synthesis and Photoluminescence Properties of Eu Doped $LiBaPO_4$ Phosphors for Solid State Lighting," *Journal of Materials Science: Materials in Electronics* 28 (2017): 8777–8783.

3. R. G. Bora, C. B. Palan, N. S. Sawala, G. V. Korpe, and S. K. Omanwar, "Photoluminescence Properties of $KSrPO_4:Sm^{3+}$ Phosphor for SSL Applications," *Journal of Materials Science: Materials in Electronics* 29 (2018): 541–545.

4. Q. Meng, Y. Liu, Y. Fu, Y. Zu, and Z. Zhou, "Synthesis and Luminescent Properties of $Tb_3Al_5O_{12}:Ce^{3+}$ Phosphors for Warm White Light Emitting Diodes," *Journal of Molecular Structure* 1151 (2018): 112–116.

5. T. Krishnapriya, A. Jose, T. A. Jose, A. C. Saritha, C. Joseph, and P. R. Biju, "Investigation of the Structural and Photoluminescence Properties of $Eu^{3+}$ Doped $Na_6CaP_2O_9$ Phosphors for Solid State Lighting," *Materials Research Bulletin* 139 (2021): 111259.

6. P. Rohilla and A. S. Rao, "Synthesis Optimisation and Efficiency Enhancement in $Eu^{3+}$ Doped Barium Molybdenum Titanate Phosphors for w-LED Applications," *Materials Research Bulletin* 150 (2022): 111753.

7. S. Kumari, A. S. Rao, and R. K. Sinha, "Investigations on Photoluminescence and Energy Transfer Studies of $Sm^{3+}$ and $Eu^{3+}$ Ions Doped $Sr_9Y_2W_4O_{24}$ Red Emitting Phosphors With High Color Purity for w-LEDs," *Journal of Molecular Structure* 1295 (2024): 136507.

8. C. Feldmann, T. Jüstel, C. R. Ronda, and P. J. Schmidt, "Inorganic Luminescent Materials: 100 Years of Research and Application," *Advanced Functional Materials* 13 (2003): 511–516.

9. J. C. G. Bünzli and C. Piguet, "Taking Advantage of Luminescent Lanthanide Ions," *Chemical Society Reviews* 34 (2005): 1048–1077.

10. C. Lin, M. T. Berry, R. Anderson, S. Smith, and P. S. May, "Highly Luminescent NIR-to-Visible Upconversion Thin Films and Monoliths Requiring No High-Temperature Treatment," *Chemistry of Materials* 21 (2009): 3406–3413.

11. J. W. Stouwdam, G. A. Hebbink, J. Huskens, and F. C. van Veggel, "Lanthanide-Doped Nanoparticles With Excellent Luminescent Properties in Organic Media," *Chemistry of Materials* 15 (2003): 4604–4616.

12. A. A. Bhat, S. A. Khandy, A. M. Ali, and R. Tomar, "Photoluminescence Emission Studies on a Lanthanum-Doped Lead-Free Double Halide Perovskite, $La:Cs_2SnCl_6$," *Journal of Physical Chemistry Letters* 14 (2023): 5004–5012.

13. A. A. Bhat, N. Singh, R. V. Nair, E. Dujardin, and J. Sharma, "Confocal Mapping of Stable Room-Temperature Emission Centers in Gadolinium Doped Vacancy-Ordered Double Halide Perovskite, $Gd:Cs_2SnCl_6$," *Optical Materials* 141 (2023): 113937.

14. P. Ghosh, J. Oliva, E. De la Rosa, K. K. Haldar, D. Solis, and A. Patra, "Enhancement of Upconversion Emission of $LaPO_4$:Er@Yb Core-Shell Nanoparticles/Nanorods," *Journal of Physical Chemistry C* 112 (2008): 9650–9658.

15. A. A. Bhat, W. Bi, A. S. Ganie, et al., "Europium Doped $SrSnO_3$ Perovskite: Structural, Spectroscopic, and Luminescent Characterization for Advanced Lighting Technologies and Beyond," *Energy & Fuels* 38 (2024): 23726–23740.

16. K. Ariga, Q. Ji, J. P. Hill, Y. Bando, and M. Aono, "Forming Nanomaterials as Layered Functional Structures Toward Materials Nanoarchitectonics," *NPG Asia Materials* 4 (2012): e17.

17. A. A. Bhat, M. B. Zaman, J. H. Malik, K. A. Malik, I. Assadullah, and R. Tomar, "Facile Way of Making Hydrothermally Synthesized Crystalline $SrSnO_3$ Perovskite Nanorods Suitable for Blue LEDs and Spintronic Applications," *ACS Omega* 6 (2021): 16356–16363.

18. S. Nishihama, T. Hirai, and I. Komasawa, "The Preparation of Rare Earth Phosphate Fine Particles in an Emulsion Liquid Membrane System," *Journal of Materials Chemistry* 12 (2002): 1053–1057.

19. X. Qin, X. Liu, W. Huang, M. Bettinelli, and X. Liu, "Lanthanide-Activated Phosphors Based on 4f-5d Optical Transitions: Theoretical and Experimental Aspects," *Chemical Reviews* 117 (2017): 4488–4527.

20. K. A. Prokop, M. Siczek, E. Tomaszewicz, et al., "Structural Ordering Studies of $Nd^{3+}$ Ion in Eulytite-Type $M_3Y(PO_4)_3$ (M= $Sr^{2+}$ or $Ba^{2+}$) Phosphates. First Translucent Ceramics From micro-Crystalline Cubic Powders," *Ceramics International* 50, no. 5 (2023): 8042–8056.

21. R. Cao, W. Wang, Y. Ren, et al., "Synthesis, Energy Transfer and Tunable Emission Properties of $Ba_2La_2ZnW_2O_{12}:Sm^{3+}$ Phosphors," *Journal of Luminescence* 235 (2021): 118054.

22. J. Zhang, C. Chen, X. Zhang, X. Wang, W. Shi, and B. Han, "Photoluminescence Properties of $Sm^{3+}$ Doped Barium Borophosphate Phosphors," *Optik* 158 (2018): 1499–1503.

23. X. Huang and H. Guo, "$LiCa_3MgV_3O_{12}:Sm^{3+}$: A New High-Efficiency White-Emitting Phosphor," *Ceramics International* 44 (2018): 10340–10344.

24. P. Sehrawat, A. Khatkar, P. Boora, et al., "Combustion Derived Color Tunable $Sm^{3+}$ Activated $BaLaAlO_4$ Nanocrystals for Various Innovative Solid State Illuminants," *Chemical Physics Letters* 758 (2020): 137937.

25. C. Ji, Z. Huang, X. Tian, H. He, J. Wen, and Y. Peng, "Novel Red Emitting Phosphors $Mg_3Y_2Ge_3O_{12}:Sm^{3+}$ With High Color Purity and Excellent Thermal Stability Used in W-LEDs," *Journal of Alloys and Compounds* 825 (2020): 154176.

26. J. C. Chang, C. T. Chen, M. Rudysh, M. G. Brik, M. Piasecki, and W. R. Liu, "$La_6Ba_4Si_6O_{24}F_2:Sm^{3+}$ Novel Red-Emitting Phosphors: Synthesis, Photoluminescence and Theoretical Calculations," *Journal of Luminescence* 206 (2019): 417–425.

27. R. Cao, X. Wang, X. Ouyang, et al., "Thermally Stable Orange-Red Emitting $Ba_2SiO_4:Sm^{3+}$ Phosphor: Synthesis and Luminescence Properties," *Journal of Luminescence* 224 (2020): 117292.

28. Z. Yahiaoui, M. A. Hassairi, and M. Dammak, "Synthesis and Optical Spectroscopy of $YPO_4:Eu^{3+}$ Orange–Red Phosphors," *Journal of Electronic Materials* 46 (2017): 4765–4773.

29. P. M. Maleka, L. Reddy, T. J. Nkosi, et al., "Structural and Morphological Characterization of Photoluminescent Cerium-Doped Near UV-Blue Sodium Ortho-Phosphate Phosphors," *Journal of Luminescence* 226 (2020): 117409.

30. S. R. Zhang, D. Zhao, Y. C. Fan, Z. Ma, Y. L. Xue, and Y. N. Li, "Color Tunability and Energy Transfer Studies of a New $Eu^{3+}$ Doped Langbeinite-Type Phosphate Phosphor for Lighting Applications," *Journal of Solid State Chemistry* 279 (2019): 120965.

31. H. Thakur, A. K. Gathania, S. Kachhap, S. K. Singh, and R. K. Singh, "Coprecipitation Synthesis, Structural, Optical Properties, and Thermometry Application of $Tm^{3+}/Yb^{3+}$ Co-Doped $YPO_4$ Phosphor," *Journal of Luminescence* 254 (2023): 119513.

32. A. Balakrishna, L. Reddy, O. M. Ntwaeaborwa, and H. C. Swart, "Remarkable Influence of Alkaline Earth Ions on the Enhancement of Fluorescence From $Eu^{3+}$ Ion Doped in Sodium Ortho-Phosphate Phosphors," *Journal of Molecular Structure* 1203 (2020): 127375.

33. H. Lai, H. Yang, C. Tao, and X. Yang, "Preparation, Characterization and Luminescence Property of $YPO_4$:Eu Nanocrystals," *Physica Status Solidi A* 204 (2007): 1178–1184.

34. W. Da-Wei, H. Shi-Hua, Y. Fang-Tian, and T. Ye, "Scintillation Properties of $YPO_4$: RE (RE= $Ce^{3+}$, $Pr^{3+}$ or $Nd^{3+}$)," *Chinese Physics C* 33 (2009): 1019–1022.

35. S. Hachani, B. Moine, A. El-Akrmi, and M. Ferid, "Energy Transfers Between $Sm^{3+}$ and $Eu^{3+}$ in $YPO_4$, $LaP_5O_{14}$ and $LaP_3O_9$ Phosphates.



15227243, 2025, 5, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/bio.70190 by Delhi Technological University, Wiley Online Library on [12/05/2025]. 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

Potential Quantum Cutters for Red Emitting Phosphors," *Journal of Luminescence* 130 (2010): 1774–1783.

36. X. C. Jiang, C. H. Yan, L. D. Sun, Z. G. Wei, and C. S. Liao, "Hydrothermal Homogeneous Urea Precipitation of Hexagonal $YBO_3$:$Eu^{3+}$ Nanocrystals With Improved Luminescent Properties," *Journal of Solid State Chemistry* 175 (2003): 245–251.

37. B. Kahouadji, L. Guerbous, A. Boukerika, S. D. Dolić, D. J. Jovanović, and M. D. Dramićanin, "Sol Gel Synthesis and pH Effect on the Luminescent and Structural Properties of $YPO_4$:$Pr^{3+}$ Nanophosphors," *Optical Materials* 70 (2017): 138–143.

38. V. Singh, J. K. Lee, M. Seshadri, A. A. Bhat, S. Watanabe, and T. G. Rao, "Photoluminescence and EPR Spectroscopic Studies on Narrowband Ultraviolet-B (NB-UVB) Emitting Trivalent Gadolinium-Doped $CaAl_4O_7$ Material for Phototherapy Lamps," *Ceramics International* 50 (2024): 4632–4639.

39. S. Rodriguez-Liviano, F. J. Aparicio, T. C. Rojas, A. B. Hungría, L. E. Chinchilla, and M. Ocaña, "Microwave-Assisted Synthesis and Luminescence of Mesoporous RE-Doped $YPO_4$ (RE= Eu, Ce, Tb, and $Ce^+$ Tb) Nanophosphors With Lenticular Shape," *Crystal Growth & Design* 12 (2012): 635–645.

40. A. A. Bhat, S. A. Khandy, I. Islam, and R. Tomar, "Optical, Electrochemical and Photocatalytic Properties of Cobalt Doped $CsPbCl_3$ Nanostructures: A One-Pot Synthesis Approach," *Scientific Reports* 11 (2021): 16473.

41. X. Yang, X. Dong, J. Wang, and G. Liu, "Glycine-Assisted Hydrothermal Synthesis of $YPO_4$:$Eu^{3+}$ Nanobundles," *Materials Letters* 63 (2009): 629–631.

42. A. A. Bhat and V. Singh, "Exploring Double Luminescence in $Bi^{3+}$ Doped $Cs_2ZrCl_6$ Perovskite Crystals: Insight Into Multiexcitonic Emission Processes for Optoelectronic Applications," *Energy & Fuels* 38 (2024): 15681–15690.

43. W. O. Milligan, D. F. Mullica, G. W. Beall, and L. A. Boatner, "Structural Investigations of $YPO_4$, $ScPO_4$, and $LuPO_4$," *Inorganica Chimica Acta* 60 (1982): 39–43.

44. S. A. Khan, A. Jalil, Q. U. Khan, et al., "New Physical Insight Into Crystal Structure, Luminescence and Optical Properties of $YPO_4$:$Dy^{3+}$, $Eu^{3+}$, $Tb^{3+}$ Single-Phase White-Light-Emitting Phosphors," *Journal of Alloys and Compounds* 817 (2020): 152687.

45. V. Singh, A. R. Kadam, and S. J. Dhoble, "Investigation of UV and VUV Excited Photoluminescence Features of Green-Emitting $YPO_4$ Doped With $Tb^{3+}$ Phosphors," *Optik* 243 (2021): 167437.

46. U. Holzwarth and N. Gibson, "The Scherrer Equation Versus the "Debye-Scherrer Equation"," *Nature Nanotechnology* 6 (2011): 534.

47. B. Kahouadji, L. Guerbous, D. J. Jovanović, et al., "Annealing Effect on the Photoluminescence Properties of $Ce^{3+}$ Doped $YPO_4$ Nanophosphors," *Optical Materials* 91 (2019): 35–41.

48. S. Arunkumar and K. Marimuthu, "Concentration Effect of $Sm^{3+}$ Ions in $B_2O_3$–PbO–$PbF_2$–$Bi_2O_3$–ZnO Glasses–Structural and Luminescence Investigations," *Journal of Alloys and Compounds* 565 (2013): 104–114.

49. D. P. Dutta, R. S. Ningthoujam, and A. K. Tyagi, "Luminescence Properties of $Sm^{3+}$ Doped $YPO_4$: Effect of Solvent, Heat-Treatment, $Ca^{2+}$/$W^{6+}$-Co-Doping and Its Hyperthermia Application," *AIP Advances* 2, no. 4 (2012): 042194.

50. D. S. Volkov, O. B. Rogova, and M. A. Proskurnin, "Organic Matter and Mineral Composition of Silicate Soils: FTIR Comparison Study by Photoacoustic, Diffuse Reflectance, and Attenuated Total Reflection Modalities," *Agronomy* 11 (2021): 1879.

51. R. Reisfeld, "Radiative and Non-Radiative Transitions of Rare-Earth Ions in Glasses," in *Rare Earths, Radiative and Non-Radiative Transitions of Rare-Earth Ions in Glasses* (Berlin, Heidelberg: Springer Berlin Heidelberg, 1975).

52. C. P. Reddy, V. Naresh, R. Ramaraghavulu, B. H. Rudramadevi, K. R. Reddy, and S. Buddhudu, "Energy Transfer Based Emission Analysis of ($Tb^{3+}$, $Sm^{3+}$): Lithium Zinc Phosphate Glasses," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 144 (2015): 68–75.

53. V. Naresh, B. H. Rudramadevi, and S. Buddhudu, "Crossrelaxations and Non-Radiative Energy Transfer From ($^4G_{5/2}$)$Sm^{3+}$→($^5D_0$) $Eu^{3+}$:$B_2O_3$–ZnO Glasses," *Journal of Alloys and Compounds* 632 (2015): 59–67.

54. R. Alcaraz de la Osa, I. Iparragirre, D. Ortiz, and J. M. Saiz, "The Extended Kubelka–Munk Theory and Its Application to Spectroscopy," *ChemTexts* 6 (2019): 2.

55. A. R. Zanatta, "Revisiting the Optical Bandgap of Semiconductors and the Proposal of a Unified Methodology to Its Determination," *Scientific Reports* 9 (2019): 11225.

56. S. Kumari, A. S. Rao, and R. K. Sinha, "Structural and Photoluminescence Properties of $Sm^{3+}$ Ions Doped Strontium Yttrium Tungstate Phosphors for Reddish-Orange Photonic Device Applications," *Materials Research Bulletin* 167 (2023): 112419.

57. Y. Ma, S. Tang, C. Ji, et al., "$Sm^{3+}$ Doped Novel $Sr_2Ga_2GeO_7$ Based High Thermal Stability Red-Emitting Phosphors for Efficient WLED," *Journal of Luminescence* 242 (2022): 118530.

58. C. B. A. Devi, K. Swapna, S. Mahamuda, et al., "Spectroscopic Studies and Lasing Potentialities of $Sm^{3+}$ Ions Doped Single Alkali and Mixed Alkali Fluoro Tungstentellurite Glasses," *Optics & Laser Technology* 111 (2019): 176–183.

59. G. Ouertani, M. Ferhi, K. Horchani-Naifer, and M. Ferid, "Effect of $Sm^{3+}$ Concentration and Excitation Wavelength on Spectroscopic Properties of $GdPO_4$:$Sm^{3+}$ Phosphor," *Journal of Alloys and Compounds* 885 (2021): 161178.

60. S. Mahamuda, K. Swapna, M. Venkateswarlu, A. S. Rao, S. Shakya, and G. V. Prakash, "Spectral Characterisation of $Sm^{3+}$ Ions Doped Oxy-Fluoroborate Glasses for Visible Orange Luminescent Applications," *Journal of Luminescence* 154 (2014): 410–424.

61. I. Kumar and A. K. Gathania, "Photoluminescence and Quenching Study of the $Sm^{3+}$ Doped $LiBaPO_4$ Phosphor," *Journal of Materials Science: Materials in Electronics* 33 (2022): 328–341.

62. J. Xie, L. Cheng, H. Tang, et al., "Synthesis and Photoluminescence Properties of NUV-Excited NaBi $(MoO_4)_2$:$Sm^{3+}$ Phosphors for White Light Emitting Diodes," *Optics & Laser Technology* 147 (2022): 107659.

63. K. R. Kalimuthu, S. M. Babu, and V. Kalimuthu, "Synthesis and Photoluminescence Properties of $Sm^{3+}$ Doped LiGd $(WO_4)_2$ Phosphors With High Color Purity," *Optical Materials* 102 (2020): 109804.



15227243, 2025, 5, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/bio.70190 by Delhi Technological University, Wiley Online Library on [12/05/2025]. 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

Plasmonics
https://doi.org/10.1007/s11468-025-02986-6

RESEARCH

# Polarization Controlled, Dual-Band, Active Plasmonic Switch Based on T-Shaped Nanopillars on a $VO_2$-Coated Plasmonic Substrate

**Asbah Masih**[1] · **Yashna Sharma**[1]

Received: 9 March 2025 / Accepted: 22 April 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

## Abstract

We propose a polarization-controlled, dual-band $VO_2$-based plasmonic switch based on an array of T-shaped nanopillars that is capable of exhibiting a high extinction ratio > 18 dB at two operating wavelengths—one in the S-band and the other in the futuristic 2000 nm short wave infrared band—such that the wavelength of operation can be selected based on the state of polarization of the incident light. The proposed switch is designed with an array of unit cells with two T-shaped gold nanopillars placed in a face-to-face alignment along their longitudinal axis on a plasmonic substrate coated with a thin layer of vanadium dioxide. On exposure to an external trigger, $VO_2$ transforms from its semiconducting ($VO_2$(S)) to its metallic state ($VO_2$(M)). This leads to a change in the reflection spectra, and hence, switching. The proposed switches exhibit an extinction ratio (ER) ~ 19 dB at a wavelength of ~ 2030 nm when the polarization of the incident light is along the axis parallel to the longitudinal axis of the T-shaped nanostructures and at a wavelength of ~ 1500 nm when the incident light is polarized perpendicular to the longitudinal axis of the T-shaped nanostructures. Further, we also demonstrate, using finite difference time domain (FDTD) modelling, that the operating wavelengths can be fine-tuned to the C-, L-, and U- band by varying the geometric parameters of the structure. Thus, the proposed switch can potentially be used for switching in the conventional optical communication band as well as the futuristic short wave infrared band, by controlling the state of polarization of the incident light. Moreover, this switch could be potentially employed in futuristic multispectral imaging devices for the 2000 nm band.

**Keywords** Plasmonic T-shaped metasurface · Phase change materials · $VO_2$ · Polarization-tuneable switch · Dual-band switching

## Introduction

Plasmonics [1, 2] deals with the study of the interaction of light with metallic nanostructures and nanoparticles. The interaction of light with nanostructured surfaces or nanoparticles leads to the collective excitation of electrons at metallic interphases called surface plasmons [1, 3] or within the metallic nanoparticles called localized surface plasmons (LSPs) [1, 2]. In the recent past, devices exploiting the enhancement of electromagnetic radiation due to the excitation of surface plasmons, localized surface plasmons, and coupled plasmons have been employed for various applications such as switching [4, 5], chemical sensing and biosensing [6–10], light harvesting [11], subwavelength imaging [12, 13], and nanomanipulation [11, 14].

More recently, active plasmonic switches [15–18] have emerged as one of the most widely known applications of plasmonics. Active plasmonic switches can dynamically and reversibly control light propagation in photonic systems and have multiple advantages such as high speed [19], lower power consumption, high integration capability, and lower footprint [20, 21]. Plasmonic switches can be employed in telecommunication systems [15, 22], optical and plasmonic logic circuits [23], and optical communication systems like free space optical (FSO) networks [24] and wavelength division multiplexing (WDM) networks [25]. Over the years, various effects like the all-optical effect [26], acousto-optic effect [27], electro-optic effect [28], thermo-optical effect [29], pockel effect [30], phase change effect [31] and free carrier dispersion effect [32] have been used to develop plasmonic switches.

✉ Yashna Sharma
yashnasharma@dtu.ac.in

[1] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

Published online: 29 April 2025

However, in the last few decades, the development of plasmonic switches using phase change materials [33–35] has gained importance. Various phase change materials like oxides of vanadium, germanium-antimony-tellurium (GeSbTe) [36, 37], silver–indium–antimony–tellurium (AgInSbTe) [38], etc. have been used because of their high-speed switching, low power consumption, low cost, and high scalability [39–41]. Out of all these PCMs, vanadium dioxide ($VO_2$) has been widely employed for active plasmonic switching due to its potential for ultrafast femtosecond switching [42], near-room temperature phase transition [43], low thermal hysteresis [44], and its capability to undergo millions of switching cycles without degradation [45]. Vanadium dioxide undergoes a phase transition from a monoclinic semiconductor state to a tetragonal metallic state (ON State) when external stimuli such as heat (transition temperature ~ 68 °C) [46], voltage (threshold voltage above a critical voltage for phase transition) [47], or optical pulses [48] are applied to it.

Previously, many researchers have proposed active far-field plasmonic switches based on $VO_2$. Some of these include the work by Moshfeghifar et al. [49] who proposed a cavity structure consisting of a $VO_2$ ring within a nano-disk to design a reconfigurable active switch for the metal–insulator-metal (MIM) plasmonic waveguides. Similarly, an extinction ratio of 11.14 dB was achieved by Singh and Datta [50], by designing a narrowband plasmonic waveguide switch using Bragg grating by employing a periodic arrangement of $VO_2$ fins in the silicon waveguide. Miller et al. [51] also designed a waveguide switch using a silicon ridge and a same-size $VO_2$ patch on silicon substrate with an extinction ratio of 14 dB. Other works on $VO_2$-based switches in the recent past include those by Agrawal et al. [52], Kepic et al. [53], Dalal et al. [54], and Zhao et al. [55]. Although moderately high extinction ratios have been demonstrated in many $VO_2$-based plasmonic switches for a particular wavelength as discussed above, a polarization-controlled dual-band plasmonic switch based on $VO_2$ has not been previously proposed. It must also be noted that although polarization selective behavior has been previously demonstrated by employing anisotropic meta surfaces [56, 57], polarization-controlled $VO_2$-based switching in the S-band and the 2000 nm waveband has not been demonstrated previously. Further, it should be noted that in the past, T-shaped plasmonic nanostructures, employed in this paper, have been proposed for applications such as plasmonic absorbers and polarization converters [58–62].

In this paper, we propose a polarization-controlled, dual-band $VO_2$-based plasmonic switch based on an array of T-shaped nanopillars such that one unit cell of the array has two T-shaped gold nanopillars placed in a face-to-face alignment along their longitudinal axis (as shown in Fig. 1a). These periodic T-shaped nanopillars are placed on a plasmonic substrate coated with a thin layer of vanadium dioxide. When the underlying $VO_2$ transforms from its semiconducting state to its metallic state, the optical properties of the switch undergo a significant change, thus leading to a change in the reflection spectra and the characteristics of the plasmonic mode of the proposed switch. The performance of the proposed switches is measured in terms of extinction ratio (ER) which is defined as the ratio of the reflectance when the switch is in ON state ($VO_2$ is in metallic phase ($R_m$)) and the reflectance when the switch is in the OFF state ($VO_2$ is in semiconductor phase ($R_s$)) in dB.

$$\mathrm{ER} = 10 * \log_{10}\left[\frac{R_m}{R_S}\right]\mathrm{dB} \tag{1}$$

As shown in Fig. 1b, for the case when the polarization of the incident light is along the *Y*-axis, a high ER ~ 19 dB can be achieved from the proposed switches at a wavelength of ~ 1500 nm in the S-band of communication. However, when X-polarized light is incident on the proposed switch, the wavelength of the highest ER changes from 1500 to 2030 nm in the extended short wave infrared range (see Fig. 1b) without compromising on the ER. Thus, the proposed polarization-selective switch is capable of offering a high ER at any of the two wavelength bands (with one in the S-band, and the other in the extended short wave infrared band) such that the wavelength band of operation can be selected based on the state of incident polarization. It must be noted that the 2000-nm band has been recently explored for next-generation fiber optic communication [53] along with imaging, radar, and ranging applications [63]. Thus, the proposed switch can be potentially used for switching in the conventional optical communication band as well as the futuristic 2000 nm short wave infrared band, by switching the state of polarization of the incident light from 0 degree (X-polarized) to 90 degrees (Y-polarized). Further, we also demonstrate, using finite difference time domain (FDTD) modelling, that the operational wavelengths for the two states of polarization can be further fine-tuned by varying the geometric parameters of the structure.

## Numerical Modelling

The polarization-controlled switches proposed in this paper have been modelled using Ansys Lumerical FDTD which is a finite difference time domain solver [64, 65]. The reflectance spectra from the switch, with $VO_2$(S) and $VO_2$(M), are measured over a wavelength range from 1200 to 2400 nm, as the switch is illuminated by a linearly polarized, broadband plane wave source (Bloch/Periodic type) that is optimized for short pulses, and has an electric field of 1 V/m with varying angles of polarization ($\theta$) as shown in Fig. 1a. Further,

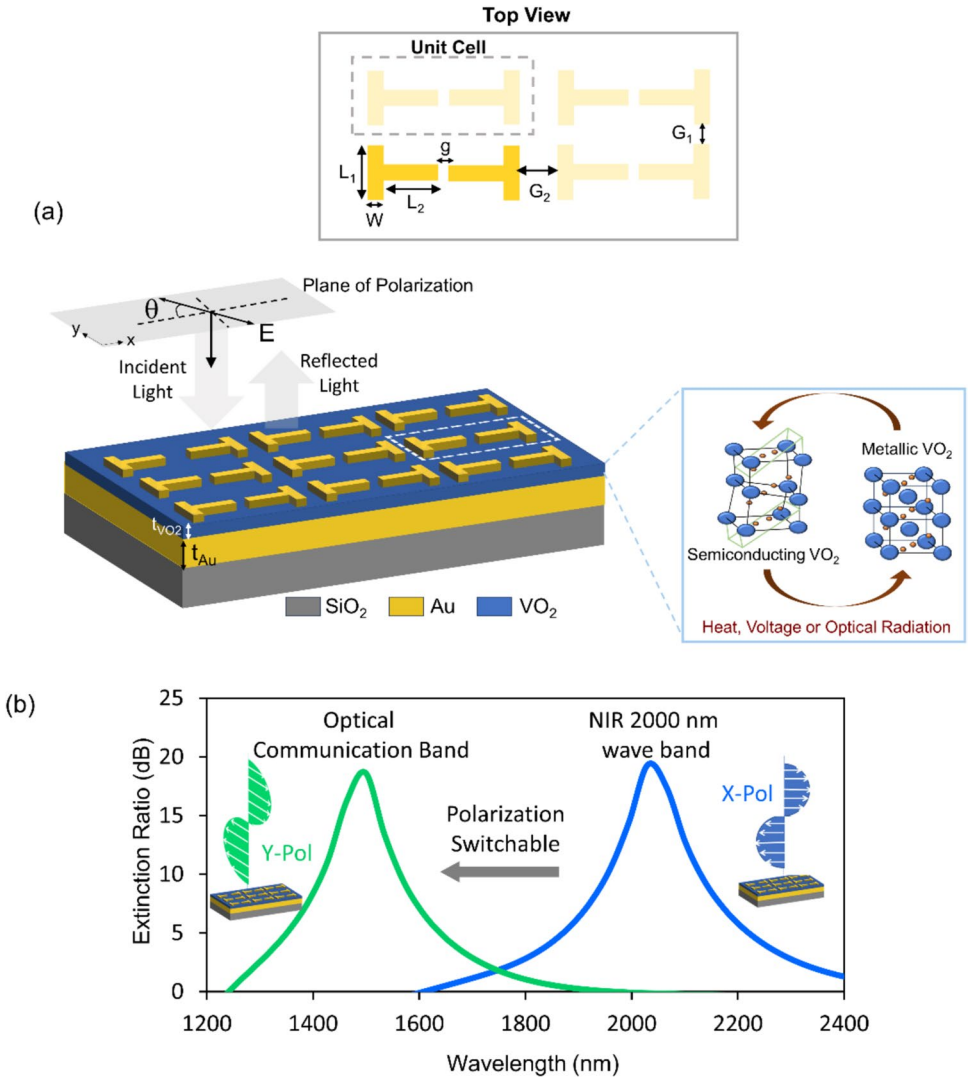



**Fig. 1** **a** Schematic of the proposed polarization-controlled, dual-band $VO_2$-based plasmonic switch based on an array of T-shaped nanopillars on a $VO_2$-coated plasmonic substrate (Au-coated $SiO_2$). The T-shaped nanopillars have lengths "$L_1$" and "$L_2$", width "$W$", and height "$h$", the gap between the two nanopillars facing each other is "$g$", the gap between the two consecutive unit cells in the $Y$- and $X$-direction are "$G_1$" and "$G_2$", respectively. The thickness of $VO_2$ layer is "$t_{VO2}$". The height of the nanopillars above the $VO_2$ layer, $h$, is taken to be 80 nm. The inset shows that on exposure to heat, voltage or optical radiation, $VO_2$ changes state from semiconducting $VO_2$ to metallic $VO_2$. **b** Extinction ratio spectra showing polarization-controlled operation of the switch at two wavelengths—~1500 nm with incident light of Y-polarization, and ~2030 nm with incident light of X-polarization

periodic boundaries are used in the $X$- and $Y$-directions while perfectly matched layer (PML) boundaries have been used in the $Z$-direction. The geometrical parameters of the proposed design are shown in Fig. 1a. The T-shaped nanostructure has length $L_1$ and $L_2$ in the $Y$- and $X$-direction respectively and width "$W$", with the height of the T-shaped nanopillar above the underlying $VO_2$ layer being $h$. Additionally, the thickness of the $VO_2$ layer is taken as $t_{VO2}$ and that of the gold layer on the $SiO_2$ substrate is taken as $t_{Au}$. One unit cell of the periodic array has two T-shaped gold nanopillars placed in a face-to-face alignment along their longitudinal axis with a separation, $g$. Further, as shown in Fig. 1, the unit cells are repeated to form an array such that the unit cells are separated from each other by a gap, $G_1$, in the $Y$-direction and gap, $G_2$ in the $X$-direction. A uniform mesh size of 1 nm has been used uniformly for the simulations after exhaustive convergence testing. It must be noted that, based on this mesh size, the theoretical maximum time step is calculated in accordance with the Courant stability criterion. For all the simulations in this paper, a time step that is 0.99 of the theoretical maximum time step is employed. The calculation and subsequent fitting of the dielectric constants for gold and $VO_2$ employed in this paper was based on the Lorentz-Drude Model [48]. The detailed discussion of the dielectric constants of the various materials employed in this manuscript is given in Supplementary Note 1 (see Supplementary Information).

## Results and Discussions

Finite difference time domain (FDTD) modelling is employed to demonstrate the polarization-controlled switching behavior in the proposed $VO_2$-based plasmonic switch based on an array of T-shaped nanopillars with a moderately high ER > 18 dB at two possible wavelengths—in the conventional communication band, and in the futuristic 2000 nm waveband. Through an exhaustive numerical analysis,

the reasons for the change in the switching behavior of the proposed switch as a function of polarization and geometrical parameters are analyzed in detail. Further, in this section, it is shown that the wavelength of operation can be fine-tuned by changing the geometrical parameters of these switches.

Figure 2a shows the reflectance spectra for the proposed switch, when $VO_2$ is in its semiconductor state for the various angles of polarization of incident light. For these simulations, a 20 nm $VO_2$ layer was taken while the geometrical lengths of the T-shaped nanopillar, $L_1$ and $L_2$ and are both taken to be 100 nm each. The width, $W$ is taken to be 30 nm, the gap between the T-shaped nanopillars facing each other in a unit cell is taken to be 10 nm, while the vertical and horizontal gap between the adjacent unit cells is taken to be 10 nm and 50 nm, respectively, to introduce an intentional anisotropy in the $X$- and $Y$-directions. It can be seen that when the angle of polarization of the incident light is 0 degree (X-polarized), a plasmonic mode at 2030 nm (Mode S1) is excited. However, when the angle of polarization of the incident light is 90 degrees (Y-polarized), the plasmonic mode is excited at 1500 nm (Mode S2). When light of any other polarization is incident on the proposed switches, with $VO_2$(S), the resultant reflectance spectra show the excitation of plasmonic modes at both these wavelengths, which is predominantly a result of the incident light splitting into its X- and Y- polarized components. As the plasmonic mode at 2030 nm is predominantly excited with X-polarized light,

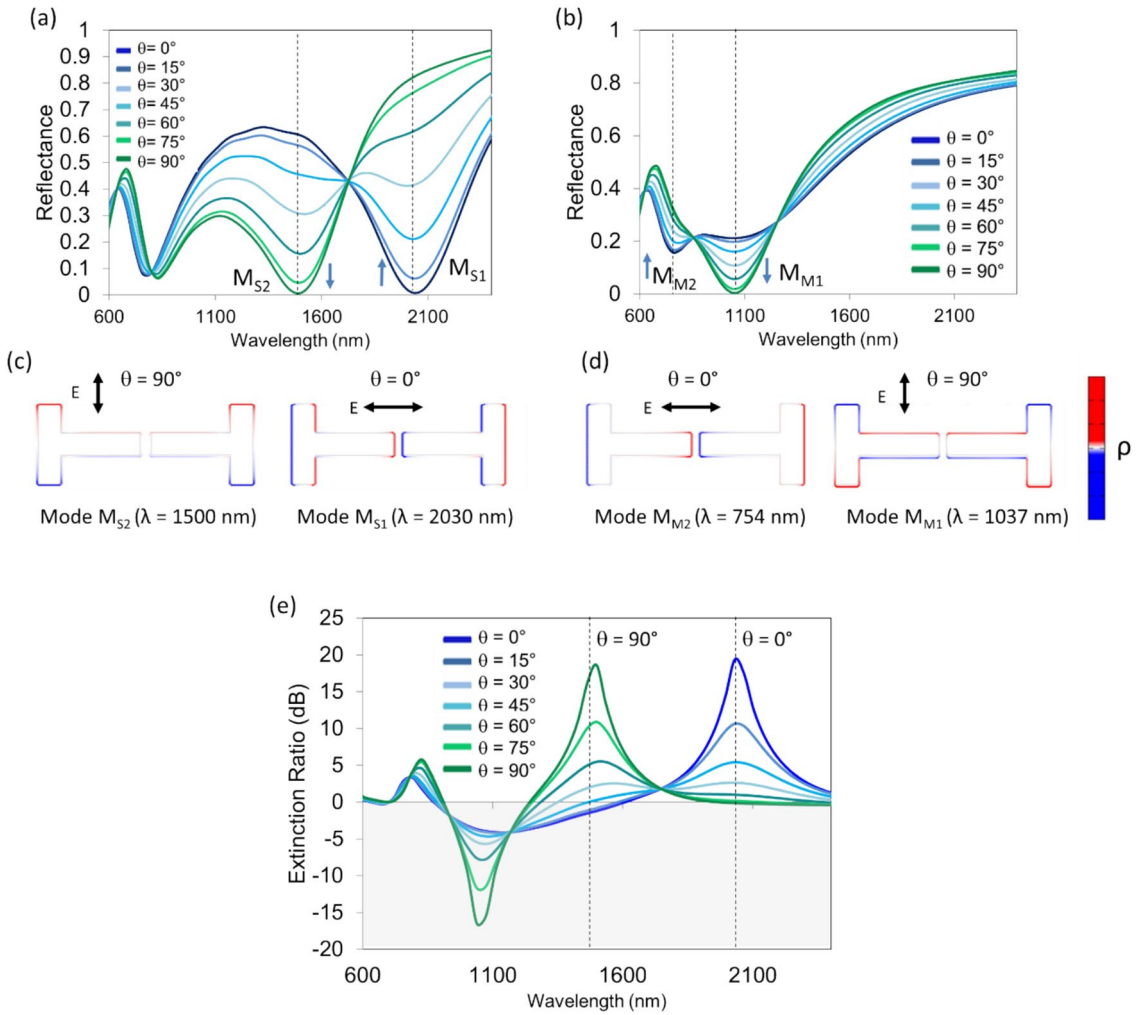



**Fig. 2** Effect of different angles of polarization ($\theta$) of the incident light on the **a** reflectance spectra of the proposed switch with VO2(S), **b** reflectance spectra of the proposed switch with $VO_2$(M), **c** charge density profiles of the proposed switch with $VO_2$(S) for $\theta = 90°$ and $\theta = 0°$ for modes $M_{S2}$ and $M_{S1}$, respectively, **d** charge density profiles of the proposed switch with $VO_2$(M) for $\theta = 0°$ and $\theta = 90°$ for modes $M_{M2}$ and $M_{M1}$, respectively, and **e** extinction ratio spectra of the proposed switch. The geometrical parameters for the switch are taken as lengths of the nanostructure arms, $L_1 = 100$ nm and $L_2 = 100$ nm; width, $W = 30$ nm; height, $h = 80$ nm; thickness of the $VO_2$ layer, $t_{VO2} = 20$ nm; gap between the two T-shaped nanopillars in a unit cell, $g = 10$ nm and vertical and horizontal gaps between the unit cells $G_1 = 10$ nm and $G_2 = 50$ nm, respectively

and the one at 1500 nm is excited with Y-polarized light, it can be seen that as the polarization angle is increased from 0 to 90 degrees, lesser incident light is coupled into the plasmonic mode at 2030 nm (the magnitude of the reflectance spectra at 2030 nm increases) as the X- polarized component of the incident light reduces. In addition, the plasmonic mode at 1500 nm becomes stronger (the magnitude of the reflectance spectra at 1500 nm reduces) as the Y-polarized component of the incident light becomes stronger. This leads to the creation of an iso-reflectance point in the reflectance spectra at 1710 nm as shown in Fig. 2a.

Figure 2b shows that when $VO_2$ transforms from its semiconductor state to its metallic state, the plasmonic modes corresponding to the X- and Y-polarized are blue-shifted to 1070 nm (Mode M1) and 790 nm (Mode M2), respectively, owing to a decrease in the real part of the refractive index of $VO_2$. Further, Mode M2 also exhibits a significant damping due to higher losses in $VO_2$(M) as compared to $VO_2$(S). As the polarization angle of the incident light is varied from 0 to 90 degrees, a similar behavior as in the case of $VO_2$(S) is observed. It can be seen that as the polarization angle of incident light is varied from 0 to 90 degrees, Mode M2 becomes progressively weaker while Mode M1 becomes stronger, leading to the creation of an iso-reflectance point at 860 nm. The charge density profiles of the modes excited in the plasmonic switch at the wavelengths corresponding to plasmonic modes S1, S2, M1, and M2 are shown in Fig. 2c, d to show the excitement of plasmonic modes in the T-shaped structures for X- and Y-polarized light, respectively. Figure 2e shows the ER spectra for the proposed switch as the polarization angle of incident light is varied. It can be seen that high ER ~ 19 dB can be selectively achieved in these switches at two wavelengths, i.e., at a wavelength of 2030 nm for X-polarization (theta = 0 degree), and at 1500 nm for Y-polarization (theta = 90 degree). As the polarization angle is varied from 0 to 90 degrees, there is a progressive decrease in the ER at 2030 nm, while a gradual increase in the ER at 1500 nm. This trend highlights the polarization selective behavior of the proposed switch, thus highlighting its application in both the conventional optical band, and the futuristic 2000-nm waveband.

The E-field enhancement ($|E|/|E_0|$) profiles for the proposed switch as a function of the polarization angle of incident light are shown in Fig. 3 for the case when $VO_2$ is in its semiconductor and its metallic state. These profiles are shown at wavelengths corresponding to the peak ER for X-polarized light (~ 2030 nm) and Y-polarized light (~ 1500 nm). It can be clearly seen from the field profiles, that when the angle of polarization of the incident light is 0° (X-polarized), the proposed switch exhibits an efficient switching behavior only at 2030 nm, where the switch exhibits an ON state with $VO_2$(S), with high field enhancement at the center of the T-shaped nanopillars, while it exhibits an OFF state with $VO_2$(M) showing a negligible field enhancement. However, it should be noted from Fig. 3b, c that, as the polarization angle is increased from 0 to 45°, the effective X-polarized component of the incident light gradually reduces, which can be seen as a reduction in the E-field enhancement between the two X-oriented arms of T-shaped nanopillars. As the polarization angle becomes larger than 45°, the switching behavior of the switch at ~ 2030 nm diminishes, and efficient switching can only be obtained at ~ 1500 nm. Figure 3b, c shows that at $\theta > 45°$, at 1500 nm, the switch exhibits an ON state with $VO_2$(S), with field localization and enhancement between adjacent T-shaped nanopillars only in the Y-direction. The field enhancement profiles show a negligibly small field enhancement in the OFF state of the switch with $VO_2$(M). Figure 4 shows that the wavelengths of maximum ER achievable using this switch can be fine-tuned by varying the periodicity of the T-shaped nanostructures in the *X*- or the *Y*-direction, i.e., by varying the gaps, $G_2$ and $G_1$, respectively (see Fig. 4a). Figure 4b, c shows that, for the polarization angle of incident light being 0°, and with the gap, $G_2$, in the *X*-direction being constant, the wavelength of maximum ER can be fine-tuned from 2030 to 1890 nm by varying the gap, $G_1$ in the *Y*-direction from 10 to 80 nm. This blue-shift in the wavelength of maximum ER can be understood from the color maps in Fig. 4a which clearly show that, with X-polarization, an increase in the gap, $G_1$, leads to a blue-shift of the plasmonic modes in the reflectance spectra of the switch, both with $VO_2$(M) and $VO_2$(S), due to the reduced dipolar interaction between the T shaped nanostructures in the adjacent unit cells. This leads to an effective fine-tuning of the wavelength of maximum ER as shown in Fig. 4c. However, it must be noted that as this gap between the adjacent unit cells of the T-shaped nanostructures is increased from 10 to 80 nm, the ER decreases from 20 to 11.5 dB, thus indicating a trade-off between the maximum achievable ER and the wavelength of operation. Figure 4d, e shows the effect of varying $G_1$, for a polarization angle of 90° on the reflectance spectra (with $VO_2$(M) and $VO_2$(S)) and on the ER spectra of the proposed switch, respectively. As the gap $G_1$ is varied from 10 to 80 nm, the plasmonic modes heavily blue-shift, resulting in the wavelength of maximum ER getting blue-shifted from 1500 to 1250 nm. This heavy blue-shift is expected as the direction of polarization of incident light (*Y*-axis) is parallel to the gap being varied in this case. Hence, the total blue-shift is significantly higher than that of the previously discussed case. However, it must be noted that the fall in the maximum ER with increasing $G_1$, is also quite significant in this case and falls below 5 dB for values of $G_1$ less than 50 nm.

The switching behavior of the proposed switch can also be tuned by varying the periodicity of the nanostructures in the *X*-direction, i.e., by varying the gap, $G_2$, while the value of gap $G_1$ remains constant. Similar to the trend observed

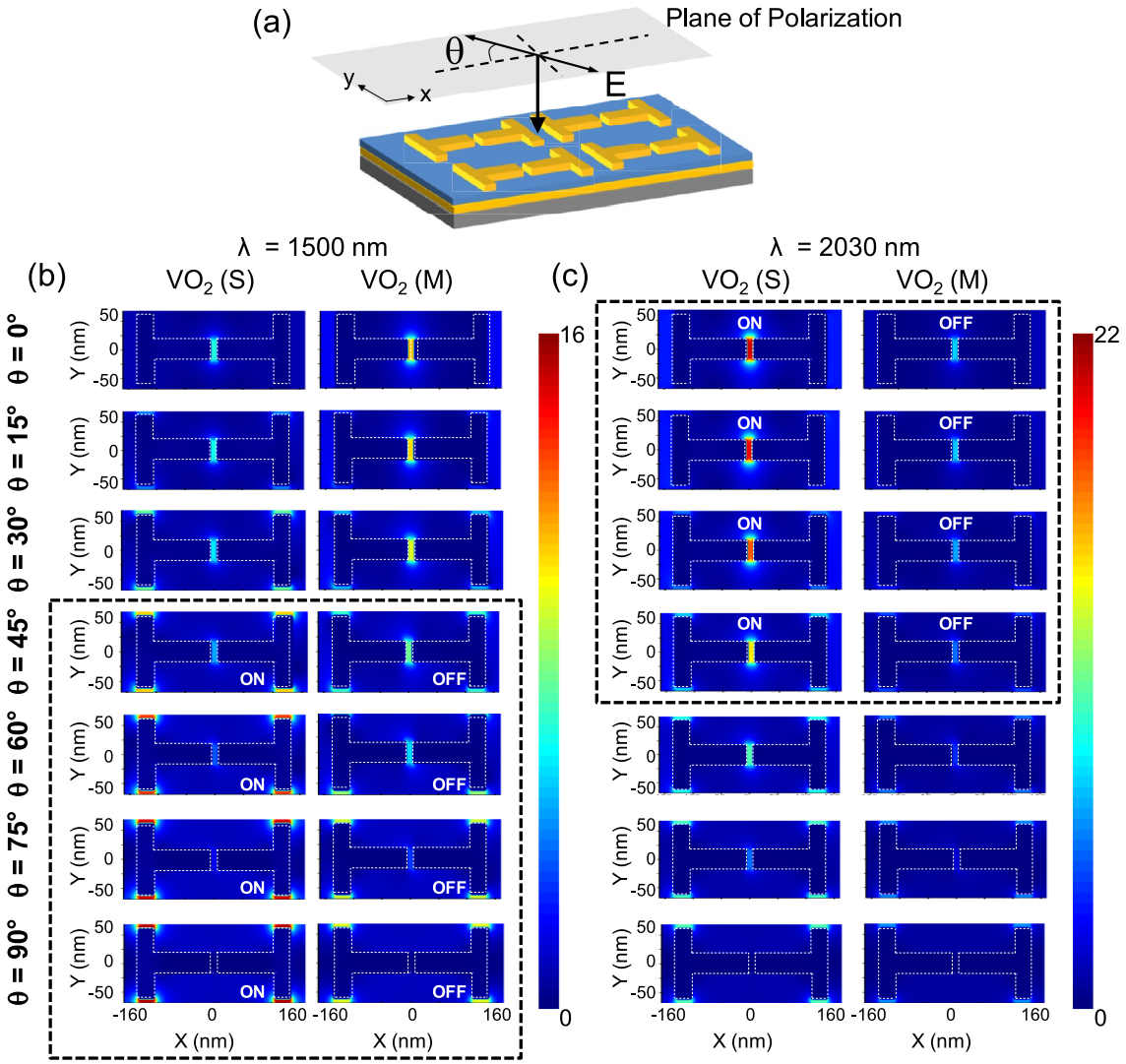



**Fig. 3** **a** Schematic of the proposed polarization-controlled, $VO_2$-based plasmonic switch showing the polarization angle of incident light, $\theta$. E-field enhancement maps for various angles of polarization of incident light, for the proposed switch with $VO_2$(S) and with $VO_2$(M) at wavelengths of **b** 1500 nm, and **c** 2030 nm. The scale bar shows the magnitude of the E-field enhancement. The geometrical parameters for the switch are taken as $L_1 = L_2 = 100$ nm, $h = 80$ nm; $W = 30$ nm; $t_{VO2} = 20$ nm; $g = 10$ nm; $G_1 = 10$ nm and $G_2 = 50$ nm

in Fig. 4b–e, with X-polarized light, Fig. 4f, g shows that the wavelength of maximum ER blueshifts from 2278 to 2030 nm when the value of $G_2$ is varied from 10 to 80 nm. However, as shown in Fig. 4h, i, when light is polarized along the *Y*-direction, i.e., perpendicular to the direction in which the gap is being varied, the wavelength of maximum ER exhibits a significantly smaller blueshift from ~ 1530 nm with $G_2$ being10 nm to ~ 1430 nm with $G_2$ being 80 nm. Hence, a variation in the periodicity of these switches in the *X*-and *Y*-direction can be utilized to fine tune the wavelength at which maximum switching is desired. In addition, the effect of gap ($g$) between the two T-shaped nanostructures and thickness of $VO_2$ layer ($t_{VO2}$) can also lead to fine-tuning of the wavelength of operation of the proposed switches and is shown in Supplementary Figs. S2 and S3, respectively (see Supplementary Information).

It was also found that the length of the two segments forming the T-shaped nanostructures, i.e., $L_1$ and $L_2$ could also be used to tune the wavelength of highest ER achievable using the proposed switches. By varying these lengths, the wavelength of maximum ER could be effectively tuned over a wavelength range > 1000 nm. In addition, it was found that the effect of lengths, $L_1$ and $L_2$, on the ER spectra was strongly dependent on the orientation of the length of the segment of the T-shaped nanopillars being varied with respect to the polarization of the incident light. Further, it should be noted that when X-polarized light is incident on the nanostructures, the ER spectra is majorly controlled by

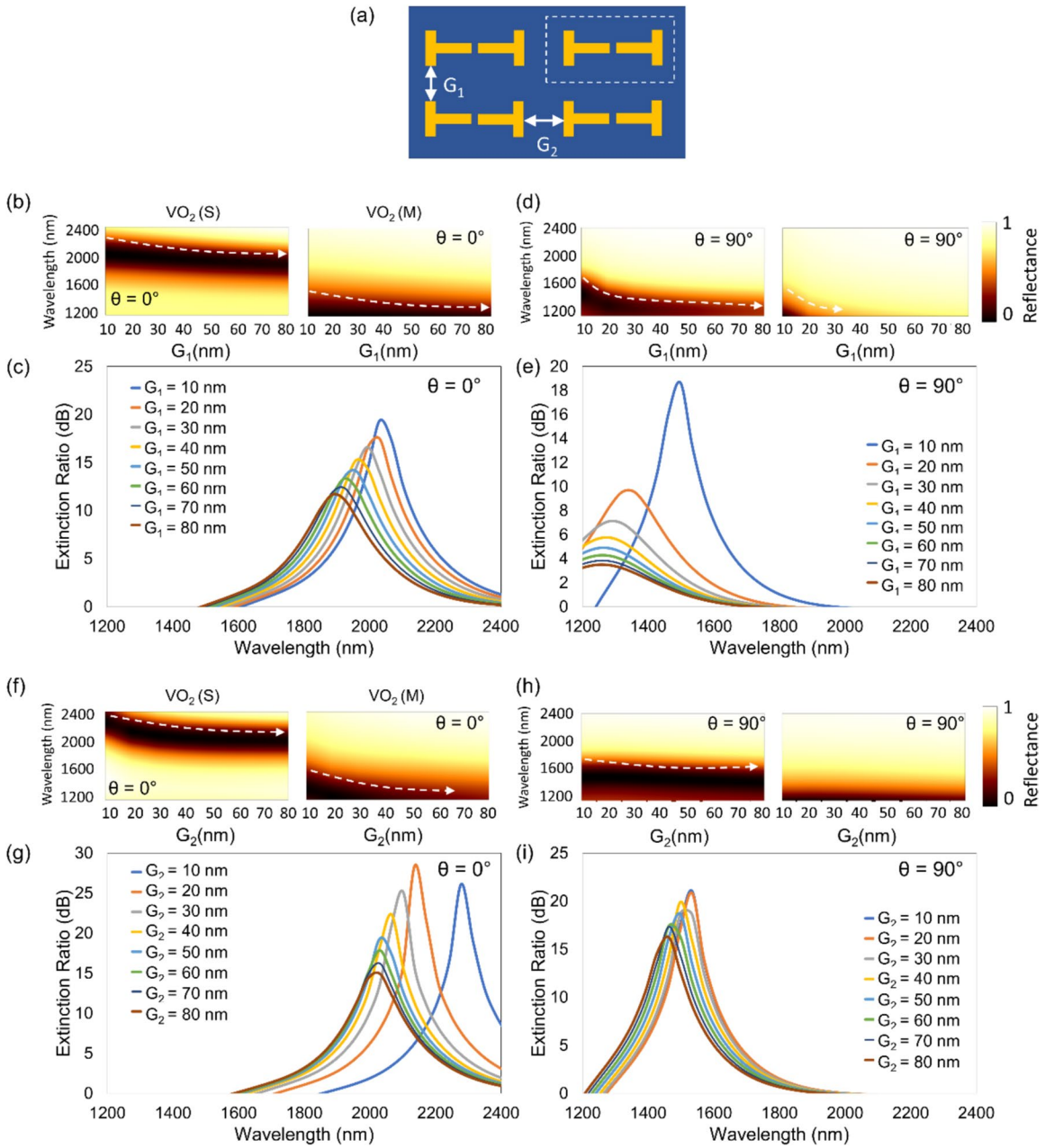



**Fig. 4** **a** Schematic of the proposed switch showing the gaps, $G_1$ and $G_2$. **b** Reflectance spectra for $VO_2$(M) and $VO_2$(S) as a function of $G_1$, and **c** extinction ratio spectra when $G_1$ is varied and $G_2$ is fixed at 10 nm for X-polarized light. **d** Reflectance spectra for $VO_2$(M) and $VO_2$(S) as a function of $G_1$, and **e** extinction ratio spectra when $G_1$ is varied and $G_2$ is fixed at 10 nm for Y-polarized light. **f** Reflectance spectra for $VO_2$(M) and $VO_2$(S) as a function of $G_2$, and **g** Extinction ratio spectra when $G_2$ is varied and $G_1$ is fixed at 10 nm for X-polarized light. **h** Reflectance spectra for $VO_2$(M) and $VO_2$(S) as a function of $G_2$ and **i** extinction ratio spectra when $G_2$ is varied and $G_1$ is fixed at 10 nm for Y-polarized light. The geometrical parameters for the switch are taken as $L_1 = L_2 = 100$ nm, $h = 80$ nm; $W = 30$ nm; $t_{VO2} = 20$ nm; and $g = 10$ nm

the plasmonic interaction between the horizontal segments of the two T-shaped nanopillars. However, for Y-polarized light, the ER spectra is majorly dependent on the plasmonic interaction between the Y-oriented vertical segments of successive unit cells. Figure S4(a)-(b) (see Supplementary Information) shows the effect of increasing length $L_1$ on the reflectance spectra of the proposed switch with $VO_2$(S) and $VO_2$(M), when X-polarized light is incident on the switch. It can be seen that increasing the length, $L_1$, from 50 to 150 nm, results in a slight blue-shift of the plasmon resonance wavelengths in the reflectance spectra of the switch. For X-polarized light, transverse plasmonic modes are excited

in the vertical arms of the T-shaped nanopillars similar to mode $M_{S1}$ discussed in Fig. 2c. Thus, when the length of these vertical segments, $L_1$, is increased, this leads to an increase in the aspect ratio of the interacting vertical segments of the T-shaped nanopillars in successive unit cells. Therefore, the blue-shift of the plasmon resonance wavelengths in Fig. S4(a)-(b) can be clearly explained by the Ganz theory which predicts that when the aspect ratio of two interacting nanorods increases, the plasmonic modes are expected to shift to smaller wavelengths [66–68]. Figure 5a shows that as the length, $L_1$, increases from 50 to 150 nm, the ER spectra for X-polarized light shows a slight blue-shift in the wavelength of maximum ER. These results are also tabulated in Table 1.

However, when Y-polarized light is incident on the structure, the effect of varying the length, $L_1$, on the reflectance spectra of the proposed switch with $VO_2$(S) and $VO_2$(M) is very significant (see Supplementary Information, Fig. S4(c)-(d)). In this case, as the incident Y-polarized light induces longitudinal modes in the vertical segments of the T-shaped nanopillars similar to mode $M_{S2}$ shown in Fig. 2d, a clear and significant red-shift is seen in the reflectance spectra of the switch, as predicted by Gans theory [66–68]. According to Ganz theory, the resonance condition gets satisfied for higher values of $\lambda$ for higher aspect ratios (or higher lengths) of interacting nanorods. Further, it can be seen from Fig. 5b, that on varying the length $L_1$ from 50 to 150 nm, the ER of the proposed switch can be tuned from being 14 dB at a wavelength of 860 nm to being 20 dB at a wavelength of 1959 nm, implying a wavelength tunability of over 1000 nm. These results are also tabulated in Table 2. A similar effect was seen on varying the length, $L_2$, of the T-shaped nanopillars. With X-polarized light, longitudinal plasmonic modes are excited in the horizontal segment of the T-shaped nanopillars. Hence, when the length of these horizontal segments, i.e., $L_2$, is increased, the

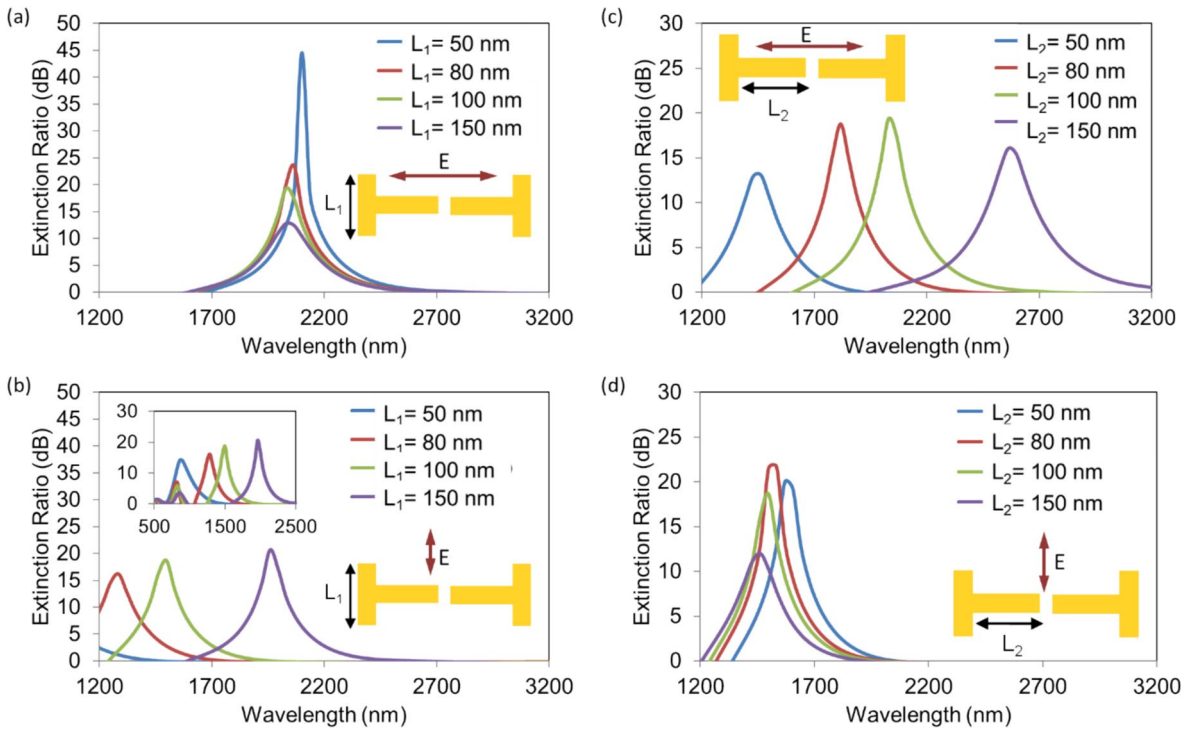



**Fig. 5** **a** Effect of the length of the T-shaped nanopillars, $L_1$, of the proposed plasmonic switch, on the extinction ratio spectra for **a** $\theta = 0°$ and **b** $\theta = 90°$, when the value of $L_2$ is kept constant as 100 nm and the effect of the length of the T-shaped nanopillars, $L_2$, on the extinction ratio spectra of the switch for **c** $\theta = 0°$ and **d** $\theta = 90°$ when the value of $L_2$ is kept constant as 100 nm. The geometrical parameters for the switch are taken as $h = 80$ nm; $W = 30$ nm; $t_{VO2} = 20$ nm; and $g = 10$ nm

**Table 1** Extinction ratio (ER) and wavelength of maximum ER for various values of length, $L_1$, of the T-shaped nanopillars

| Polarization angle | $L_1 = 50$ nm | $L_1 = 80$ nm | $L_1 = 100$ nm | $L_1 = 150$ nm |
|---|---|---|---|---|
| 0° | 44.5 dB at 2101 nm | 23.5 dB at 2065 nm | 19 dB at 2030 nm | 12.7 dB at 2030 nm |
| 90° | 14 dB at 860 nm | 16 dB at 1286 nm | 18.6 dB at 1498 nm | 20.6 dB at 1959 nm |

**Table 2** Extinction ratio (ER) and wavelength of maximum ER for various values of length, $L_2$, of the T-shaped nanopillars

| Polarization angle | $L_2 = 50$ nm | $L_2 = 80$ nm | $L_2 = 100$ nm | $L_2 = 150$ nm |
|---|---|---|---|---|
| 0° | 12.9 dB at 1427 nm | 18.7 dB at 1817 nm | 19 dB at 2030 nm | 15 dB at 2597 nm |
| 90° | 20 dB at 1569 nm | 21.6 dB at 1534 nm | 18.6 dB at 1498 nm | 12 dB at 1463 nm |

plasmon resonance wavelength in the reflectance spectra for the switch, with $VO_2$(S) and $VO_2$(M), red-shifts to longer wavelengths (see Fig. S5(a)-(b)). The ER spectrum demonstrates a similar effect, where the wavelength of maximum ER can be tuned from 1427 to 2597 nm as $L_2$ is varied from 50 to 150 nm. However, with Y-polarized light, transverse plasmonic modes are excited in the horizontal segments of the T-shaped nanopillars. Hence, an increase in the length of these horizontal segments, leads to a blue-shift in the plasmon resonance wavelength in the reflectance spectra of the switch, with $VO_2$(M) and $VO_2$(S), as predicted by Gans theory [66–68] (see Fig. S5(c)-(d)). These results are also tabulated in Table 2.

The effect of the width, $W$, of the T-shaped nanopillars, on the ER spectra is shown in Fig. 6 for X-polarized and Y-polarized incident light. It can be seen that a variation in the width of these nanopillars from 30 to 80 nm can be effectively used to significantly tune the wavelength of maximum ER for the proposed switches from ~ 2030 to ~ 2700 nm for X-polarized light, and from ~ 1498 to 1746 nm for Y-polarized light. The tunability of the wavelength of maximum ER over the various optical wavelength bands (S, C, L, and U) by varying the width, $W$, is summarized in Table 3. This progressive red-shift in the wavelength of maximum ER on varying the width, $W$, can be understood by analyzing the effect of width on the reflectance spectra for the proposed switch with $VO_2$(M) and $VO_2$(S) as shown in the Supplementary Fig. S6 (see Supplementary Information) for X- and Y-polarized light. It can be seen from Fig. S6 (a)-(b) that the plasmonic modes excited in the

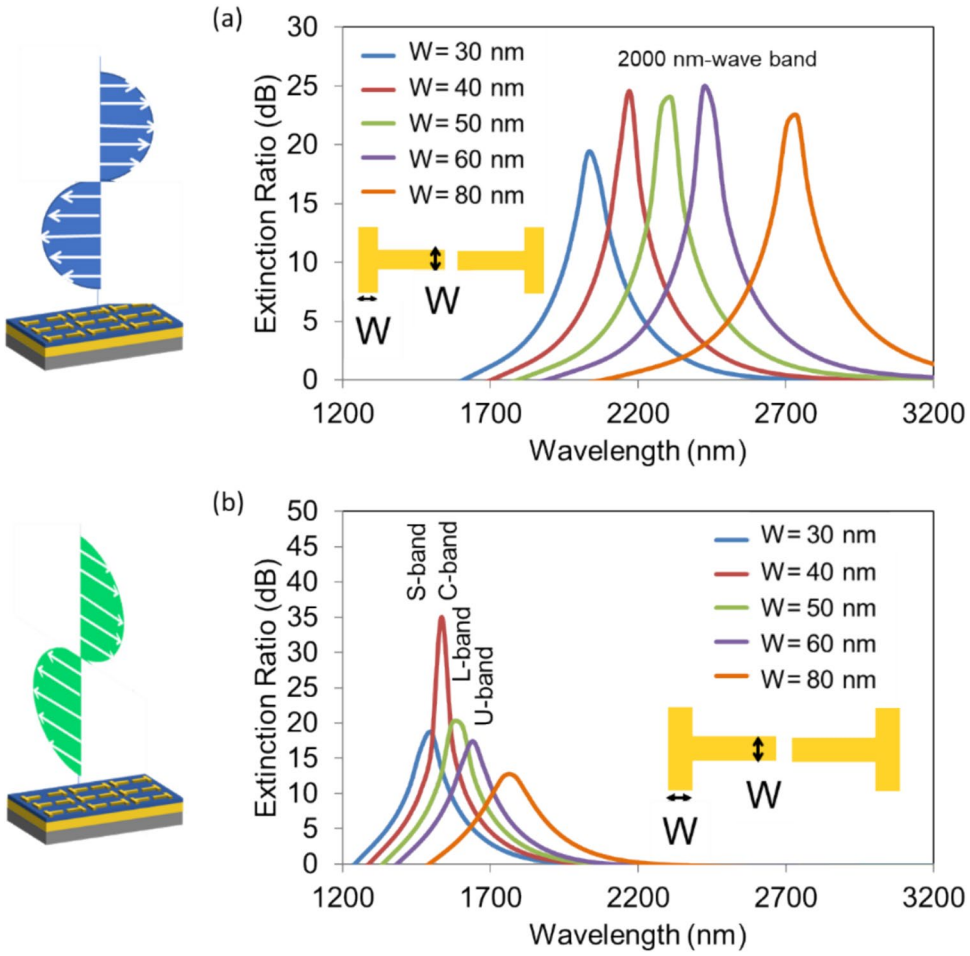



**Fig. 6** Effect of the width of the T-shaped nanopillar ($W$) of the proposed switch on the extinction ratio spectra for **a** $\theta = 0°$ and **b** $\theta = 90°$ with W varying from 30 to 80 nm. The geometrical parameters for the switch are taken as $L_1 = L_2 = 100$ nm, $h = 80$ nm; $t_{VO2} = 20$ nm; $g = 10$ nm, $G_1 = 10$ nm, and $G_2 = 50$ nm

**Table 3** Extinction ratio (ER) and wavelength of maximum ER for various values of width, *W*, of the T-shaped nanopillars

| Polarization angle | $T = 30$ nm | $T = 40$ nm | $T = 50$ nm | $T = 60$ nm | $T = 80$ nm |
|---|---|---|---|---|---|
| 0° | 19 dB at 2030 nm | 24.5 dB at 2172 nm | 23.9 dB at 2313 nm | 24.8 dB at 2420 nm | 21.7 dB at 2703 nm |
| 90° | 18.6 dB at 1498 nm (S-band) | 37.6 dB at 1534 nm (C-band) | 19.9 dB at 1569 nm (L-band) | 17.4 dB at 1640 nm (U-band) | 12.5 dB at 1746 nm |

proposed switch for X-polarized light, with both $VO_2$(M) and $VO_2$(S), exhibit a red-shift on increasing the width, *W*, from 30 to 80 nm. It should be noted that an increase in the width, *W*, effectively leads to an increase in the total length of the T-shaped nanopillars in the *X*-direction, when $L_2$ is kept constant. If the T-shaped nanopillars interacting with *X*-polarized light are approximated to be a system of interacting nanorods, this red-shift can be explained by using the Ganz theory, which infers that longer lengths of interacting nanorods imply the resonance condition getting satisfied at higher values of $\lambda$, when the incident light is polarized along the direction of the nanorod length. Thus, the plasmonic modes of the switch for both $VO_2$(S) and $VO_2$(M) exhibit a red shift with increasing, *W*, an effective red-shift is also seen in the wavelength of maximum ER as shown in Table 3. A similar effect is seen for the case when the polarization of incident light is 90 degrees (Fig. S6 (c)-(d)). This can be understood by the fact that when Y-polarized light is incident on the T-shaped nanostructures, modes similar to the ones shown in Fig. 2c, d are excited. Since the modes excited for Y-polarized light show an oscillation of the charges in the *Y*-direction, an increase in the width *W* for this case implies an increased separation of the charges in the *Y*-direction for the horizontal arms of the "T"-shaped nanostructures, thus implying a reduced intra-particle restoring force, and hence a red-shift in the resonant wavelength. For this case, it is interesting to see that the change in width of the T-shaped nanopillars from 30 to 80 nm, effectively leads to the wavelength of maximum ER spanning over the S, C, L, U-bands of optical communication as shown in Table 3.

Thus, the proposed switch can be effectively used as a polarization controlled active plasmonic switch capable of exhibiting a high ER over multiple wavelength bands, i.e., the 2000 nm waveband for X-polarized light, and the S, C, L, and U-bands with Y-polarized light. It must also be noted that for values of polarization angles other than 0 and 90 degrees, the switch exhibits two simultaneous wavelengths of operation, but with lower extinction ratios. In addition, the values of G1 less than 50 nm must be chosen to allow for a significant ER to be achieved. Further, it should be noted that the rounding radius of the corners of the T-shaped nanostructures is taken to be 3 nm in these simulations. However, variable rounding radii might be obtained based on the fabrication methodology and could thus result in slightly lower ER than those reported in this manuscript.

## Conclusion

We propose a polarization-tuneable, dual-band $VO_2$-based plasmonic switch based on an array of T-shaped nanopillars that is capable of exhibiting a high extinction ratio > 18 dB at two operating wavelengths—one in the S-band and the other in the extended short wave infrared band—such that the wavelength of operation can be selected based on the state of polarization of the incident light. The proposed switch is designed with an array of unit cells with two T-shaped gold nanopillars placed in a face-to-face alignment along their longitudinal axis on a plasmonic substrate coated with a thin layer of vanadium dioxide. On exposure to an external thermal, electrical or optical trigger, $VO_2$ transforms from its semiconducting state ($VO_2$(S)) to its metallic state ($VO_2$(M)). This leads to a significant change in the optical properties of the switch, thus leading to a change in the reflection spectra, and hence, switching. The performance of the proposed switches is measured in terms of extinction ratio which is defined as the ratio of the reflectance when the switch is in ON (with $VO_2$(M)) state to its OFF state (with $VO_2$(S)). The proposed switches exhibit an ER ~ 19 dB at a wavelength of ~ 2030 nm when the polarization of the incident light is along the axis parallel to the longitudinal axis of the T-shaped nanostructures and at a wavelength of ~ 1500 nm when the incident light is polarized perpendicular to the longitudinal axis of the T-shaped nanostructures. Further, we also demonstrate, using finite difference time domain modelling, that the operating wavelengths can be fine-tuned to the C-, L-, and U- band by varying the geometric parameters of the structure. Thus, the proposed switch can potentially be used for switching in the conventional optical communication band as well as the futuristic short wave infrared band, by controlling the state of polarization of the incident light.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11468-025-02986-6.

**Author Contributions** A.M. carried out the numerical simulations, collected the data and helped in the analysis of the data. Y. S. contributed to conceptualization of the idea, analysis of data, and manuscript

editing. The first draft of the manuscript was prepared by A. M. and was critically reviewed and edited by Y. S. All authors read and approved the final manuscript.

**Funding** The authors wish to express their gratitude towards the Delhi Technological University (DTU) for their support and resources.

**Data Availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing Interests** The authors declare no competing interests.

## References

1. Maier SA (2007) Plasmonics: fundamentals and applications. Springer, New York
2. Novotny L, Hecht B (2012) Principles of nano-optics. Cambridge University Press
3. Dereux A, Ebbesen TW (2003) Surface plasmon subwavelength optics. Nature 424:824–830. https://doi.org/10.1038/nature01937
4. Gupta N, Savaliya PB, Dhawan A (2020) Plasmonic nanoantennas on $VO_2$ films for active switching of near-field intensity and radiation from nanoemitters. Opt Express 28:27476–27494. https://doi.org/10.1364/OE.391655
5. Kim S, Yun H, Park K, Hong J, Yun JG, Lee K, Kim J, Jeong SJ, Mun SE, Sung J, Lee YW, Lee B (2017) Active directional switching of surface plasmon polaritons using a phase transition material. Sci Rep 7:43723. https://doi.org/10.1038/srep43723
6. Kravets GV, Schedin F, Jalil R, Britnell L, Gorbachev VR, Ansell D, Thackray B, Novoselov SK, Geim KA, Kabashin VA, Grigorenko NA (2013) Singular phase nano-optics in plasmonic metamaterials for label-free single-molecule detection. Nat Mater 12:304–309. https://doi.org/10.1038/nmat3537
7. Gao YK, Gan QQ, Xin ZM, Cheng XH, Bartoli FJ (2011) Plasmonic Mach-Zehnder interferometer for ultrasensitive on-chip biosensing. ACS Nano 5:9836–9844. https://doi.org/10.1021/nn2034204
8. Gopinath A, Boriskina VS, Premansiri WR, Ziegler L, Reinhard MB, Negro DL (2009) Plasmonic nanogalaxies: multiscale aperiodic arrays for surface-enhanced Raman sensing. Nano Lett 9:3922–3929. https://doi.org/10.1021/nl902134r
9. Feng J, Siu SV, Roelke A, Mehta V, Rhieu YS, Palmore GTR, Pacific D (2012) Nanoscale plasmonic interferometers for multispectral, high throughput biochemical sensing. Nano Lett 12:602–609. https://doi.org/10.1021/nl203325s
10. Tan VT, Nguyen H, Kim Y, Ok G, Lee J (2020) Photonic–plasmonic nanostructures for solar energy utilization and emerging biosensors. Nanomaterials 10:11. https://doi.org/10.3390/nano10112248
11. Dvorak P, Neuman T, Brinek L, Samoril T, Kalousek R, Dub P, Varga P, Sikola T (2013) Control and near-field detection of surface plasmon interference patterns. Nano Lett 13:2558–2563. https://doi.org/10.1021/nl400644r
12. Wei F, Lu D, Shen H, Wan W, Posetto LJ, Huang E, Liu Z (2014) Wide field super-resolution surface imaging through plasmonic structured illumination microscopy. Nano Lett 14:4634–4639. https://doi.org/10.1021/nl501695c
13. Kawata S, Inouye Y, Verma P (2009) Plasmonics for near-field nano-imaging and superlensing. Nat Photonics 3:388–394. https://doi.org/10.1038/nphoton.2009.111
14. Lin J, Mueller J P B, Wang Q, Yuan G (2013) Polarization-controlled tunable directional coupling of surface plasmon polaritons. Science 340:331–334. https://www.science.org/doi/10.1126/science.1233746
15. Carvalho WOF, Salazar JRM (2020) Plasmonics for telecommunications applications. Sensors 20:2488. https://doi.org/10.3390/s20092488
16. Mynbaev DK, Sukharenko V (2012) Plasmonic-based devices for optical communications. Int J High Speed Electron Syst 21:1250006. https://doi.org/10.1142/S0129156412500061
17. Yang J, Zhang X (2021) Optical fiber delivered ultrafast plasmonic optical switch. Adv Sci 8:2100280. https://doi.org/10.1002/advs.202100280
18. Dickson W, Wurtz GA, Evans PR, Pollard RJ, Zayats AV (2008) Electronically controlled surface plasmon dispersion and optical transmission through metallic hole arrays using liquid crystal. Nano Lett 8:281–286. https://doi.org/10.1021/nl072613g
19. Michel AU, Zalden P, Chigrin DN, Wuttig M, Lindenberg AM, Taubner T (2014) Reversible optical switching of infrared antenna resonances with ultrathin phase-change layers using femtosecond laser pulses. ACS Photon 1:833–839. https://doi.org/10.1021/ph500121d
20. Savaliya PB, Thomas A, Dua R, Dhawan A (2017) Tunable optical switching in the near-infrared spectral regime by employing plasmonic nanoantennas containing phase change materials. Opt Express 25:23755–23772. https://doi.org/10.1364/OE.25.023755
21. Sharma Y, Dhawan A (2022) Active near-field plasmonic switches based on Sierpiński-fractal nanoantennas on $VO_2$ films. J Opt 24:065001. https://doi.org/10.1088/2040-8986/ac6208
22. Karimi Y, Kaatuzian H, Tooghi A, Danaie M (2021) All optical plasmonic switches based on Fano resonance in an X-shaped resonator coupled to parallel stubs for telecommunication applications. Optik 243:167424. https://doi.org/10.1016/j.ijleo.2021.167424
23. Ghosh RR, Dhawan A (2021) Integrated non-volatile plasmonic switches based on phase-change-materials and their application to plasmonic logic circuits. Sci Rep 11:18811. https://doi.org/10.1038/s41598-021-98418-6
24. Jamali V, Ajam H, Najafi M, Schmauss B, Schober R, Poor HV (2021) Intelligent reflecting surface assisted free space optical communications. IEEE Commun Mag 59:57–63. https://doi.org/10.1109/MCOM.001.2100406
25. Papaioannou S, Kalavrouziotis D, Vyrsokinos K, Weeber JC, Hassan K, Markey L, Dereux A, Kumar A, Bozhevolnyi SI, Baus M, Tekin T, Apostolopoulos D, Avramopoulos H, Pleros N (2012) Active plasmonics in WDM traffic switching applications. Sci Rep 2:652. https://doi.org/10.1038/srep00652
26. Khani S, Danaie M, Rezaei P (2020) All-optical plasmonic switches based on asymmetric directional couplers incorporating Bragg gratings. Plasmonics 15:869–879. https://doi.org/10.1007/s11468-019-01106-5
27. Smith DA, Chakravarthy RS, Bao Z, Baran JE, Jackel JL, d'Alessandro A, Li KD (1996) Evolution of the acousto-optic wavelength routing switch. J Lightwave Technol 14:1005–1019. https://doi.org/10.1109/50.511601
28. Wang Q, Yao J (2007) A high speed 2 × 2 electro-optic switch using a polarization modulator. Opt Express 15:16500–16505. https://doi.org/10.1364/OE.15.016500
29. Nielsen MG, Bernardin T, Hassan K, Kriezis EE, Weeber J (2014) Silicon-loaded surface plasmon polariton waveguides for nanosecond thermo-optical switching. Opt Lett 39:2282–2285. https://doi.org/10.1364/OL.39.002282
30. Bishop AI, Barker PF (2006) Subnanosecond Pockels cell switching using avalanche transistors. Rev Sci Instrum 77:044701. https://doi.org/10.1063/1.2194472

31. Lee SY (2017) Design of a plasmonic switch using ultrathin chalcogenide phase-change material. Curr Opt Photon 1:239–246. https://doi.org/10.3807/COPP.2017.1.3.239
32. Sun D (2013) A proposal for digital electro-optic switches with free-carrier dispersion effect and Goos-Hanchen shift in silicon-on-insulator waveguide corner mirror. J Appl Phys 114:104502. https://doi.org/10.1063/1.4820378
33. Dalal K, Sharma Y (2024) Plasmonic switches based on $VO_2$ as the phase change material. Nanotechnology 35:14. https://doi.org/10.1088/1361-6528/ad1642
34. Farmakidis N, Youngblood N, Lee JS, Feldmann J, Lodi A, Li X, Aggarwal S, Zhou W, Bogani L, Pernice W, Wright CD, Bhaskaran H (2022) Electronically reconfigurable photonic switches incorporating plasmonic structures and phase change materials. Adv Sci 9:2200383. https://doi.org/10.1002/advs.202200383
35. Thomas A, Savaliya P, Kumar K, Ninawe A, Dhawan A (2018) Au nanowire-$VO_2$ spacer-Au film based optical switches. J Opt Soc Am B 35:1687–1697. https://doi.org/10.1364/JOSAB.35.001687
36. Heidari S, Nozhat N (2022) Wideband polarization independent plasmonic switch based on GST phase-change material. Appl Opt 61:4068–4073. https://doi.org/10.1364/AO.456423
37. Messaoud TB, Landry G, Gariepy JP, Ramamoorthy B, Ashrit PV, Hache A (2008) High contrast optical switching in vanadium dioxide thin films. Opt Commun 281:6024–6027. https://doi.org/10.1016/j.optcom.2008.09.027
38. Xiao ZF, Kun DW, Sujuan L, Nanan L, Yunqi H, Yufen R (2019) Bipolar resistive switching effect in laser-induced crystallization AgInSbTe films. Optik 180:271–275. https://doi.org/10.1016/j.ijleo.2018.11.100
39. Du Z, He C, Xin J, Song Z (2024) Terahertz dynamic multichannel holograms generated by spin-multiplexing reflective metasurface. Opt Express 32:248–259. https://doi.org/10.1364/OE.510046
40. Guo K, Zhang M, Du Z, Song Z (2024) Terahertz spin-selective metasurface for multichannel switching of OAM. Physics 56:107279. https://doi.org/10.1016/j.rinp.2023.107279
41. Xin J, Du Z, Zhou Z, Song Z (2023) Optical reflective metasurfaces enable spin-decoupled OAM and focusing physical chemistry. Chem Phys 25:27008–27016. https://doi.org/10.1039/D3CP02321D
42. Cueff S, John J, Zhang Z, Parra J, Sun J, Orobtchouk R, Ramanathan S, Sanchis P (2020) $VO_2$ nanophotonics. APL Photon 5:110901. https://doi.org/10.1063/5.0028093
43. Jorgenson GV, Lee JC (1986) Doped vanadium oxide for optical switching films. Sol Energy Mater 14:205–214. https://doi.org/10.1016/0165-1633(86)90047-X
44. Liu K, Lee S, Yang S, Delaire O, Wu J (2018) Recent progresses on physics and applications of vanadium dioxide. Mater Today 21:875–896. https://doi.org/10.1016/j.mattod.2018.03.029
45. Solanki U, Mandal P (2022) Phase sensitive $VO_2$-metal switchable plasmonic metasurface for thermal controlling of broad band near-infrared absorption. Opt Quantum Electron 54:794. https://doi.org/10.1007/s11082-022-04180-y
46. Sharma Y, Tiruveedhula VA, Muth JF, Dhawan A (2015) $VO_2$ based waveguide-mode plasmonic nano-gratings for optical switching. Opt Express 23:5822–5849. https://doi.org/10.1364/OE.23.005822
47. Kruger BA, Joushaghani A, Poon JKS (2012) Design of electrically driven hybrid vanadium dioxide ($VO_2$) plasmonic switches. Opt Express 20:23598–23609. https://doi.org/10.1364/OE.20.023598
48. Kumar N, Rua A, Aldama J, Echeverria K, Fernandez FE, Lysenko D (2018) Photoinduced surface plasmon switching at $VO_2$/Au interface. Opt Express 26:13773–13782. https://doi.org/10.1364/OE.26.013773
49. Moshfeghifar S, Abbasian K, Gilalrlue MM, Jahani MATG (2021) Active tunable plasmonic switch designed by metal–insulator–metal waveguides connected to a nanodisk cavity enabled by a phase-change material ring. Opt Eng 60:045104. https://doi.org/10.1117/1.OE.60.4.045104
50. Singh M, Datta A (2018) Modeling of a vertical hybrid plasmonic switch with $VO_2$ fin Bragg grating. IEEE Photon Technol Lett 30:997–1000. https://doi.org/10.1109/LPT.2018.2826845
51. Miller KJ, Hallman KA, Haglund RF, Weiss SM (2017) Silicon waveguide optical switch with embedded phase change material. Opt Express 25:26527–26536. https://doi.org/10.1364/OE.25.026527
52. Agrawal AK, Das A, Gupta N, Kumar K, Dhawan A (2022) Large-area plasmonic switches based on crystalline au nanoparticles between $VO_2$ layers for enhanced switching in visible-near-infrared regime. Adv Photon Res 3:2100253. https://doi.org/10.1002/adpr.202100253
53. Kepic P, Ligmajer F, Hrton M, Ren H, Menezes LS, Maier SA, Sikola T (2021) Optically tunable Mie resonance $VO_2$ nanoantennas for metasurfaces in the visible. ACS Photon 8:1048–1057. https://doi.org/10.1021/acsphotonics.1c00222
54. Dalal K, Sharma Y (2023) Broadband plasmonic switches based on nanodisc-dimers with progressively increasing diameters on a plasmonic film with a $VO_2$ spacer. Opt Commun 530:129121. https://doi.org/10.1016/j.optcom.2022.129121
55. Zhao Y, Yang R, Wang J, Wei X, Tian J, Zhang W (2022) Dual-mode terahertz broadband polarization conversion metasurface with integrated graphene-$VO_2$. Opt Commun 510:127895. https://doi.org/10.1016/j.optcom.2021.127895
56. Pelzman C, Cho YS (2015) Polarization-selective switching of extraordinary optical transmission though a metasurface. IEEE Photonics Conference (IPC), Reston, VA, USA, pp 204–205. https://doi.org/10.1109/IPCon.2015.7323651
57. Kannatassen A, Haglund FR Jr (2014) Polarization selective phase-change nanomodulator. Sci Rep 4:6771. https://doi.org/10.1038/srep06771
58. Dopf K, Moosmann C, Kettlitz SW, Schwab PM, Ilin K, Siegel M, Lemmer U, Eisler HJ (2015) Coupled T-shaped optical antennas with two resonances localized in a common nanogap. ACS Photon 2:1644–1651. https://doi.org/10.1021/acsphotonics.5b00446
59. Xuan G, Gu L, Peng R, Hu H (2024) The optical chiral properties of double-layer T-shaped plasmonic array. Plasmonics 19:159–165. https://doi.org/10.1007/s11468-023-01971-1
60. Xi Z, Lu Y, Yu W, Wang P, Ming H (2013) Improved sensitivity in a T-shaped nanodimer plasmonic sensor. J Opt 15:025004. https://doi.org/10.1088/2040-8978/15/2/025004
61. Liu Gd, Zhai X, Wang LL, Bx W, Lin Q, Xj S (2016) Actively tunable fano resonance based on a T-shaped graphene nanodimer. Plasmonics 11:381–387. https://doi.org/10.1007/s11468-015-0058-4
62. Cao W, Yang X, Gao J (2017) Broadband polarization conversion with anisotropic plasmonic metasurfaces. Sci Rep 7:8841. https://doi.org/10.1038/s41598-017-09476-8
63. Wu HY, Huang TY, Shen TP, Lee H, Oketani R, Yonemaru Y, Yamanaka M, Shoji S, Lin HK, Chang WC, Kawata S, Fujita K, Chu WS (2016) Ultrasmall all-optical plasmonic switch and its application to superresolution imaging. Sci Rep 6:24293. https://doi.org/10.1038/srep24293
64. FDTD: 3D electromagnetic simulator, Lumerical Inc. [Online]. Available: https://www.ansys.com/en-in/products/optics/fdtd
65. Trivedi R, Sharma Y, Dhawan A (2015) Plane wave scattering from a plasmonic nanowire-film system with the inclusion of nonlocal effects. Opt Express 23:26064–26079
66. Link S, El-Sayed MA (2005) Simulation of the optical absorption spectra of gold nanorods as a function of their aspect ratio

and the effect of the medium dielectric constant. J Phys Chem B 109:10531–10532. https://doi.org/10.1021/jp058091f
67. Lu H, Li L, Zhang J, Xia S, Kang X, Huang M, Shen K, Dong C, Zhang X (2019) The generalized analytical expression for the resonance frequencies of plasmonic nanoresonators composed of folded rectangular geometries. Sci Rep 9:52. https://doi.org/10.1038/s41598-018-37275-2
68. Payne KE, Shuford LK, Park S, Schatz CG, Mirkin AC (2006) Multipole plasmon resonances in gold nanorods. J Phys Chem B 110:2150–2154. https://doi.org/10.1021/jp056606x

**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.

OXFORD

*The Computer Journal*, 2025, 1–25
https://doi.org/10.1093/comjnl/bxaf065
**Original Article**

# Privacy-preserving data sharing in blockchain-enabled IoT healthcare management system

Himanshu Nandanwar and Rahul Katarya*

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

*Corresponding author. E-mail: rahulkatarya@dtu.ac.in

## Abstract

Blockchain technology offers a secure solution for managing sensitive data with Artificial Intelligence, supply chain, cloud computing, and healthcare applications. Its key features, confidentiality, decentralization, security, and privacy, enhance healthcare systems, especially when integrated with Internet of Things (IoT) devices. This integration improves communication between healthcare systems and IoT devices, increasing security, privacy, and operational efficiency. However, traditional healthcare systems face challenges like phishing, identity theft, and masquerading attacks. We propose a blockchain-based decentralized application to mitigate these risks and generate, maintain, and validate healthcare medical certificates. The application enables secure communication between hospitals, patients, doctors, and IoT devices using smart contracts for confidentiality and authentication. Our architecture utilizes Non-Interactive Zero-Knowledge Proof to maintain data integrity and privacy. We further integrate Blockchain Data Storage and the Inter-Planetary File System to reduce storage costs and enhance security through Ethereum smart contracts. An Intrusion Detection System monitors IoT traffic to detect potential security threats. Performance analysis demonstrates that this solution addresses key security and privacy challenges, offering an efficient and scalable framework for healthcare data management.

## 1. INTRODUCTION

The rapid growth and adoption of the Internet of Things (IoT) in healthcare systems have significantly advanced the smart healthcare industry by providing effective solutions for saving, maintaining, securing, and sharing healthcare documents [1]. Despite these advancements, the proliferation of IoT devices and the massive influx of healthcare data pose substantial security and privacy challenges for confidential medical documents. IoT-based healthcare models typically use cloud storage systems for data storage and maintenance [2]. However, the centralized architecture of these cloud systems introduces vulnerabilities, such as single-point failures, and an increased risk of security attacks, including masquerades, phishing, identity theft, and data breaches. Consequently, the privacy and security of healthcare documents remain primary concerns as nearly all healthcare data are stored online.

Traditional approaches in the healthcare industry have relied on encoding schemes and cryptographic methods to safeguard healthcare documents [3]. Often maintained on centralized databases or cloud storage systems, these methods are prone to single-point failure issues and various attacks [4]. Furthermore, attackers frequently clone or duplicate medical documents, creating significant challenges for users who need to access their medical records for treatment, especially if they misplace critical documents [5].

Blockchain technology offers a promising solution, innovative cryptography, and distributed systems development. Initially introduced by Santoshi Nakamoto through the cryptocurrency Bitcoin, blockchain technology features a peer-to-peer network that maintains data in a distributed ledger [6]. This decentralized approach, characterized by data integrity, transparency, and security through cryptographic algorithms, has shown potential across various domains, including finance, healthcare, and supply chain management [7]. In a blockchain, data are stored in hash format within blocks, each linked to the previous one, forming an immutable chain validated by consensus algorithms.

Integrating blockchain technology with IoT-based healthcare systems can enhance transparency, privacy, and security while reducing costs. Blockchain enables the healthcare system to maintain trust among its entities by securely managing healthcare data and verifying the authenticity of healthcare documents, thereby preventing fraud [8]. It stores healthcare details, such as unique identification numbers and certifications, in a blockchain structure, strengthening the healthcare system against counterfeiting and unauthorized modifications [9]. Smart contract functionalities within blockchain technology further enhance the security and privacy of healthcare documents by automating agreements and ensuring they are executed only when specified conditions are met. The generic architecture of blockchain-based IoT-based Healthcare Management System is shown in Fig. 1. The decentralized nature of blockchain technology and its cryptographic strengths make it an ideal solution for addressing the security and privacy challenges in IoT-based healthcare systems [10]. Researchers have proposed various blockchain-based models to enhance the security of medical records [11–13]. For instance, Rhayem *et al.* [11] developed a blockchain-based medical system using the Ethereum framework to achieve traceability without involving trusted third parties, guaranteeing data security, privacy, and transparency. Other

**Received:** August 5, 2024. **Revised:** December 20, 2024. **Accepted:** March 6, 2025
© The British Computer Society 2025. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

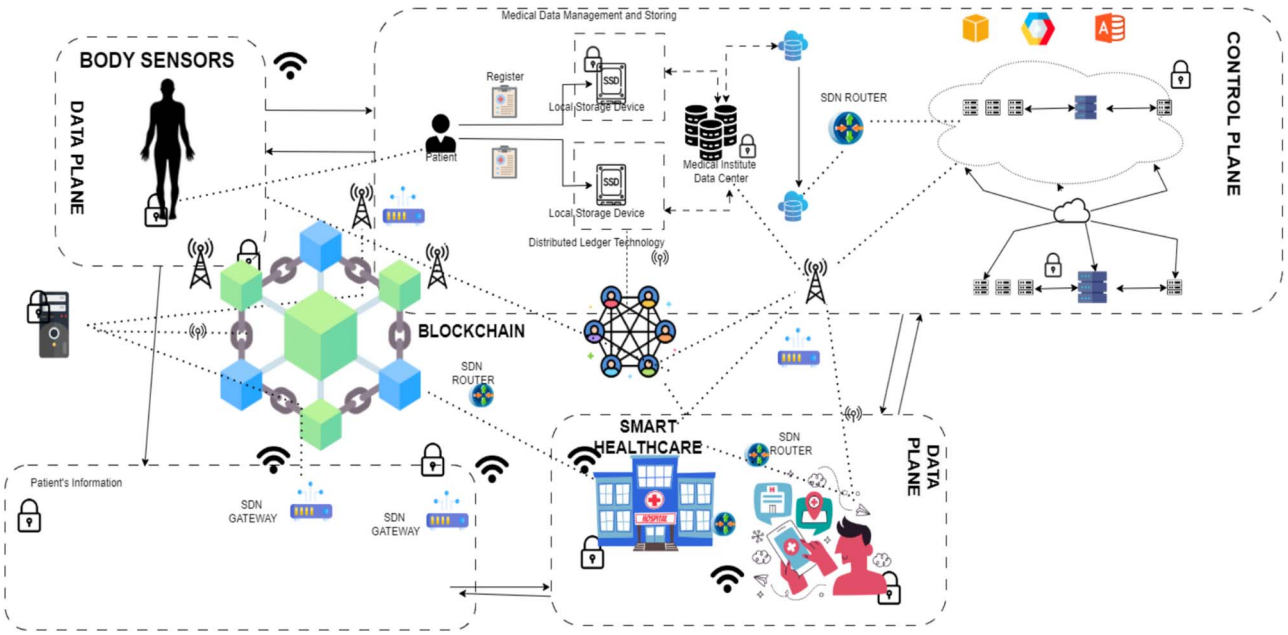



Figure 1. Blockchain-based IoT-based healthcare management system.

applications, such as the Internet of Medical Things (IoMT), utilize smart contracts to manage healthcare records, monitor sensitive data, and ensure secure access to global data.

Despite these advancements, existing systems still require comprehensive security and privacy analysis to manage healthcare data effectively. Integrating blockchain technology with healthcare can revolutionize the industry by providing a robust IoT healthcare system that ensures transparency, security, and traceability. This paper explores the various aspects of blockchain-based healthcare systems and proposes a framework for implementing a blockchain-based IoT system in the healthcare sector, emphasizing the need for secured centralized systems to protect electronic health records (EHRs) from privacy leaks, data breaches, and other security threats. The proposed solution aims to reduce storage costs, ensure patient data security, and prevent unauthorized access while maintaining data integrity and availability through blockchain's decentralized approach.

## 1.1. Motivation

In the modern healthcare ecosystem, ensuring the security, privacy, and accessibility of medical records is a critical challenge. Traditional centralized systems for managing healthcare data and medical certificates are plagued by vulnerabilities such as single points of failure, data breaches, unauthorized access, and document tampering. These inefficiencies not only compromise patient trust but also hinder seamless healthcare delivery. Additionally, the rapid proliferation of IoT devices in healthcare generates vast amounts of sensitive data, which centralized systems struggle to manage efficiently, leading to latency and scalability issues. Blockchain technology, with its decentralized, immutable, and transparent nature, offers a promising solution to address these challenges. By leveraging advanced cryptographic techniques and consensus mechanisms, blockchain ensures data integrity, enhances privacy, and eliminates the risks associated with centralized control. Despite the potential of blockchain-based solutions, existing implementations often fail to address critical requirements such as high transaction throughput, real-time scalability, and comprehensive privacy preservation. This research aims to bridge these gaps by proposing a robust blockchain-based framework that integrates IoT devices, hybrid consensus mechanisms, and privacy-preserving cryptographic techniques, thereby revolutionizing medical record management in healthcare.

This paper introduces a novel decentralized application that uses blockchain technology to enhance medical certificate management security, privacy, and efficiency in the healthcare sector. The key contributions of this work include:

- **Proposing a novel blockchain-based IoT application for healthcare:** We propose an innovative decentralized application that utilizes blockchain technology to create a secure communication medium between healthcare entities such as hospitals, patients, and doctors. This system integrates various IoT devices to maintain and update healthcare information seamlessly.
- **Introducing unique identification for medical certificates:** We introduce a mechanism to generate unique identification numbers for each medical certificate, ensuring the integrity and traceability of medical records.
- **Enhancing consensus mechanisms:** We deploy a hybrid consensus mechanism that combines Proof of Work (PoW) and Practical Byzantine Fault Tolerance (PBFT). This hybrid approach improves security and transaction speed, making the system capable of handling high transaction volumes in healthcare settings.
- **Implementing privacy-preserving computations:** We implement Fully Homomorphic Encryption (FHE) to allow secure computations on encrypted medical data. This ensures that sensitive information remains private throughout the processing stages.
- **Utilizing Non-Interactive Zero-Knowledge Proofs (NIZKPs):** We utilize NIZKPs to verify medical certificates without disclosing sensitive patient data, thus enhancing privacy and trust in the system.
- **Incorporating Interplanetary File System (IPFS):** We use IPFS as a distributed file system to distribute and store all IoT data due to the limited storage capacity available on each block.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

This approach enables the proposed architecture to handle large amounts of IoT data and scale effectively.

- **Integrating an Intrusion Detection System (IDS):** We propose an IDS to monitor IoT traffic, detecting potential security attacks and anomalies in real time. The IDS enhances overall system security by identifying and mitigating potential threats before they can cause harm.
- **Conducting comprehensive evaluation:** We perform extensive experimental tests to evaluate the proposed application on various parameters such as latency, computation time, processing time, throughput, and network usage, demonstrating its effectiveness and robustness.

The proposed framework distinguishes itself by integrating blockchain technology with IoT-based healthcare systems using a novel combination of mature techniques, including FHE, NIZKPs, and a hybrid consensus mechanism combining PoW and PBFT. While these techniques have been explored independently in other domains, their synergistic application within the context of IoT healthcare systems addresses critical gaps. Specifically, the hybrid consensus mechanism balances scalability and security, overcoming the high latency of PoW and the vulnerability of PBFT in high-volume systems. The use of FHE ensures secure computations on encrypted data, enabling privacy-preserving operations, while NIZKPs further enhance trust by allowing verification of sensitive data without exposing it. This integrated approach ensures robust performance and privacy preservation, which traditional systems fail to achieve due to their reliance on isolated or centralized methods. This research addresses these gaps by proposing a novel blockchain-based decentralized application that integrates IoT devices, implements a hybrid consensus mechanism combining PoW and PBFT, and utilizes advanced cryptographic techniques, homomorphic encryption, and NIZKPs. The proposed system is thoroughly evaluated based on multiple performance parameters, demonstrating its effectiveness and robustness in real-world healthcare settings. This research contributes to developing a more secure, efficient, and privacy-preserving healthcare data management system by addressing these gaps.

### 1.2. Paper structure

The rest of this paper is structured as follows: In Section 2, we review the existing state-of-the-art work and identify research gaps. Section 3 details the proposed architecture and its workflow. In Section 4, we analyze the security and privacy aspects of the proposed blockchain-based framework. Section 5 presents the experimental setup, results analysis, and discussion of our findings. Finally, Section 6 concludes the paper and outlines future work directions.

## 2. RELATED WORK

### 2.1. Privacy preserving in healthcare management system

In recent literature, privacy-preserving methods in the IoMT have emerged as crucial for safeguarding sensitive patient information while enhancing healthcare management efficiency [14]. Existing research extensively discusses encryption techniques, such as Advanced Encryption Standard (AES) and Rivest, Shamir, Adleman (RSA), employed to secure data at rest and in transit, effectively preventing unauthorized access [15,16]. Another noteworthy approach is signcryption [17], which combines encryption and digital signatures to provide confidentiality and data authenticity.

Studies on access control mechanisms highlight role-based access control and attribute-based access control as effective methods to restrict data access based on user roles and attributes, ensuring that only authorized personnel can access sensitive medical information [18,19]. Data anonymization techniques such as differential privacy [20] and pseudonymization [21] are frequently cited. These methods protect patient identities by adding statistical noise to datasets and replacing identifying information with tokens.

Blockchain technology has gained significant attention in the literature for its potential to enhance security in healthcare management. It introduces a decentralized, immutable ledger for recording all medical data transactions, providing security through time-stamped, tamper-resistant records. Smart contracts, as described in several studies, automate data-sharing agreements, ensuring compliance with predefined rules and patient consent.

Zero-knowledge proofs (ZKPs) are another area of interest in current research [22]. They allow the verification of data transactions without revealing sensitive information, thereby maintaining privacy. Edge computing, which processes data locally on IoMT devices, is noted for its ability to reduce the need for central storage and minimize data breach risks.

The literature also discusses patient-controlled data access systems [23] that enable individuals to manage who can access their data and for what purposes. Continuous auditing and real-time monitoring are crucial for enhancing security by promptly detecting and responding to suspicious activities. Collectively, these methods, as reviewed in existing studies, ensure robust protection and trust in the digital healthcare ecosystem.

### 2.2. Literature review

This literature review examines various blockchain-based architectures and models designed to enhance security, privacy, and efficiency in healthcare data management and sharing. One researcher details a blockchain-based telehealth architecture [23] focusing on secure storage, decentralized access control, and logging, with verified security by AVISPA (is a push-button tool for the Automated Validation of Internet Security Protocols and Applications), demonstrating computational efficiency and strong cryptographic keys. Another researcher introduces a triple-subject purpose-based access control (TS-PBAC) model [18] for the IoMT, integrating blockchain for secure, privacy-preserving data sharing. This model employs a hierarchical purpose tree, local differential privacy, and mutual evaluation metrics, providing superior patient privacy protection and stable access control decisions compared to traditional methods.

The review also covers a blockchain-assisted, secure, and privacy-preserving health data-sharing scheme for edge-based IoMT, which uses a bloom filter with hash functions for keyword ciphertext verification, key-policy attribute-based encryption for profile matching, and an incentive mechanism with a two-phase Stackelberg game for optimal pricing [24]. This protocol achieves high security, scalability, and feasibility in IoMT scenarios. Additionally, a blockchain-based smart healthcare system is highlighted for privacy-preserved electronic medical record (EMR) exchange and sharing [25]. This system employs dynamic access control, local differential privacy, and multilevel smart contracts for secure, reliable, and traceable transactions, validated by experimental results with 200 000 real-world EMRs.

The review further explores a blockchain and Artificial Intelligence (AI)-enabled model for secure medical data transmission in IoT networks called blockchain and artificial intelligence-enabled secure medical data transmission (BAISMDT) [26]. This

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

model enhances data security and privacy using signcryption, blockchain, and modified particle swarm optimization with wavelet kernel extreme learning machine, achieving high accuracy in disease detection. Another blockchain-based scheme for privacy-preserving medical data-sharing balances patient privacy with research needs, employing ZKPs and proxy re-encryption to maintain confidentiality while using PBFT for efficient distributed consensus [27–29].

Several papers focus on blockchain-based frameworks and models for secure and privacy-preserving healthcare data management. HealthRec-Chain combines blockchain and IPFS for secure, scalable health data storage, utilizing Java-enabled GPG encryption (GNU Privacy Guard (GnuPG)) and a personalized Ethereum dashboard [30]. Another blockchain-based e-health system secures EHRs using pairing-based cryptography and blockchain smart contracts for reliable transactions, addressing security challenges with minimal computation overhead [10]. Additionally, a CNN-based model combined with blockchain and federated learning [31] enhances EHR privacy, improving accuracy, data privacy, and malicious activity detection.

BCHealth, a blockchain-based architecture, enhances privacy in IoT healthcare by allowing data owners to define access policies [32]. It utilizes two chains for data and policy storage, clustering for scalability, and a modified consensus algorithm for performance. A consortium blockchain-based scheme integrates IPFS and ZKPs to ensure security and privacy in sharing personal health records, using attribute-based encryption and smart contracts for secure, personalized access [33].

Lastly, the review discusses various innovative schemes such as OHE-MSC [34], which leverages blockchain and homomorphic encryption for secure IoT medical data sharing, and BDSDT, which combines blockchain with deep learning for secure data transmission [9]. Lightweight data-sharing schemes integrating outsourcing attribute-based encryption and smart contracts are also examined [35]. These approaches enhance security, efficiency, and fault tolerance in IoMT data sharing, demonstrating the transformative potential of blockchain and AI technologies in healthcare management [8,36–38].

Existing blockchain-based healthcare systems typically focus on either enhancing data security or improving system scalability but rarely address both comprehensively. For instance, systems employing PBFT offer faster consensus but struggle with scalability, while PoW-based models prioritize security at the cost of high computational overhead. Moreover, while cryptographic techniques like FHE or NIZKPs have been explored individually, their integration into a unified framework remains underexplored. This paper bridges these gaps by combining these techniques into a cohesive system, providing both scalability and security without compromising efficiency.

### 2.3. Research gaps

Despite the potential benefits of blockchain technology in healthcare, several gaps exist in current research and implementations:

- Integration with IoT devices: While there is a significant focus on blockchain technology in healthcare, there needs to be more research on effectively integrating blockchain with IoT devices to ensure real-time, secure data acquisition and management.
- Scalability and performance: Most existing blockchain-based healthcare applications use traditional consensus mechanisms like PoW, which are inefficient in handling high transaction volumes typical in healthcare settings. There is a need to explore and implement more efficient consensus mechanisms that can enhance the scalability and performance of these systems.
- Privacy-preserving techniques: Current implementations often fail to adequately protect patient privacy while allowing necessary data verification and access. Techniques like ZKPs and homomorphic encryption have been proposed in theory but lack practical implementation and evaluation in healthcare applications.
- Comprehensive evaluation: Many blockchain-based healthcare solutions lack thorough experimental evaluations to test their effectiveness, robustness, and efficiency under real-world conditions. There is a need for extensive testing and analysis based on various performance metrics to demonstrate the practical viability of these solutions.

This research addresses these gaps by proposing a novel blockchain-based decentralized application that integrates IoT devices, implements a hybrid consensus mechanism combining PoW and PBFT, and utilizes advanced cryptographic techniques such as homomorphic encryption and ZKPs. The proposed system is thoroughly evaluated based on multiple performance parameters, demonstrating its effectiveness and robustness in real-world healthcare settings. This research contributes to developing a more secure, efficient, and privacy-preserving healthcare data management system by addressing these gaps.

## 3. PROPOSED METHODOLOGY

This section proposes a decentralized application that utilizes blockchain technology to generate and maintain medical certificates through various phases, such as data acquisition, validation, representation, and justification. User can maintain their healthcare information using different IoT devices. The proposed applications aim to prevent fraud in maintaining and issuing user medical records, including medical test reports, discharge summaries, and operative reports. Figure 2 presents the workflow of the proposed blockchain-based IoT decentralized application for the healthcare management system.

The diagram presents a decentralized healthcare management system that leverages blockchain technology, NIZKPs, and the IPFS to enhance the privacy and security of medical certificate management. It delineates the interactions among IoT users, devices, and the healthcare management system, which collectively facilitate secure access and validation of medical data. On the left, the IoT user engages with the healthcare system through an IoT app and connected devices that transmit user data. Conversely, the right side illustrates the blockchain network's critical role in ensuring data integrity and security through nodes, identity managers, and consensus mechanisms. Smart contracts automate certificate issuance and verification, while ZKP ensures data verification without exposing sensitive information.

The workflow initiates with user registration (Step 1) and the issuance of unique credentials (Step 3), secured by the blockchain's identity manager. Upon requesting a medical certificate (Step 5), the healthcare system returns credentials (Step 4), enabling the user to present them to the decentralized application (Step 6). The application verifies identity and processes the request (Step 7), leading to the issuance of a unique certificate ID (Step 8). This ID links to the medical certificate stored in IPFS, from which the user retrieves the certificate (Steps 9 and 10). The system's architecture, supported by a

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

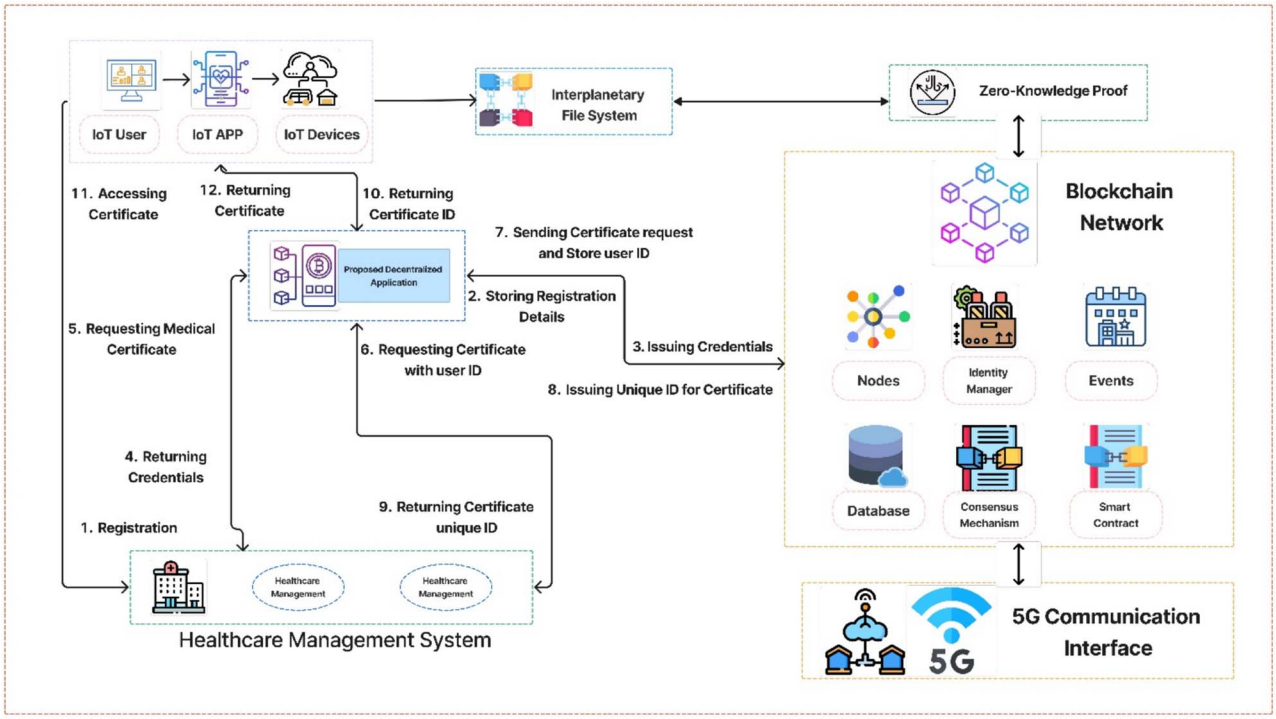



Figure 2. Proposed workflow of the novel Blockchain-based IoT application for healthcare management system.

5G communication interface, facilitates real-time interactions among its components, ensuring a robust, secure, and efficient healthcare management solution.

## 3.1. Architecture and workflow

The proposed blockchain-based application is designed to operate within a distributed network that seamlessly integrates blockchain technology with IoT devices to enhance healthcare management. The blockchain network is at the heart of the system, which provides a secure and immutable ledger for managing healthcare data.

User registration and verification is the first step in the process. Hospitals, patients, doctors, and IoT devices are registered using a smart contract on the Ethereum blockchain. The hospital authority manages this registration, issuing credentials to each entity. For hospitals and doctors, the registration involves sending a request through the user portal, which the blockchain network verifies. Registration requests are also processed for patients, generating credentials that allow them to access application services. IoT devices, such as smartwatches, are similarly registered by transmitting a request containing device-specific details like Device ID. The blockchain verifies and stores these credentials, associating each device with its user profile.

Integration of IoT devices plays a critical role in the system. IoT devices collect real-time health data from users, such as heart rate and activity levels. This data are securely transmitted to the blockchain network, where it is encrypted using advanced techniques to ensure privacy. Each data point is assigned a unique identifier on the blockchain, facilitating secure and immutable storage. This integration ensures that health metrics are recorded accurately and are protected against tampering or unauthorized access.

When patients seek medical services, they submit a service request through the application. The hospital authority then verifies the request and the patient's data via the blockchain network, ensuring that all interactions are secure and authenticated. Once verification is complete, the hospital can issue a medical certificate. The application processes this certificate, which assigns a unique identification number and records it on the blockchain. This ensures the certificate's authenticity and provides a tamper-proof record that authorized parties can easily access. Certificate access is streamlined through the use of unique identifiers. Both users and hospitals can retrieve the generated medical certificate using this unique ID. The blockchain's distributed ledger ensures that access to this information is secure and controlled, reducing the risk of unauthorized access or modifications.

The application integrates a hybrid consensus mechanism combining PoW and PBFT to secure the network. The PoW component requires computational work to validate and record transactions, enhancing data integrity and preventing unauthorized modifications. Simultaneously, PBFT ensures consensus among network participants by providing robustness against faulty or malicious nodes, improving overall system reliability and efficiency. This hybrid approach leverages the strengths of both algorithms to maintain a secure, resilient, and efficient blockchain network.

The system features a graphical user interface that facilitates interaction with the blockchain network. This intuitive interface allows hospitals, patients, and doctors to quickly manage and access their data, enhancing user experience and operational efficiency. Finally, the Distributed Ledger Management ensures that all healthcare data are protected from unauthorized access. The blockchain supports secure operations such as data insertion, deletion, and updates, maintaining the confidentiality and integrity of sensitive medical information. Figure 3 represents a decentralized healthcare system using blockchain and IoT for secure medical certificate management. It includes key components like IoT devices, consensus mechanisms, homomorphic encryption, IPFS, NIZKP, and IDS for secure communication, certificate validation, and real-time threat detection.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

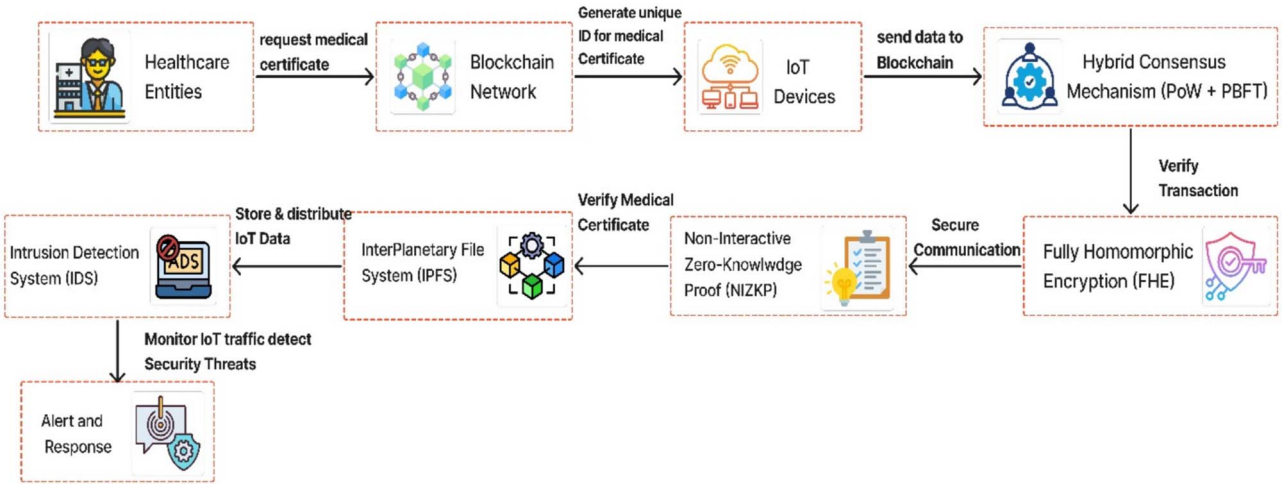



Figure 3. Workflow of a Blockchain-based IoT healthcare application with homomorphic encryption, zero-knowledge proofs, and intrusion detection for secure medical certificate management.

By integrating blockchain technology with IoT devices, the application creates a robust, secure, and efficient system for managing medical records and certificates. This innovative approach enhances healthcare data management's accuracy, security, and transparency, addressing many challenges associated with traditional centralized systems.

## 3.2. Integration of Proof of Work and Practical Byzantine Fault Tolerance

The PoW consensus algorithm creates new blocks in the blockchain network, ensuring that blocks are added securely and consistently. Each block contains a hash of the previous block, a timestamp, and transaction data. The goal is to find a nonce $n$ such data that the hash of the block denoted as $H$, meets a specific target $T$:

$$H = hash(prev_{hash}, timestamp, transaction, n)$$
$$H < T$$

Minor repeatedly changes the nonce $n$ and recompute the hash until they find a valid one. Following are the steps in the proposed model:

Step 1: When a new medical certificate is generated, it forms a new term transaction.

Step 2: Miners in the blockchain network collect these transactions and form a new block.

Step 3: They compute the hash with different nonce values until they find a valid one that meets the target $T$.

Step 4: The first miner to find a valid nonce broadcasts the block to the network.

Step 5: Other nodes verify the block and add it to their blockchain network.

The PBFT is used to enhance the security and speed of the consensus process, ensuring that the system can handle high transaction volume. It contains five phases:

**Phase 1: Initialization phase**

A client (e.g. a hospital) sends a request to the primary node (leader) to execute a service operation. This can be represented as:

$$Request = (operation, Timestamp, Client\ ID)$$

$$Client \rightarrow Primary : \{Request\}$$

**Phase 2: Pre-prepare Phase**

The primary note receives the client request and multicasts a pre-prepare message to all replicas. The pre-prepared message is denoted as:

$$\{pre-prepare, V, n, d\}$$
$$Primary \rightarrow Replicas : \{pre-prepare, V, n, d\}$$

where $V$ is the view number, $N$ is the sequence number assigned to the request, and $d$ is the digest (hash) of the request completed as $d = \text{hash (request)}$.

**Phase 3: Prepare phase**

Each replica verifies the pre-prepared message. If valid the replica multicasts a prepare message to all other replicas.

$$\{prepare, V, n, d, i\}$$
$$\forall Replica\ i : \{prepare, V, n, d, i\}$$

where $i$ is the replica ID.

Each replica waits to receive $2f + 1$ prepare message (including its own), where $f$ is the maximum number of faulty replicas the system can tolerate.

$$Wait\ for\ 2f + 1\ prepare\ message$$

**Phase 4: Commit phase**

Once a replica has received $2f + 1$, prepare message, it multicasts a commit message as:

$$\{Commit, V, n, d, i\}$$
$$\forall\ Replica\ i : \{Commit, V, n, d, i\}$$

Each replica waits to receive $2f + 1$ commit message (including its own).

$$Wait\ for\ 2f + 1\ commit\ message$$

**Phase 5: Reply phase**

When a replica has received $2f+1$ commit message, it executes the requested operations and sends a reply to the client:

$$Reply = (Result, Timestamp, Client\ ID)$$

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

$$Replica \rightarrow Client : (Reply)$$

The following are the steps in the proposed model:

Step 6: When a hospital requests to add a new medical certificate, the request is processed through the PBFT consensus.

Step 7: The primary node initiates the PBFT process.

Step 8: Replicas go through pre-prepare and commit phases to agree on adding the new block.

Step 9: The request is executed, and the result (new block) is added to the blockchain.

Considering a scenario where patient $P$ requests a medical certificate from the hospital $H$. The PBFT consensus ensures that this request is processed securely and efficiently. Patient $P$ sends a request $R$ to the primary node (hospital $H$): $P \rightarrow H : R$. Hospital $H$ (primary node) assigns a sequence number and creates a pre-prepared message as:

$$d = hash(R)$$

Hospital $H$ broadcasts the pre-prepared message to all replica nodes (other hospitals/doctors):

$$H \rightarrow Replicas : (Pre - prepare, V, n, d)$$

Replica nodes verify the pre-prepared message and broadcast the prepared message:

$$Replica \rightarrow Replicas : (prepare, V, n, d, i)$$

The replica node counts the prepare message, and upon receiving $f + 1$ prepares, they broadcast the commit message:

$$Replica \rightarrow Replicas : (commit, V, n, d, i)$$

The replica node counts the commit message; upon receiving the $2f + 1$ commit, they execute the transactions and send the result to the patient:

$$Replica \rightarrow P : (Medical\ Certificate)$$

The proposed blockchain-based healthcare system can achieve higher security and faster transaction speeds by enhancing the consensus mechanism. Figure 4 illustrates the workflow of the PBFT mechanism, detailing the interaction phases between the client and nodes during transaction processing and consensus. The algorithm 1 outlines a consensus process involving request, pre-prepare, prepare, commit, and reply phases, ensuring transaction validity and execution across distributed nodes.

Algorithm 1. Working of Consensus Mechanism in the proposed application

**Algorithm 1.1.: Request Phase**
**Input:** Client request R
**Output:** Transaction initiation

1. Client sends request R to the primary node: $Client \rightarrow Primary : \{Request\}$

**Algorithm 1.2.: Pre-Prepare Phase**
**Input:** Request R
**Output:** Pre-prepare message

1. Primary node assigns a sequence number $n$ and creates a pre-prepare message:$(Pre - prepare, V, n, d)$, where V is the view number, and $d$ is the digest of R: $d = hash(R)$
2. Primary node broadcasts the pre-prepare message to all replica nodes: $Primary \rightarrow Replicas : (Pre - prepare, V, n, d)$

**Algorithm 1.3.: Prepare Phase**
**Input:** Pre-prepare message $(Pre - prepare, V, n, d)$
**Output:** Prepare message

1. Replica nodes verify the pre-prepare message: if $(Pre - prepare, V, n, d)$ is valid
2. Replica nodes create and broadcast prepare messages: $(prepare, V, n, d, i)$ , where i is the node identifier.
3. Replica nodes send prepare messages to all other nodes: $Replica \rightarrow Replicas : (prepare, V, n, d, i)$

**Algorithm 1.4.: Commit Phase**
**Input:** Prepare messages $(prepare, V, n, d, i)$
**Output:** Commit message

1. Replica nodes count the prepare messages. If $f + 1$ prepares are received (where $f$ is the maximum number of faulty nodes), they create and broadcast commit messages: $(commit, V, n, d, i)$
2. Replica nodes send commit messages to all other nodes: $Replica \rightarrow Replicas : (commit, V, n, d, i)$

**Algorithm 1.5.: Reply Phase**
**Input:** Commit messages $(commit, V, n, d, i)$
**Output:** Transaction execution and client reply

1. Replica nodes count the commit messages. If $2f + 1$ commits are received, they execute the transaction: $execute(R)$
2. Replica nodes send the result to the client: $Replica \rightarrow Client : (Reply(R))$

## 3.3. Fully homomorphic encryption for data privacy

FHE is a powerful cryptographic method that enables computations on encrypted data without decryption, supporting addition and multiplication operations. This capability makes FHE more suited to complex medical scenarios than semihomomorphic techniques like Paillier, which are limited to addition operations. FHE protects sensitive medical data in the proposed healthcare blockchain architecture, allowing comprehensive analyses and computations. This subsection explains the utilization of fully homomorphic encryption and illustrates an example of how FHE works in the proposed architecture.

**Encryption and decryption process in FHE**

In FHE, plain text m is encrypted into cipher text C using an encrypted function E:

$$C = E(m)$$

FHE allows both addition and multiplication operations on cipher texts, making it highly flexible for complex computations:

- Addition: $E(m_1).E(m_2) mod\ n^2 = E(m_1 + m_2)$
- Multiplication: $E(m_1)^{m2}\ mod\ n^2 = E(m_1 * m_2)$

The decryption process using a decryption function D, retrieves the original plain text m from cipher text C:

$$m = D(C)$$

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

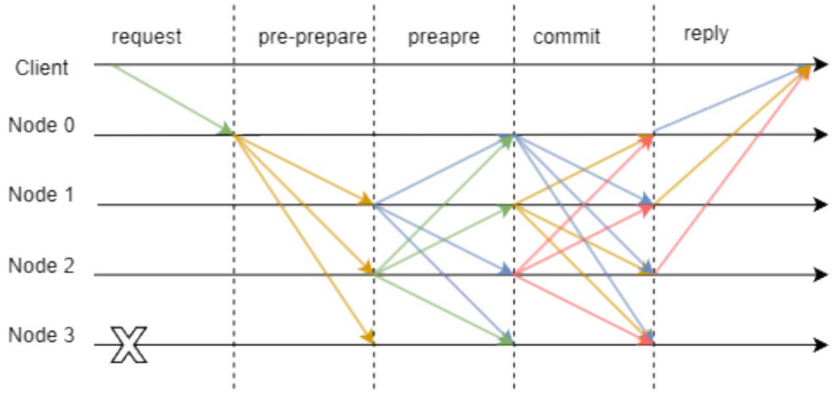



Figure 4. Workflow of PBFT mechanism in the proposed framework at different phases.

**Example: Computation of total medical expenses using FHE**

Consider a scenario where a hospital must calculate a patient's total medical expenses while keeping individual expenses confidential. With FHE, such computations can be carried out on encrypted data without compromising privacy.

**Key generation**

- Two large prime numbers $p = 31$ *and* $q = 37$ are chosen.
- Compute $n = p * q = 31 * 37 = 1147$.
- Public and private keys are generated to support addition and multiplication operations on encrypted data using $\lambda = lcm\,(p-1, q-1), g \in Z^*_{n^2}, \mu = (L(g^\lambda \bmod n^2))^{-1}$ mod n, Where $L(x) = \frac{x-1}{n}$

**Encryption:** Two medical expenses, $m_1 = 200$ *and* $m_2 = 150$, are encrypted using random values $r_1 = 5$ *and* $r_2 = 7$.

$$C_1 = E(m_1) = g^{m_1} * r_1^n \bmod n^2$$
$$C_2 = E(m_2) = g^{m_2} * r_2^n \bmod n^2$$

**Homomorphic operations:** Using FHE, both addition and multiplication can be performed on the encrypted data without requiring decryption:

$$C_{sum} = C_1.C_2 \bmod n^2 = E(m_1 + m_2)$$
$$C_{product} = C_1^{m2} \bmod n^2 = E(m_1 * m_2)$$

**Decryption:** After performing the operations, the ciphertext is decrypted using the private key to reveal the results:

$$m_{sum} = D(C_{sum}) = m_1 + m_2 = 350$$
$$m_{product} = D\left(C_{product}\right) = m_1 * m_2 = 30,000$$

This example illustrates the versatility of FHE, allowing for complex computations beyond basic addition, which is crucial in scenarios that require secure financial calculations or statistical analysis of sensitive medical data. By utilizing FHE, the proposed healthcare blockchain system facilitates privacy-preserving computations on encrypted data, allowing advanced medical operations without revealing sensitive information. FHE enhances system security by supporting secure computations, such as billing, patient statistics, and resource management, within a decentralized architecture. This ensures that patient data remains confidential even in complex medical scenarios while enabling more flexible computations than semi-homomorphic encryption methods can provide.

Implementing the FHE for data privacy in the healthcare blockchain system ensures the privacy of sensitive medical information while enabling secure computation on encrypted data. This approach enhances security and maintains patient confidentiality throughout the data processing phase, as shown in Fig. 5. The above example demonstrates how homomorphic encryption is utilized in the proposed architecture, enabling secure and private computation. Algorithm 2 describes a cryptographic system's key generation, encryption, decryption, and homomorphic addition, enabling secure and privacy-preserving computations on encrypted data.

Algorithm 2. Homomorphic Encryption process in the proposed application

**Algorithm 2.1.: Key Generation**
**Input:** Two large prime numbers: $p$ *and* $q$
**Output:** Public key $(n, g)$ and a private key is $(\lambda, \mu)$.

1. Choose $p$ *and* $q$
2. Compute $n = pq$
3. Compute $\lambda = \mathrm{lcm}\ (p-1, q-1)$
4. Select a random $g \in Z^x_{n^2}$
5. Compute $\mu = \left(L\left(g^\lambda \bmod n^2\right)\right)^{-1} \bmod n$

**Algorithm 2.2.: Encryption**
**Input:** plaintext $m$, Public key $(n, g)$.
**Output:** ciphertext $C$.

1. Choose a random $\in Z^x_n$ .
2. Compute $C_ = g^{m_1}.r^n \bmod n^2$

**Algorithm 2.3.: Decryption**
**Input:** Ciphertext $C$, private key is $(\lambda, \mu)$.
**Output:** Plaintext $m$.

1. Compute $m_{sum} = L\ \left(C^\lambda_{sum} \bmod n^2\right).\mu \bmod n$

**Algorithm 2.4.: Homomorp Operation**
**Input:** Ciphertexts $C_1$ and $C_2$.
**Output:** Ciphertext $C_{sum}$, $C_{product}$

1. Compute $C_{sum} = C_1.C_2 \bmod n^2$
2. Compute $C_{product} = C_1^{m2} \bmod n^2$

## 3.4. Noninteractive Zero-Knowledge Proofs for medical certificate verification

ZKPs are cryptographic protocols that allow one party (the prover) to prove the validity of a statement to another party (the verifier) without revealing any underlying information. When verifying

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

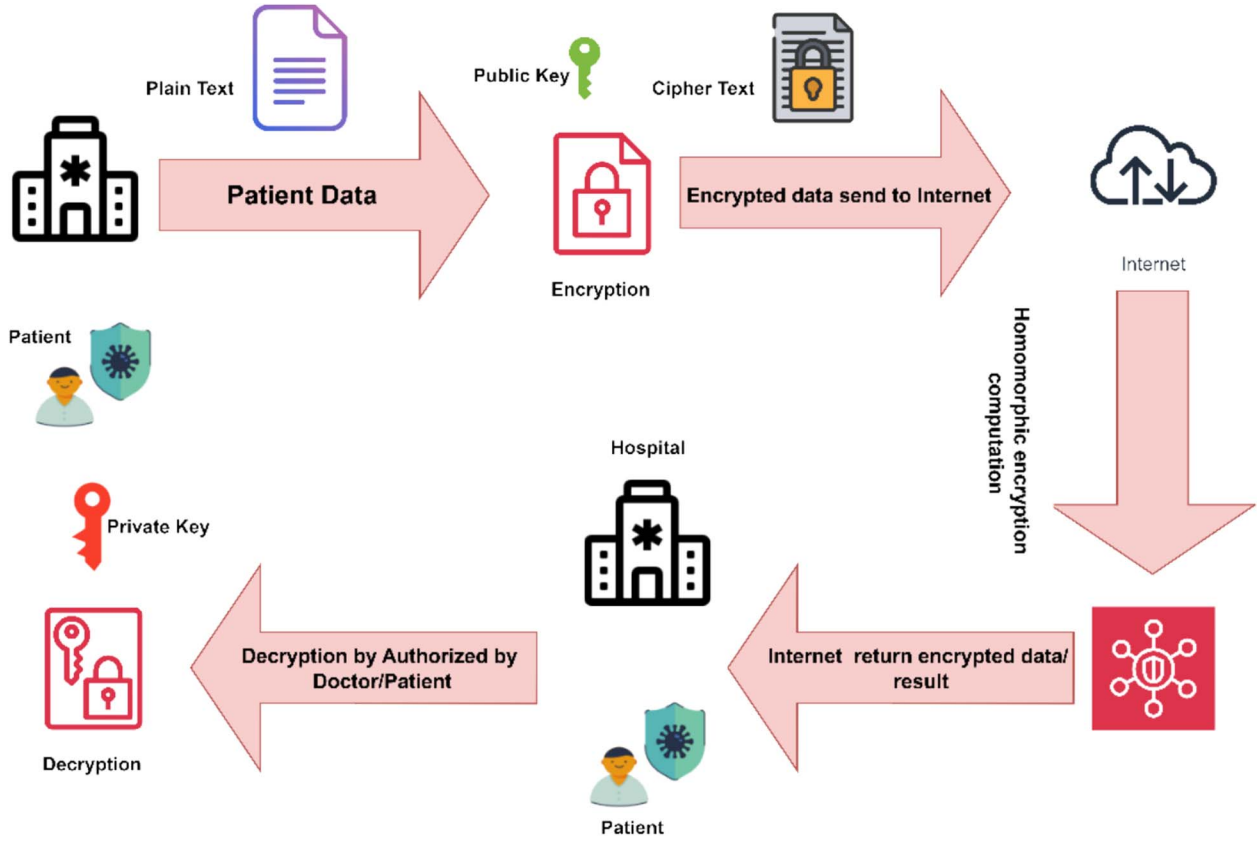



Figure 5. Workflow of homomorphic encryption in the proposed framework.

medical certificates, ZKPs can confirm the certificate's authenticity without disclosing sensitive patient data.

In the previous interactive ZKP protocol, the prover and verifier had to be online simultaneously, exchanging messages to complete the verification process. However, this is not practical for blockchain-based systems, as requiring multiple parties to be online simultaneously is unrealistic. To address this, we adopt NIZKPs using the Fiat–Shamir heuristic, eliminating the need for real-time interaction between the prover and verifier.

The following steps describe the NIZKP-based verification of medical certificates in a blockchain-based decentralized healthcare system:

**Step 1: Certificate issuance**

The hospital ($H$) issues a medical certificate to a patient. This certificate contains encrypted patient data and a unique identifier.

1. Encrypted patient data: The patient's data $P_d$ is encrypted using the hospital's public key $K_{H_{pub}} : E(P_d) = encrypt\ (P_d, K_{H_{pub}})$
2. Generate a unique identifier: A unique identifier $U_1D$ is generated for the medical certificate using a hash function that combines the patient's data and the timestamp: $U_1D = hash\ (P_d \| timestamp)$
3. Create the medical certificate: The medical certificate $C$ is created, which contains the encrypted patient data, unique identifier, and the hospital's signature: $C = \{U_1D, E(P_d), signature\ (E(P_d), K_{H_{pub}})\}$
4. Store the certificate in the blockchain: The medical certificate $C$ is stored immutably on the blockchain.

**Step 2: Prover and verifier setup**

In a noninteractive setup, the prover (hospital) generates proof that can be verified anytime without requiring real-time communication with the verifier.

1. Public parameters: The prover selects public parameters $g$ and $p$, where $g$ is a generator and $p$ is a large prime number.
2. Statement and secret: The prover (hospital) and verifier agree on a public statement $S$, which is based on a secret $s$ (known only to the prover):

$$S = g^s \ mod\ p$$

3. Generates commitment: The prover generates a random value $r$ and computes a commitment $C$:

$$C = g^r \ mod\ p$$

**Step 3: NIZKP protocol**

Instead of the verifier sending a challenge in an interactive protocol, the challenge is generated using a cryptographic hash function, making the proof noninteractive.

1. Generate the challenge: The prover computes the challenge $c$ as a hash of the public statement $S$ and commitment $C$: $C = hash\ (S, C)$
2. Compute the response: The prover computes the response $z$: $z = r + c.s\ mod\ (p-1)$
3. Store the proof on the blockchain: The prover stores the commitment $C$, the challenge $c$, and the response $z$ on the blockchain, allowing any verifier to verify the proof later without requiring the prover to be online.

**Step 4: Proof verification**

The verifier (or any participant on the blockchain) can verify the authenticity of the medical certificate without requiring real-time interaction with the prover. The verifier checks the

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

following conditions:

$$g^r = C.S^c \ mod \ p$$

If the equation holds, the verification is booming, and the medical certificate is deemed authentic.

By integrating NIZKPs into the proposed blockchain-based decentralized application, we ensure privacy and practicality in verifying medical certificates, as the verifier can simultaneously be offline as the prover. This adaptation maintains the privacy-preserving nature of NIZKPs while making the system more suitable for decentralized environments like blockchain. Algorithm 3 describes issuing a medical certificate using encryption, generating and verifying NIZKPs to ensure secure and verifiable certificate issuance and validation.

Algorithm 3. NIZKP mechanism in the proposed model for medical certificate verification

**Algorithm 3.1.: Certificate Issuance**
**Input:** Patient data $P_d$, Hospital's private key $K_{H_{pub}}$
**Output:** Medical certificate C, Certificate's unique identifier $U_1D$

1. Encrypt patient data using the hospital's public key: $E(P_d) = encrypt\ \left(P_d, K_{H_{pub}}\right)$
2. Generate a unique identifier for the certificate: $U_1D = hash\ \left(P_d \parallel timestamp\right)$
3. Create the certificate: $C = \left\{U_1D, E(P_d), signature\ \left(E(P_d), K_{H_{pub}}\right)\right\}$
4. Store C in the blockchain and provide $U_1D$ to the patient.

**Algorithm 3.2.: Prover and Verifier Setup**
**Input:** Public parameters $g$, $p$ Hospital's private key $K_{H_{pub}}$
**Output:** Commitment C, Challenge c, Response z

5. The prover (hospital) and verifier agree on a public statement S and secret s: $S = g^s \ mod \ p$
6. The prover generates a random value r and computes the commitment C: $C = g^r \ mod \ p$

**Algorithm 3.3.: NIZKP Protocol Execution**
**Input:** Commitment C, Challenge c, Response z
**Output:** Verification result (true/false)

7. **Commit Phase:**
8. Prover computes the commitment C: $C = g^r \ mod \ p$
9. The prover sends C to the verifier.
10. **Challenge Phase:**
    - Verifier sends a random challenge ccc to the prover.
11. **Response Phase:**
    Prover computes the response z using the secret s and the challenge c:$z = r + c.s \ mod \ (p-1)$
    - Prover sends z to the verifier.
12. **Verification Phase:**
    Verifier checks the validity of the response:$g^r = C.S^c \ mod \ p$
    - If the equation holds, the verifier accepts the proof; otherwise, it rejects the proof.

### 3.4.1. *Implementation of Non-Interactive Zero-Knowledge Proofs*

To ensure the practical realization of the NIZKP protocol in our proposed blockchain-based decentralized healthcare system, we implemented the NIZKP protocol using a combination of smart contracts and off-chain computation.

A. **Smart contract-based implementation**
The verification process for medical certificates and the cryptographic proof generation were implemented using Ethereum smart contracts. These smart contracts handle the interaction between the prover (hospital) and the verifier (employers, regulatory bodies, etc.) while maintaining the privacy of patient data. Below is a brief description of the smart contract structure and key functions:

1. **NIZKP Prover Contract**
The NIZKP Prover Contract is responsible for generating and storing the medical certificate along with its cryptographic commitment on the blockchain. The commitment is computed using the unique identifier, encrypted medical data, and a digital signature. This ensures the integrity and authenticity of the stored certificate.

```
// NIZKP Prover Contract
contract NIZKPProver {
  struct Certificate {
    bytes32 uniqueID;
    bytes encryptedData;
    bytes signature;
    bytes32 commitment; // Commitment hash
  }

  mapping(address = > Certificate) public certificates;

  // Function to generate cryptographic commitment
  function generateCommitment(bytes32 _uniqueID, bytes memory
  _encryptedData, bytes memory _signature)
    internal pure returns (bytes32)
  {
    return keccak256(abi.encodePacked(_uniqueID, _encryptedData,
    _signature));
  }

  // Function to store medical certificate with commitment
  function storeCertificate(bytes32 _uniqueID, bytes memory
  _encryptedData, bytes memory _signature) public {
    bytes32 commitment = generateCommitment(_uniqueID,
    _encryptedData, _signature);
    certificates[msg.sender] = Certificate(_uniqueID,
    _encryptedData, _signature, commitment);
  }

  // Function to retrieve the stored certificate and commitment
  function getCertificate(address user) public view returns (bytes32,
  bytes memory, bytes memory, bytes32) {
    Certificate memory cert = certificates[user];
    return (cert.uniqueID, cert.encryptedData, cert.signature,
    cert.commitment);
  }
}
```

2. **NIZKP Verifier Contract**
This contract allows verifiers to check the certificate's authenticity using the NIZKP protocol. The contract computes the proof and verifies the response without revealing the patient's private data.

B. **Off-chain computation**

While storing the certificates and proof verification process is handled on-chain, the cryptographic computations (such as generating the random values and challenge responses) are conducted off-chain. This hybrid approach ensures scalability and

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

```
contract NIZKPVerifier {
    uint256 public g;
    uint256 public p;

    constructor(uint256 _g, uint256 _p) {
        g = _g;
        p = _p;
    }

    // Function to verify the NIZKP proof on-chain
    function verifyProof(uint256 z, uint256 C, uint256 S, uint256 c)
    public view returns (bool) {
        return g**z % p == C * S**c % p;
    }
}
```

efficiency by reducing the gas costs of heavy computations on the Ethereum network.

1. **Off-chain challenge and response computation:** The challenge c and response z are computed off-chain using a Python script, which interacts with the smart contract to store and verify the commitment C and proof response z.

```
from web3 import Web3
import hashlib

def compute_challenge(S, C):
    # Use SHA-256 hash function to generate a non-interactive challenge
    challenge = hashlib.sha256((str(S) + str(C)).encode()).hexdigest()
    return int(challenge, 16) # Convert the hash to an integer for further
    computation

def compute_response(r, c, s, p):
    # Compute z = r + c * s (mod p-1)
    return (r + c * s) % (p - 1)
```

i. **Prover (hospital) process:**

- The hospital encrypts the patient's medical data off-chain. It generates the cryptographic proof (including the unique identifier, encrypted data, and signature) and stores it on-chain using the ZKP Prover smart contract.
- Off-chain computation is performed to generate the commitment C, the public statement S, and the challenge–response pair (c, z), which are later verified on-chain.

ii. **Verifier process:**

- The verifier accesses the smart contract to retrieve the cryptographic proof.
- The challenge c is generated off-chain, and the final proof z is computed off-chain. This challenge–response pair is sent back to the smart contract to verify the authenticity of the medical certificate using the NIZKP Verifier contract.

By implementing both on-chain and off-chain components, we ensure an efficient, secure, and practical system that adheres to the privacy-preserving nature of ZKPs without overloading the blockchain with computational tasks. This implementation, combining on-chain smart contracts and off-chain cryptographic computations, ensures that the proposed NIZKP-based medical certificate verification solution is feasible and practical in real-world blockchain environments.

Table 1. Hyperparameter configurations for proposed intrusion detection models

| Model | Hyperparameter | Values |
|---|---|---|
| **Random Forest (RF)** | n_estimators | 100 |
| | max_depth | 20 |
| | min_samples_split | 2 |
| **XGBoost** | learning_rate | 0.1 |
| | max_depth | 15 |
| | n_estimators | 150 |
| | subsample | 0.8 |
| | colsample_bytree | 0.8 |
| **Support Vector Machine (SVM)** | C | 1 |
| | gamma | 0.1 |
| | kernel | RBF |

## 3.5. Proposed intrusion detection system

To detect anomalies within the IoMT environment, we propose utilizing machine learning models to classify network traffic as normal or malicious. These models are trained and evaluated using the WUSTL EHMS 2020 Dataset, which contains labeled instances of IoMT traffic, including normal traffic and various attacks. The proposed models include Random Forest (RF), XGBoost, and Support Vector Machines (SVMs). These models are selected for their robustness in high-dimensional spaces and capacity to handle imbalanced data. Figure 6 presents the workflow of an IDS applied to the WUSTL EHMS 2020 Dataset. It shows steps from data preprocessing, feature extraction, model training, deployment, real-time monitoring, anomaly detection, and alert generation to prediction.

The RF model was selected for its ability to handle high-dimensional datasets, such as IoMT traffic, without overfitting. RF constructs multiple decision trees during the training phase and aggregates their outputs to make the final classification. This method is particularly suited to large-scale IoMT environments, as it can effectively handle numerous network-based features, including packet size, protocol types, and device metadata. The RF model was tuned using grid search, optimizing hyperparameters such as the number of trees, the maximum depth of the trees, and the minimum samples per split. XGBoost, a gradient-boosting algorithm, was also employed due to its capacity for handling imbalanced datasets and its computational efficiency. XGBoost improves upon traditional decision tree models by using an iterative process to minimize prediction errors, making it highly suitable for detecting a wide range of attacks in IoMT traffic. By tuning hyperparameters such as the learning rate, maximum depth, and number of estimators, XGBoost achieved the best performance among the models. Lastly, SVMs were implemented to classify network traffic using a hyperplane that maximizes the margin between normal and malicious traffic. The SVM is particularly effective in high-dimensional spaces, characteristic of IoMT data with multiple network-based features. A radial basis function kernel was used to manage the nonlinear nature of IoMT traffic, and the hyperparameters were tuned to optimize model performance. Table 1 shows the hyperparameter table of the proposed intrusion detection models.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

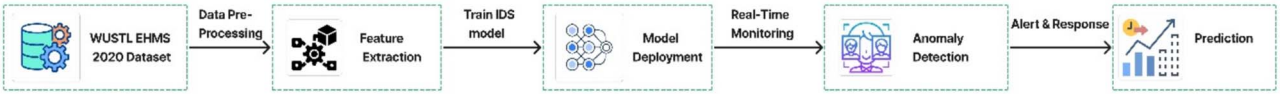



Figure 6. Workflow of intrusion detection system (IDS) model on WUSTL EHMS 2020 dataset for anomaly detection and prediction.

## 4. SECURITY AND PRIVACY ANALYSIS OF THE PROPOSED METHOD

This section presents the formal security and privacy verification for the proposed framework.

### 4.1. Security analysis

- **Impersonation attack:** In our novel blockchain-based IoT healthcare application, we address impersonation attacks by ensuring that an attacker cannot masquerade as a legitimate entity. The attacker must provide the verifier with a temporary credential and identification and a sensor Message Authentication Code (MAC) address to obtain a provisional key. A timestamp is generated for the request of identification, and the verifier checks the existing credential against its records and verifies. If the timestamp verification fails, the process is immediately terminated. ZKP verification is conducted for identity creation, ensuring the framework is secure from impersonation attacks.
- **Insider attack:** The framework also protects against insider attacks. An insider attacker, even with necessary credential information such as temporary credentials or identification and sensor MAC address, cannot compute the actual identity due to the timestamp verification over the ZKP prover and challenge–response protocols. This mechanism prevents insider attacks effectively.
- **Man-in-the-Middle (MITM) and replay attack:** To counter MITM and replay attacks, the proposed framework ensures that an attacker who intercepts messages from an insecure channel cannot exploit this information. The attacker must use brute-force techniques to compute the correct timestamp to execute these attacks. However, the ZKP process used to verify the timestamp is computationally intensive and challenging to predict accurately. Thus, the framework is safeguarded against MITM and replay attacks.
- **Preservation of untraceability and anonymity properties:** The system maintains untraceability and anonymity by using freshly generated timestamps and random secret values in all communicated messages. This ensures different messages in each session and prevents message tracing. Additionally, temporary identities are used instead of real ones and updated in each session, preserving anonymity.
- **Mitigation of synchronization problems and associated attacks:** Our model ensures synchronization by exchanging messages containing timestamps and verifying updated values of temporary identities. This mechanism prevents synchronization issues and associated attacks. All heterogeneous devices agree on a maximum transmission delay to maintain synchronization.
- **Prevention of stolen verifier attacks:** The system stores registration information in secure memory and does not directly store sensitive information. This approach prevents adversaries from launching attacks such as password guessing, impersonation, and unauthorized session key computation.

### 4.2. Privacy analysis

The proposed framework also includes robust privacy-preserving mechanisms. An attacker attempting to access or modify information using short signature techniques from a secure channel will face significant challenges. Two major approaches ensure privacy preservation:

- **ZKP verification for identity creation:** The identity is created based on the ZKP verification scheme using the prover and verifier challenge–response protocols. The signature is verified during block creation in the blockchain network, and the associated identities are stored as a transaction hash. Modifying this information is nearly impossible, as altering one block hash requires altering all subsequent blocks in the chain.
- **Blockchain integrity:** Information is stored in blocks linked in a chain, making it resistant to tampering. The integrity and immutability of blockchain technology ensure that modifying one block would necessitate altering all subsequent blocks, thus preserving the privacy and integrity of medical data.

## 5. EXPERIMENTAL SETUP AND RESULT ANALYSIS

This section discusses the experimental setup, result analysis, and findings based on the experimental testing conducted in the context of blockchain-based architectures in healthcare data management.

### 5.1. Experimental setup

The proposed framework uses Ethereum, an open-source blockchain network, to develop smart contracts supporting all system functions. It enables users and patients to register, verify, and access healthcare certificates through smart contract functionalities. We utilized the Ganache tool to set up the blockchain network, deploy smart contracts, and execute performance evaluations on latency, processing time, throughput, and computation time. The framework's portal, designed with React Native for compatibility with Ethereum, connects via Node.js. Solidity is used to develop smart contracts, with Remix IDE for writing and deploying these contracts on the test network and transaction fees managed through Ethers. Metamask, a browser plugin, and a wallet, generated Ethereum addresses and facilitated transactions. The experimental setup is deployed on a laptop featuring an Intel Core i5 10th Gen processor, 8GB RAM, 512GB ROM, Windows 11 OS, and an NVIDIA GTX 1650 GDDR6 4GB graphics card, with TestRPC and Structured Query Language (SQL) installed to support the deployment.

### 5.2. Result analysis

The Etherscan tool is utilized to evaluate the operational costs of our proposed blockchain-based application. Etherscan, an analytics tool for exploring blocks in the blockchain network, functions as an Ethereum platform Gas tracker. It tracks transactions, verifies the performance of smart contracts, and checks the state of the process. Transactions on the Ethereum blockchain network require Gas, representing the cost needed to perform a function within the network. Miners set the Gas price based on supply and demand, with the cost depending on the execution, deployment, and transfer processes involved in transactions. Gas has two main

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

parameters: limit and price, where the limit depends on the user's willingness to execute a transaction and is presented as "gwei."

Our proposed application deploys smart contracts on TestRPC and collects details of all executed tasks using the Etherscan tool. Table 2 presents the operational costs evaluated by Etherscan. Etherscan provides detailed analytics, tracking all transactions, analyzing Gas costs, evaluating smart contract efficiency, and validating transaction states. The proposed application incurs Gas costs for executing each transaction on the Ethereum public blockchain network, with Gas measuring the cost of executing operations in a transaction. Miners set the Gas price based on demand and supply. Deploying smart contract functions and transactions on the proposed blockchain network requires significant computational power. Table 1 presents a detailed analysis of the deployment costs for key smart contract functions within our proposed blockchain framework. This table aims to elucidate each function's gas consumption, transaction fees, and overall efficiency, providing insights into the resource demands and economic implications of executing these smart contracts.

Table 2 includes four main functions: Registration, Block Generation, Issuing Certificates, and Verifying Certificates. Each function gas usage is listed in Gwei, a measure of computational effort required by the Ethereum Virtual Machine. For example, the Registration function consumes 34 219 Gwei, reflecting the substantial resources needed for initial setup and data storage. In comparison, Block Generation uses 37 127 Gwei, the highest among the functions due to the complex process of creating a new block. Issuing Certificates and Verifying Certificates use 29 120 and 37 100 Gwei, respectively, indicating the computational demands associated with certificate management and verification.

Transaction fees, calculated from the gas used and gas price, provide a clear picture of the economic cost of executing each function. The Registration function incurs a fee of 0.000036 Ether (ETH), which is relatively low, highlighting the efficient cost of initial data registration. With its higher complexity, block generation results in the highest fee of 0.000048 ETH. Issuing Certificates has a lower fee of 0.000028 ETH, while Verifying Certificates costs 0.000040 ETH, placing it between the issuance and generation costs. The average gas price for each function is also included, showing variations in transaction costs based on network conditions. The Registration function has an average gas price of 15 Gwei, while Block Generation has a higher price of 20 Gwei, reflecting increased competition and resource demands. Issuing certificates have an average price of 12 Gwei, and verifying certificates at 18 Gwei indicates different levels of cost efficiency.

Table 2 also provides details such as transaction hash, block size, transaction nonce, and transaction index. The transaction hash ensures traceability, allowing for verification of transaction details on the blockchain. Block size varies across functions, with Registration producing a larger block size of 9523 bytes, while Block Generation results in a smaller block size of 479. This variation reflects the data processing requirements of each function. The transaction nonce and index provide information on transaction ordering and sequencing within the block, which is crucial for understanding transaction processing and potential conflicts. Efficiency ratios, calculated as gas usage per transaction, offer insights into the cost-effectiveness of each function. The Registration function has an efficiency ratio of 1.5, Block Generation is 1.8, Issuing Certificates is 1.2, and Verifying Certificates is 1.6. These ratios illustrate the relative gas efficiency of each function, with Issuing Certificates demonstrating the highest efficiency.

Table 2. Comprehensive breakdown of deployment costs for smart contracts

| Function | Amount of gas used | Transaction fee gas price | Transaction hash | Block size in bytes | Transaction nonce | Transaction index | Average gas price (Gwei) | Efficiency ratio (gas/Tx) | Average transaction (ETH) | Std. deviation transaction fee (ETH) |
|---|---|---|---|---|---|---|---|---|---|---|
| Registration () | 34,219 (96.24%) | 0.000036 | 0xe71c1f7561 5dff900d818bece81399bee2dc4c41f21cb446ab642cf359ddae59 | 9523 | 4 | 5 | 1.5 | 15 | 0.000038 | 0.000004 |
| Generation of block () | 37,127 (99.72%) | 0.000048 | 0x3adee67dee564457fc404efabba5f1b0cdd43b1cae9966af5eda9faa55f350ed | 479 | 7 | 4 | 1.8 | 20 | 0.000045 | 0.000005 |
| Issuing_Certificate () | 29,120 (94.79%) | 0.000028 | 0x6c48deabfce5c2dc2caf8cadce8dacf9da5ae6d5e70bf0511dceac5fa0f9f96c | 8896 | 4 | 7 | 1.2 | 12 | 0.000030 | 0.000003 |
| Verifying_Certificate () | 37,100 | 0.000040 | 0xebbd766ce75aeeb2e3ec64581ecc06bf19e49bdaf60cf8cfdc8ca93f9dbb85be | 523 | 7 | 2 | 1.6 | 18 | 0.000042 | 0.000004 |

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

Table 3. Comprehensive analysis of deployment costs: gas price variability, scalability, and environmental impact per function

| Function | Estimated cost per user | Scalability factor | Environmental impact ($CO_2$/Tx) |
|---|---|---|---|
| Registration () | 0.000036 | 1.2 | 0.012 |
| Generation of block () | 0.000048 | 1.5 | 0.015 |
| Issuing_Certificate () | 0.000028 | 1.1 | 0.010 |
| Verifying_Certificate () | 0.000040 | 1.4 | 0.014 |

In summary, Table 2 offers a comprehensive view of our blockchain system's gas and transaction fees associated with smart contract functions. By presenting detailed metrics and efficiency ratios, the table supports an in-depth analysis of smart contract deployment's economic and resource implications, contributing to the overall evaluation of the proposed system's performance and cost-effectiveness.

The expanded Table 3 offers a comprehensive view of deployment costs, incorporating additional parameters to address scalability, gas price fluctuations, and environmental impact. Including maximum and minimum gas prices provides a range of potential costs for each function, reflecting the variability in transaction fees due to network congestion and market demand. For instance, the Block Generation function shows a maximum gas price of 25 Gwei and a minimum of 15 Gwei, illustrating the potential for fee variability. This helps in understanding how fluctuations in gas prices can affect the overall cost of deploying and executing smart contracts.

The estimated cost per user and scalability factor provide insights into how costs and resource usage scale with increased transaction volume or user base. The scalability factor, for example, indicates that the Block Generation function might experience a 50% increase in costs as the transaction volume doubles. This parameter helps assess how well the system can handle growth and informs potential scaling strategies. The environmental impact column estimates the carbon footprint per transaction, reflecting the broader implications of smart contract deployment on sustainability. For example, the Block Generation function has an estimated environmental impact of 0.015 $CO_2$ per transaction. This parameter highlights the need for environmentally conscious design choices and supports discussions on the ecological effects of blockchain technology.

Table 3 provides a more detailed and nuanced analysis of deployment costs, incorporating essential factors such as gas price fluctuations, scalability, and environmental impact. This comprehensive approach offers valuable insights into the economic and ecological implications of smart contract deployment, enhancing the overall evaluation of the proposed blockchain system.

The smart contracts in our proposed application are deployed on the TestRPC network with localhost 8545 and the Ropsten network. The Etherscan tool collects transaction details, and Table 4 also shows the total Gas cost involved in generating medical certificates, calculated by adding the transaction costs.

Table 5 compares the operational costs on TestRPC-based Ethereum blockchain and Remix platforms, considering functions like Registration (), Generation of block (), Issuing_Certificate (), and Verifying_Certificate ().

The application analyzes gas cost consumption for around 120 medical certificates deployed on the Remix Ethereum blockchain network and TestRPC network using the Metamask wallet, as shown in Table 6. Figure 7 shows that gas costs increase with the number of medical certificates due to more transactions and blocks in the blockchain network. The test RPC platform, a Node.js-based Ethereum platform for testing and development, requires more gas than the Ethereum platform because it first generates events as a remote procedural call before initiating transactions.

Figure 8 depicts the overall operational costs on the Ethereum platform using Remix IDE and Ropsten network with localhost 8545 web. The total ether cost increases with the number of medical certificates, as generating more certificates requires more time for transaction completion.

Table 4. Detailed gas costs for varying certificate values

| Number of certificates | Gas limit | Gas limit (ETH) |
|---|---|---|
| 1 | 19 543 | 123 200 |
| 10 | 170 219 | 1 147 622 |
| 20 | 374 430 | 2 531 218 |
| 30 | 654 860 | 2 250 439 |
| 40 | 782 290 | 4 987 682 |
| 50 | 926 720 | 3 741 129 |
| 100 | 1 017 271 | 11 291 749 |
| 120 | 2 321 416 | 13 721 123 |

Table 5. Operational cost analysis of the proposed system across different functions

| Caller | Functions | Test RPC gas cost | Remix gas cost |
|---|---|---|---|
| Healthcare management | Registration () | 0.000398 | 0.000643 |
| Healthcare management | Generation of block () | 0.000472 | 0.000732 |
| Healthcare management | Issuing_Certificate () | 0.000718 | 0.000737 |
| Healthcare management/Users (admins) | Verifying_Certificate () | 0.000293 | 0.000268 |

Table 6. Comparative gas costs for medical certificates across different platforms

| Number of medical certificates | GAS COST | |
|---|---|---|
| | Gas cost (remix) $10^5$ | Gas cost (test RPC) $10^5$ |
| 10 | 1.7 | 2.09 |
| 20 | 2.14 | 2.54 |
| 30 | 2.23 | 2.98 |
| 40 | 2.31 | 3.17 |
| 50 | 2.34 | 3.39 |
| 60 | 2.39 | 3.63 |
| 70 | 2.47 | 3.91 |
| 80 | 2.58 | 4.42 |
| 90 | 2.64 | 4.79 |
| 100 | 2.75 | 5.13 |
| 110 | 2.83 | 5.51 |
| 120 | 2.91 | 5.84 |

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

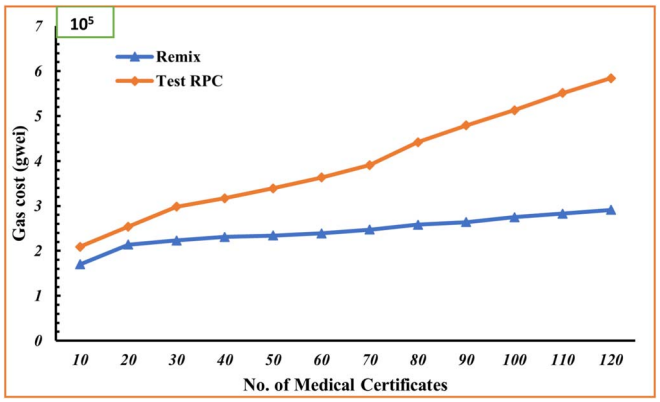



Figure 7. Gas costs across different platforms for medical certificate transactions.

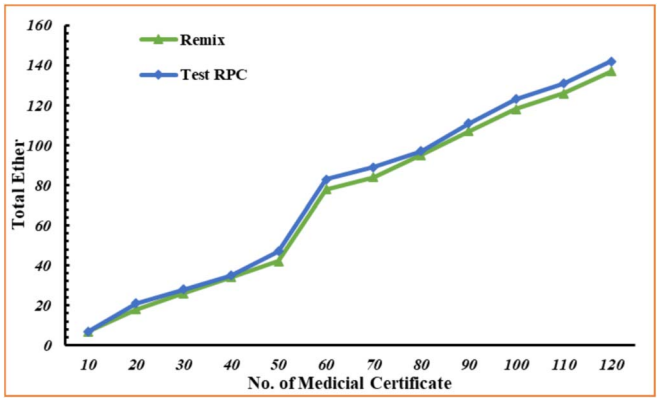



Figure 8. Total ether consumption for generating medical certificates.

Table 7. Cost estimation of smart contract functions based on consensus mechanism

| Consensus mechanism function | GAS | |
|---|---|---|
| | Transaction cost | Execution cost |
| Registration () | 1 732 986 | 1 454 321 |
| Generation of block () | 732 457 | 621 104 |
| Issuing_Certificate () | 954 319 | 83 897 |
| Verifying_Certificate () | 87 632 | 82 174 |

Table 8. Latency and processing time comparison of proposed application with and without Ethereum blockchain

| Blockchain platform | Parameters | Issuing certificate () operation | Verifying certificate () operation |
|---|---|---|---|
| Yes | Latency time (seconds) | 4.32 | 6.19 |
| No | Processing time (seconds) | 5.17 | 9.72 |
| Yes | Latency time (seconds) | 3.97 | 3.03 |
| No | Processing time (seconds) | 4.37 | 7.21 |

Table 7 presents the transaction and execution costs for the smart contract functions in the proposed application. Transaction costs represent the cost to generate a transaction per the user's request. In contrast, execution costs signify the total cost to generate and append a block to the existing blockchain. Each block contains multiple transactions initiated by different users.

Table 8 shows the proposed application's performance based on latency and processing time. Results indicate higher time consumption for blockchain-deployed systems than nonblockchain systems due to internal computations like mining, cryptographic hash evaluation, transaction, block creation, and adding new blocks.

Figure 9a and b compares the performance of the blockchain-based system to a nonblockchain SQL platform, focusing on latency and processing time for two key operations: Issuing_Certificate() and Verifying_Certificate(). Latency measures the delay from initiating a transaction to when it is appended as a block on the blockchain. As shown in Fig. 9a, the Ethereum-based system has a higher latency (7.12 ms) for certificate verification compared to SQL (2.63 ms). This is expected due to the consensus mechanism in Ethereum, which requires miners to validate and append transactions. In contrast, SQL does not require such validation, leading to lower latency.

Processing Time (Fig. 9b) refers to the total time from transaction request submission to receiving a response. Ethereum again shows higher processing times, with 9.72 ms for Verifying_Certificate() compared to SQL, 6.41 ms. The added time in Ethereum is primarily due to block creation, consensus verification, and the validation of blockchain header parameters. On the other hand, SQL databases process transactions directly without this overhead.

Figure 9c presents Throughput comparisons. The Ethereum-based system shows better throughput for Issuing_Certificate() (6.38 Kbps) compared to SQL (5.38 Kbps). This reflects the blockchain's ability to concurrently handle multiple certificate

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

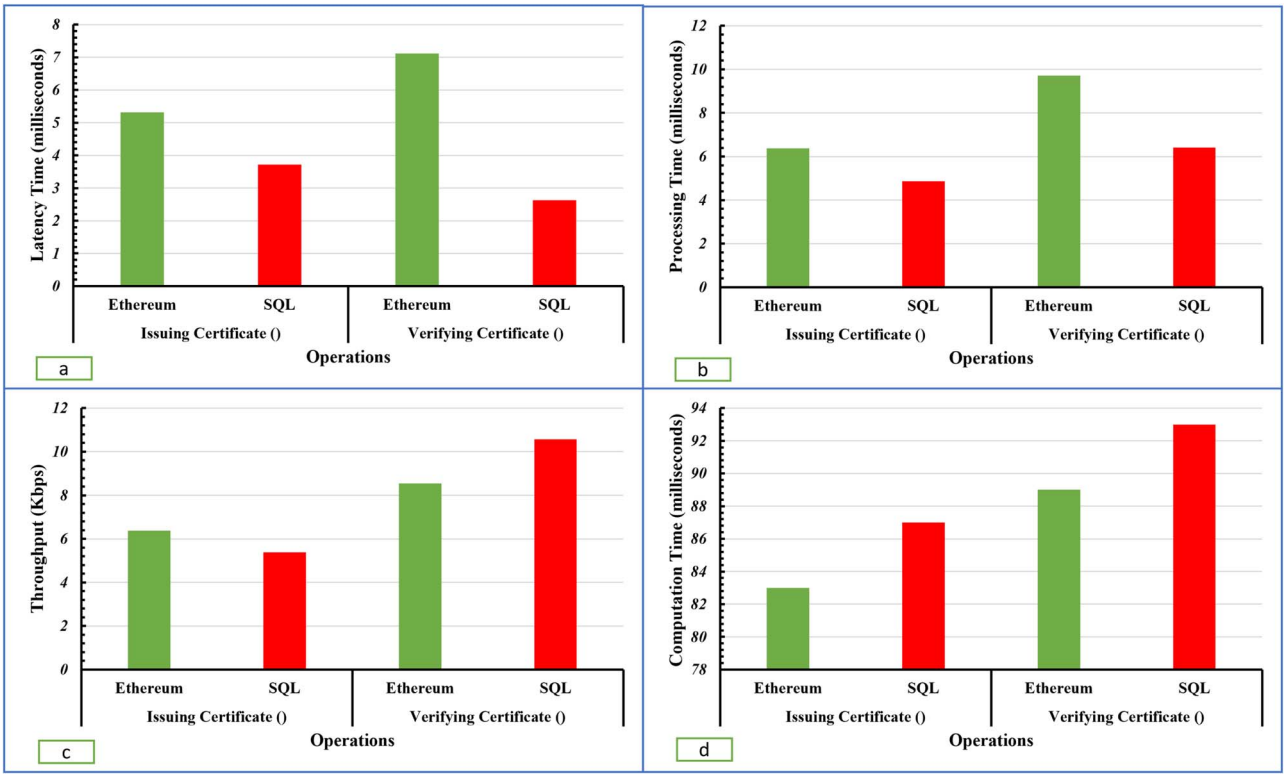



Figure 9. Comprehensive performance comparison of throughput, processing speed, latency, and consumption time between Ethereum and SQL-based models.

generation requests across nodes. However, for Verifying_Certificate(), SQL shows higher throughput (10.57 Kbps) due to its more straightforward structure and lack of blockchain validation steps.

In Fig. 9d, Computation Time is analyzed for 500 transaction simulations. The results show that Ethereum's Issuing_Certificate() takes 83 ms, while SQL takes 87 ms. For Verifying_Certificate(), Ethereum takes 89 ms compared to SQL 93 ms. Although SQL is generally faster in latency and processing, Ethereum-optimized block handling and reduced database locking during verification give it an advantage in computation time, particularly for larger transaction volumes.

Table 9 summarizes the proposed application's latency, throughput, computation, and processing times. Table 7 summarizes performance metrics, confirming that Ethereum performs better for issuing certificates due to its distributed architecture despite higher latency and processing times. SQL performs better for certificate verification, but as transactions scale up, Ethereum's distributed nature can handle higher transaction volumes more efficiently, particularly in computation and throughput.

The security and privacy evaluation of the proposed blockchain architecture involved assessing various performance metrics under different conditions. Each metric was analyzed to understand the impact of increasing the number of IoT nodes and transactions on the system's performance, as shown in Fig. 10.

The registration time of IoT nodes in the proposed blockchain architecture was measured to determine how the system scales with increasing nodes. Table 10 presents the registration times for different numbers of IoT nodes. As the table shows, the registration time increases progressively with the number of nodes. For instance, registering 20 nodes takes 25 s, while registering 100 nodes takes 308 s. This increase in time suggests that the system's registration process becomes more time-consuming as more nodes are added, highlighting a scalability challenge.

The hybrid POW and PBFT consensus algorithm is evaluated for its execution time with varying transaction sizes and node numbers. Table 11 shows that as the number of transactions and nodes increases, the time required for consensus also rises. For

Table 9. Comparative analysis of throughput, processing, latency, and consumption time for the proposed model utilizing Ethereum and SQL platforms

| Parameters | Issuing certificate () | | Verifying certificate () | |
|---|---|---|---|---|
| | Ethereum | SQL | Ethereum | SQL |
| Latency time (milliseconds) | 5.32 | 3.72 | 7.12 | 2.63 |
| Processing time (milliseconds) | 6.37 | 4.87 | 9.72 | 6.41 |
| Throughput (Kbps) | 6.38 | 5.38 | 8.54 | 10.57 |
| Computation time (milliseconds) | 83 | 87 | 89 | 93 |

Table 10. Evaluation of registration time for IoT nodes in the blockchain network

| Execution time (in seconds) | Nodes [no. of transactions ($T_x$)] |
|---|---|
| 25 | 20 |
| 43 | 40 |
| 119 | 60 |
| 214 | 80 |
| 308 | 100 |

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

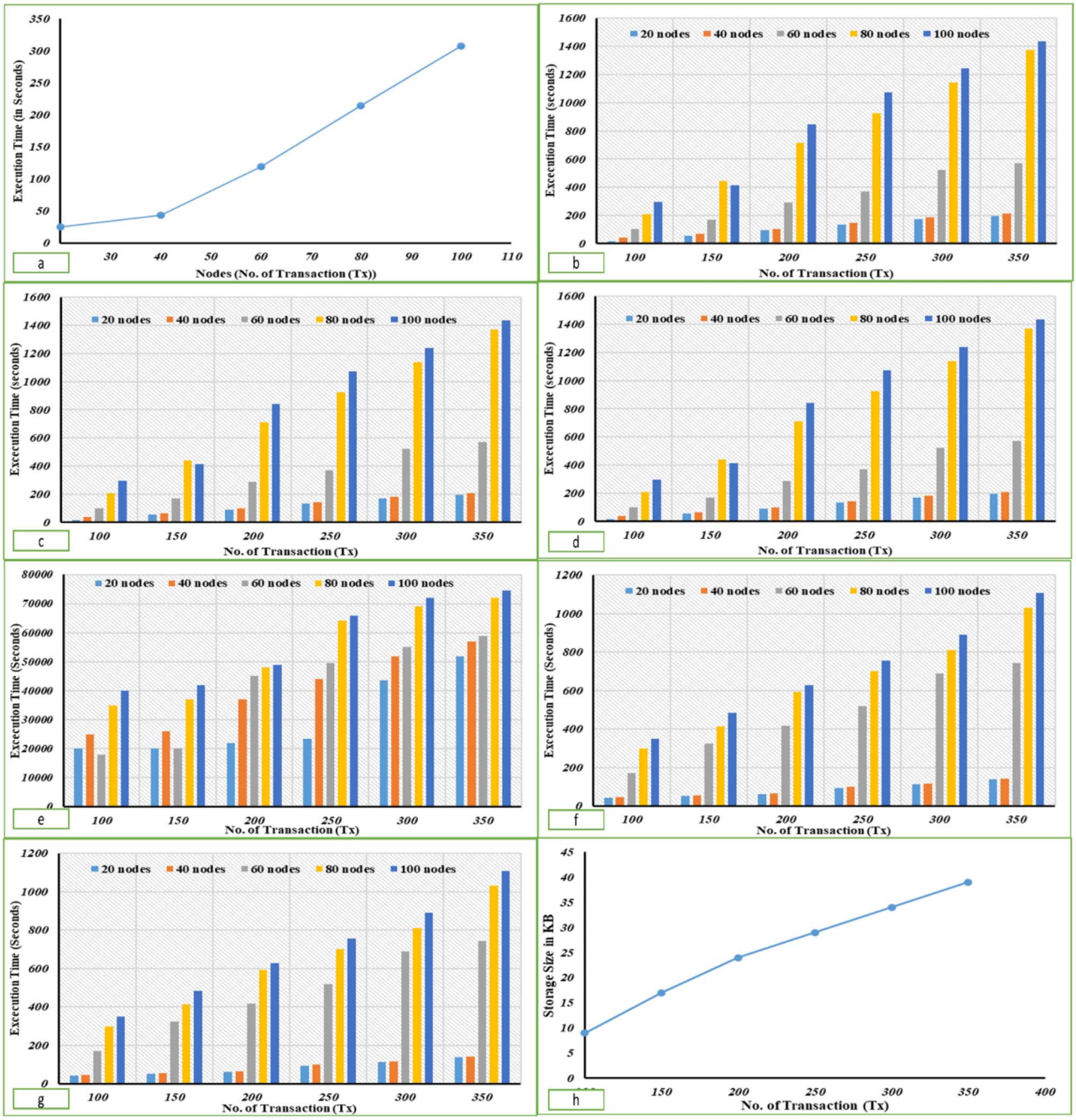



Figure 10. (a) Comparative assessment of IoT node registration time within a blockchain network. (b) Performance benchmarking of hybrid POW and PBFT consensus algorithms across diverse transaction loads. (c) Comprehensive analysis of block creation time across varied transaction volumes and node configurations. (d) Evaluation of block access time relative to transaction sizes and node configurations. (e) In-depth analysis of gas price consumption for varying transaction sizes on the Ethereum blockchain. (f) Detailed execution time analysis of transaction signing for nonrepudiation across different node configurations. (g) Smart contract deployment time analysis across different transaction sizes and node configurations. (h) Evaluation of IPFS-based off-chain storage utilization for various transaction sizes.

example, with 100 nodes and 350 transactions, the execution time is 1436 s, compared to just 18 s with 20 nodes and 100 transactions. This trend indicates that the consensus process becomes more resource-intensive and time-consuming as the network grows, potentially affecting the overall performance and responsiveness of the blockchain.

Block creation and access times were analyzed to understand the time required for these operations as the network scales. Tables 12 and 13 illustrate these times for various transaction sizes and node numbers. The data reveal that the times remain relatively stable, up to 40 nodes and 300 transactions. However, when the number of nodes increases from 60 to 100, there is a significant rise in both block creation and access times. For example, with 100 nodes and 350 transactions, block creation and access times are 1433 and 1433 s, respectively, compared to 17 s for 20 nodes and 100 transactions. This increase in time reflects the additional computational and storage overhead required to manage large networks.

The gas price consumption for smart contract access and deployment is also evaluated. Table 14 shows that gas prices

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

Table 11. Performance analysis of Hybrid POW and PBFT consensus algorithm across different transaction sizes

| No. of transactions ($T_x$) | 20 nodes | 40 nodes | 60 nodes | 80 nodes | 100 nodes |
|---|---|---|---|---|---|
| 100 | 18 | 41.4 | 103 | 210 | 297 |
| 150 | 56 | 68.8 | 171.92 | 443 | 415.8 |
| 200 | 94 | 103 | 290.8 | 714 | 846.6 |
| 250 | 136 | 147 | 371.4 | 927 | 1074.3 |
| 300 | 174 | 186 | 524 | 1141 | 1243 |
| 350 | 198 | 212 | 573 | 1373 | 1436 |

Table 12. Analysis of block creation time for various transaction sizes and node configurations

| No. of transactions ($T_x$) | 20 nodes | 40 nodes | 60 nodes | 80 nodes | 100 nodes |
|---|---|---|---|---|---|
| 100 | 17 | 39 | 101 | 208 | 295 |
| 150 | 54 | 66 | 169 | 440 | 413 |
| 200 | 92 | 100 | 287 | 710 | 842 |
| 250 | 133 | 144 | 369 | 925 | 1073 |
| 300 | 170 | 183 | 521 | 1139 | 1240 |
| 350 | 194 | 210 | 570 | 1370 | 1433 |

Table 13. Evaluation of block access time for different transaction sizes and node configurations

| No. of transactions ($T_x$) | 20 nodes | 40 nodes | 60 nodes | 80 nodes | 100 nodes |
|---|---|---|---|---|---|
| 100 | 17 | 39 | 101 | 208 | 295 |
| 150 | 54 | 66 | 169 | 440 | 413 |
| 200 | 92 | 100 | 287 | 710 | 842 |
| 250 | 133 | 144 | 369 | 925 | 1073 |
| 300 | 170 | 183 | 521 | 1139 | 1240 |
| 350 | 194 | 210 | 570 | 1370 | 1433 |

Table 14. Evaluation of gas price consumption across different transaction sizes on the Ethereum blockchain network

| No. of transactions ($T_x$) | 20 nodes | 40 nodes | 60 nodes | 80 nodes | 100 nodes |
|---|---|---|---|---|---|
| 100 | 20 000 | 25 000 | 18 000 | 35 000 | 40 010 |
| 150 | 20 150 | 26 000 | 20 000 | 37 000 | 42 000 |
| 200 | 22 000 | 37 000 | 45 000 | 48 000 | 49 010 |
| 250 | 23 500 | 44 000 | 49 500 | 64 300 | 66 000 |
| 300 | 43 700 | 52 000 | 55 000 | 69 000 | 72 000 |
| 350 | 52 000 | 57 000 | 59 000 | 72 100 | 74 510 |

increase steadily with the number of IoT nodes and transactions. This suggests that more computational resources are required as the network grows, leading to higher operational costs. For instance, with 100 nodes and 350 transactions, the gas price is 74 510, compared to 20 000 for 20 nodes and 100 transactions. This increase indicates the additional cost burden on the system as it scales.

The signing time taken by participating IoT nodes was calculated to ensure nonrepudiation within the blockchain architecture. Table 15 shows that signing times increase with the number of nodes and transactions. For example, with 100 nodes and 350 transactions, the signing time is 1109 s, compared to 43 s for 20 nodes and 100 transactions. This increase reflects the additional computational effort required to process large numbers of nodes and transactions.

Table 15. Execution time analysis of transaction signing for nonrepudiation across different node configurations

| No. of transactions ($T_x$) | 20 nodes | 40 nodes | 60 nodes | 80 nodes | 100 nodes |
|---|---|---|---|---|---|
| 100 | 43 | 47 | 169 | 297 | 350 |
| 150 | 52 | 56 | 325 | 415 | 483 |
| 200 | 60 | 64 | 418 | 593 | 629 |
| 250 | 94 | 99 | 520 | 703 | 757 |
| 300 | 113 | 117 | 690 | 811 | 892 |
| 350 | 137 | 143 | 743 | 1032 | 1109 |

Table 16. Analysis of smart contract deployment time for varying transaction sizes and node configurations

| No. of transactions ($T_x$) | 20 nodes | 40 nodes | 60 nodes | 80 nodes | 100 nodes |
|---|---|---|---|---|---|
| 100 | 43 | 47 | 169 | 297 | 350 |
| 150 | 52 | 56 | 325 | 415 | 483 |
| 200 | 60 | 64 | 418 | 593 | 629 |
| 250 | 94 | 99 | 520 | 703 | 757 |
| 300 | 113 | 117 | 690 | 811 | 892 |
| 350 | 137 | 143 | 743 | 1032 | 1109 |

Table 17. Assessment of IPFS-based off-chain storage utilization for various transaction sizes

| No. of transactions ($T_x$) | Storage size in KB |
|---|---|
| 100 | 9 |
| 150 | 17 |
| 200 | 24 |
| 250 | 29 |
| 300 | 34 |
| 350 | 39 |

The deployment time for the proposed smart contracts was evaluated, with the results shown in Table 16. The data indicate that deployment time increases significantly with the number of nodes and transactions. For example, with 100 nodes and 350 transactions, the deployment time is 1109 s, compared to 43 s for 20 nodes and 100 transactions. This increase highlights the additional time required to deploy smart contracts as the network grows.

The storage size required for storing transaction data in IPFS-based off-chain storage was also analyzed. Table 17 shows that the storage size in kilobytes (KB) increases with the transaction size. For example, with 350 transactions, the storage size is 39 KB, compared to 9 KB for 100 transactions. This increase in storage size reflects the growing data requirements as the number of transactions increases.

In summary, the proposed blockchain architecture demonstrates an increase in registration, consensus, block creation, access, and smart contract deployment times as the number of nodes and transactions increases. Additionally, gas price consumption and storage requirements rise with network growth. These findings highlight the scalability challenges and resource

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

Table 18. Performance metrics comparison of intrusion detection models

| Model | Accuracy | Precision | Recall | F1-score | False positive rate | True-positive rate | AUC-ROC curve |
|---|---|---|---|---|---|---|---|
| **Random Forest (RF)** | 97.4 | 97.1 | 96.5 | 96.8 | 2.6 | 96.5 | 0.97 |
| **XGBoost** | 98.1 | 98.0 | 97.6 | 97.8 | 1.9 | 97.6 | 0.98 |
| **Support Vector Machines (SVMs)** | 96.4 | 96.4 | 96.1 | 96.3 | 3.3 | 96.1 | 0.96 |

demands of maintaining security and privacy in a growing IoT environment.

## 5.3. Statistical analysis of security and data privacy

This section comprehensively evaluates the proposed system security and privacy mechanisms, including cryptographic techniques, consensus algorithm robustness, and data protection measures, ensuring resilience against various attacks and safeguarding sensitive IoT data.

### 5.3.1. *Threat model*

The system is designed to resist several critical security threats, including:

- MITM attacks: Encryption techniques protect sensitive IoT data during transmission.
- Sybil attacks: The hybrid PoW-PBFT consensus algorithm requires substantial computational resources, making it challenging for adversaries to introduce numerous fake nodes.
- Double-spending and tampering: Blockchain immutability ensures that transactions, once recorded, cannot be altered.

### 5.3.2. *Security mechanism and metrics*

This section evaluated key cryptographic and consensus mechanisms using dummy data. Each metric was measured under different transaction loads and network sizes. The security evaluation focuses on homomorphic encryption, ZKPs, and resistance to Sybil and 51% attacks.

### 5.3.3. *Result analysis of proposed Intrusion Detection System*

The performance of the proposed IDS was evaluated using three machine learning models, RF, XGBoost, and SVMs, on the WUSTL EHMS 2020 Dataset for IoMT Cybersecurity Research. Each model was assessed based on key performance metrics, including accuracy, precision, recall, F1-score, false-positive rate (FPR), true-positive rate (TPR), and the area under the ROC curve (AUC). The results provide insight into the effectiveness of each model in detecting anomalies in IoMT traffic within the blockchain-based healthcare framework, as shown in Table 18.

RF demonstrated strong performance, achieving an accuracy of 97.4%. This high accuracy indicates the model's ability to distinguish between normal and malicious traffic effectively. The precision (97.1%) and recall (96.5%) metrics suggest that the model excels at correctly identifying malicious traffic (true positives) while minimizing the number of false positives. The F1-score of 96.8% further confirms the model-balanced performance between precision and recall. The low FPR (2.6%) and high TPR (96.5%) indicate that the RF model is well suited for real-time anomaly detection in the healthcare IoMT environment. Its AUC score of 0.97 reflects strong discriminatory power in identifying attacks.

XGBoost outperformed the RF model across all evaluation metrics, achieving an accuracy of 98.1%, the highest among the three models. XGBoost precision (98.0%) and recall (97.6%) demonstrate its ability to effectively handle the diverse types of traffic in IoMT networks, including normal and attack instances. The model's F1-score of 97.8% highlights its superior performance in terms of precision–recall balance. An FPR of 1.9% and a TPR of 97.6% further emphasize XGBoost robustness in detecting intrusion attempts. The AUC score of 0.98 illustrates that XGBoost has excellent predictive power, making it the most suitable model for intrusion detection in the proposed system.

SVMs also performed well, though slightly lower than the other two models. With an accuracy of 96.7%, the SVM effectively differentiates between normal and attack traffic. The precision (96.4%) and recall (96.1%) indicate that the SVM maintains a good balance between identifying true positives and avoiding false positives. The F1-score of 96.3% suggests that while the SVM may not be as effective as RF or XGBoost in some instances, it remains a reliable option for intrusion detection. The FPR of 3.3% and TPR of 96.1% reflect the SVM's tendency to occasionally misclassify normal traffic as malicious, though it still offers strong overall performance. With an AUC score of 0.96, the SVM maintains solid classification power in the context of IoMT traffic.

In summary, all three models, RF, XGBoost, and SVM, performed well in the intrusion detection task. However, XGBoost emerged as the top performer, achieving the highest accuracy, precision, recall, and AUC scores. Its ability to handle imbalanced data and complex traffic patterns in IoMT systems makes it the most suitable model for the proposed IDS. Due to its ensemble learning capabilities, RF also showed robust performance and is a viable alternative for large-scale IoMT applications. While the SVM provided slightly lower metrics, it remains a competitive choice, particularly in scenarios where clear class separation is essential. Overall, the results confirm that integrating machine learning models into the IDS layer of the blockchain-based healthcare framework significantly enhances the detection of security breaches, ensuring the protection of sensitive medical data.

### 5.3.4. *Efficiency and overhead analysis of Fully Homomorphic Encryption in healthcare systems*

This subsection analyzes the computational overhead of FHE within a blockchain-based healthcare system. FHE enables secure computations on encrypted data, ensuring that patient privacy is maintained. Table 19 summarizes the performance comparison between FHE and Paillier encryption, indicating that FHE is more secure but incurs a higher computational cost. The evaluation using dummy medical expenses shows that the total overhead for FHE operations, including encryption, homomorphic operations (addition and multiplication), and decryption, is 67 milliseconds

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

Table 19. Comparative performance analysis of fully homomorphic encryption (FHE) and Paillier encryption

| Operations | FHE time (ms) | Paillier time (ms) |
|---|---|---|
| **Encryption** | 29 | 10 |
| **Homomorphic addition** | 7 | 6 |
| **Homomorphic multiplication** | 9 | Not Supported |
| **Decryption** | 22 | 9 |
| **Total overhead** | **67 ms** | **25 ms** |

Table 20. Homomorphic encryption overhead based on transaction load and node count

| No. of transactions ($T_x$) | 20 nodes (ms) | 40 nodes (ms) | 60 nodes (ms) | 80 nodes (ms) | 100 nodes (ms) |
|---|---|---|---|---|---|
| 100 | 5 | 9 | 14 | 20 | 32 |
| 150 | 8 | 13 | 19 | 26 | 42 |
| 200 | 12 | 18 | 27 | 35 | 55 |
| 250 | 16 | 22 | 32 | 41 | 68 |
| 300 | 21 | 27 | 38 | 49 | 83 |
| 350 | 25 | 34 | 45 | 58 | 120 |

(ms). Specifically, encryption took 29 ms, while the homomorphic operations added 16 ms, and decryption required 22 ms. These times reflect the complexity of FHE, which performs operations on encrypted data without exposing sensitive information. Although the total computational time may seem high compared to nonencrypted operations, the privacy and security benefits make this overhead acceptable in privacy-critical environments like healthcare.

In comparison, the Paillier semihomomorphic encryption scheme was evaluated, which supports only addition operations. The Paillier scheme exhibited a lower overall overhead of 25 ms, with 10 ms for encryption, 6 ms for homomorphic addition, and 9 ms for decryption. However, the limited capability of Paillier to handle only addition makes it inadequate for more complex medical computations, such as multiplication, often required in healthcare scenarios.

The results highlight a tradeoff between performance and functionality. While introducing a higher computational cost, FHE provides the flexibility to perform addition and multiplication on encrypted data, making it a more suitable solution for secure, privacy-preserving computations in healthcare applications. In contrast, Paillier's lower overhead makes it faster but limits its use to more straightforward scenarios. Ultimately, the choice between FHE and Paillier depends on the required balance between computational efficiency and the complexity of operations needed.

### 5.3.5. *Cryptographic strength evaluation*

The encryption strength was analyzed by evaluating the system's resistance to known cryptographic attacks. We tested the system's homomorphic encryption by measuring computational overhead and encryption time for varying transaction sizes and node configurations. As seen in Table 20, the overhead associated with homomorphic encryption increased with the number of nodes and transactions. For example, with 20 nodes and 100 transactions, the encryption overhead was only 5 ms, while for 100 nodes and 350 transactions, the overhead increased to 120 ms. This shows a linear increase in computational time as the system scales, demonstrating the cryptographic robustness. The results show that encryption overhead remains manageable even as the number of nodes and transactions increases, ensuring that the system can securely process IoT data without significant performance degradation.

### 5.3.6. *Non-interactive Zero-Knowledge Proof privacy evaluation*

To evaluate the performance of the NIZKP scheme within the blockchain-based healthcare framework, a comprehensive statistical analysis was conducted. The key focus areas of the analysis were gas consumption, latency, and privacy validation, all of which are crucial to determining the system's feasibility and efficiency in a real-world blockchain environment.

Table 21. Gas consumption and transaction fees for NIZKP operations in blockchain-based healthcare system

| Action | Gas computation (Gwei) | Transaction fee (ETH) |
|---|---|---|
| Smart contract deployment | 2 350 000 | 0.021 |
| Certificate issuance | 150 000 | 0.00135 |
| Proof generation | 80 000 | 0.00072 |
| Prof verification | 110 000 | 0.001 |

Table 22. Latency analysis of certificate issuance and NIZKP verification in real-time blockchain environment

| Operation | Average latency (ms) |
|---|---|
| Certificate issuance | 1200 |
| Proof generation | 350 |
| Proof verification | 500 |

Table 21 shows the gas consumption and corresponding transaction fees for various NIZKP operations in the blockchain environment. Smart contract deployment, a one-time cost, required the highest gas consumption (2 350 000 Gwei), translating into a transaction fee of 0.021 ETH. Certificate issuance, which represents the process of generating and issuing a medical certificate on the blockchain, consumed 150 000 Gwei with a nominal transaction fee of 0.00135 ETH. Similarly, proof generation and proof verification required 80 000 Gwei (0.00072 ETH) and 110 000 Gwei (0.001 ETH), respectively, indicating efficient gas usage for the ongoing operations. This analysis demonstrates that after the initial deployment cost, the system operates with minimal gas consumption, making it feasible and cost-effective for large-scale healthcare applications.

Table 22 presents the average latency experienced during certificate issuance, proof generation, and proof verification. Certificate issuance recorded the highest latency at 1200 ms, reflecting the time required to execute an on-chain transaction. Proof generation, performed off-chain, had a low latency of 350 ms, highlighting the efficiency of off-chain computations in the NIZKP framework. Proof verification, an on-chain process, was also efficient, with an average latency of 500 ms. The low latency values for proof generation and verification indicate that the NIZKP-based system can meet the real-time demands of healthcare environments, where rapid validation of medical certificates is crucial.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

Table 23. Privacy validation results for NIZKP-based medical certificate verification

| Metrics | Percentage (%) |
|---|---|
| Successful verification | 100% |
| Privacy breaches detected | 0% |

Table 24. NIZKP verification time across different node and transaction configurations

| No. of transactions ($T_x$) | 20 nodes (ms) | 40 nodes (ms) | 60 nodes (ms) | 80 nodes (ms) | 100 nodes (ms) |
|---|---|---|---|---|---|
| 100 | 3 | 5 | 8 | 12 | 18 |
| 150 | 5 | 7 | 11 | 17 | 25 |
| 200 | 8 | 11 | 15 | 21 | 33 |
| 250 | 10 | 14 | 19 | 25 | 40 |
| 300 | 13 | 18 | 23 | 31 | 47 |
| 350 | 15 | 21 | 27 | 35 | 55 |

Table 23 summarizes the privacy validation results, confirming the system's secure handling of sensitive medical data. The experiments showed that 100% of the verifications were successful, with no privacy breaches detected. This reinforces the effectiveness of the NIZKP approach in ensuring that patient data remain private during certificate verification. By using cryptographic proofs without revealing sensitive information, the system guarantees high data privacy, addressing one of the core challenges in healthcare data management.

NIZKPs were analyzed to determine their effect on privacy during transaction verification. We evaluated the NIZKP verification time under different node and transaction loads. As seen in Table 24, the verification time for NIZKPs increased with the number of transactions and nodes but remained within acceptable limits. The verification time for 20 nodes and 100 transactions was 3 ms, while for 100 nodes and 350 transactions, it increased to 55 ms. The NIZKP verification time remains low enough to maintain privacy without significantly affecting system performance, even with increasing transaction loads and node counts.

The statistical analysis of the NIZKP execution demonstrates that the proposed framework offers a highly efficient and secure solution for medical certificate verification in blockchain-enabled healthcare systems. The low gas consumption ensures cost-effective operations, while the minimal latency for key processes such as proof generation and verification supports the system's ability to operate in real-time environments.

The privacy validation confirmed that no sensitive patient data was exposed during the proof verification process, affirming the framework's capacity to maintain high data privacy standards. These findings collectively address the concerns regarding scalability, cost, and security, highlighting the practical applicability of the NIZKP framework in real-world healthcare systems.

#### 5.3.7. *Consensus algorithm security evaluation*

The security of the hybrid PoW-PBFT consensus mechanism was evaluated by simulating attacks such as Sybil attacks and double-spending attempts. The consensus delay was measured based on the number of malicious nodes attempting to disrupt the network. Table 25 presents the delay introduced by the hybrid consensus mechanism in the presence of malicious nodes. With no malicious nodes, the delay was 25 ms. The delay increased as the number of malicious nodes increased, reaching 115 ms with 30 malicious nodes. The results demonstrate that the hybrid PoW-PBFT consensus algorithm is robust against Sybil and double-spending attacks, as the system continues to achieve consensus even in the presence of malicious nodes, albeit with an increased delay.

Table 25. Consensus delay in the presence of malicious nodes for Hybrid PoW-PBFT mechanism

| No. of malicious nodes | Consensus delay (ms) |
|---|---|
| 0 | 25 |
| 5 | 32 |
| 10 | 48 |
| 15 | 65 |
| 20 | 85 |
| 25 | 99 |
| 30 | 115 |

#### 5.3.8. *Data confidentiality and integrity*

We conducted penetration tests to assess the system's resistance to unauthorized data access. During the tests, the system's encryption and access control measures blocked attempts to access encrypted data stored on the blockchain and IPFS. In all test cases, unauthorized attempts to access data were unsuccessful. Table 26 summarizes the number of unauthorized attempts blocked by the system across different node configurations. The system effectively prevents unauthorized access, maintaining data confidentiality and integrity across increasing node and transaction loads.

This security and privacy evaluation demonstrates that the proposed blockchain architecture successfully addresses critical security concerns. Homomorphic encryption, ZKPs, and the hybrid PoW-PBFT consensus mechanism provide robust protection against common blockchain attacks. Additionally, the system maintains data confidentiality and integrity under increasing transaction and node loads, ensuring scalability while preserving privacy.

### 5.4. Comparison with state-of-artwork

In this section, we evaluate the performance of the proposed hybrid PoW-PBFT consensus mechanism in comparison with other widely used consensus algorithms, including Traditional PoW, Pure PBFT, Delegated Proof of Stake (DPoS), Proof of Authority (PoA), and Raft. The analysis focuses on key metrics such as gas consumption, transaction fees, processing speed, error rate, scalability, and consensus finality. The comparative performance data highlight the advantages and trade-offs of each approach, offering a comprehensive understanding of how the proposed method stands against state-of-the-art solutions in Table 27.

The Hybrid PoW-PBFT (Proposed) consensus mechanism demonstrates superiority in several key performance metrics. For instance, average gas consumption is reduced to 28 500 Gwei, significantly lower than the Traditional PoW (45 000 Gwei) and Pure PBFT (38 000 Gwei). This makes the hybrid model more energy-efficient and cost-effective. Regarding average transaction fees, the proposed mechanism achieves the lowest cost at 0.000022 ETH, significantly outperforming other models like DPoS and PoA, further enhancing its cost-efficiency for high-volume transactions.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

Table 26. Number of unauthorized access attempts blocked by the system under varying node configurations

| No. of transactions ($T_x$) | 20 nodes (no. of blocked attempts) | 40 nodes (no. of blocked attempts) | 60 nodes (no. of blocked attempts) | 80 nodes (no. of blocked attempts) | 100 nodes (no. of blocked attempts) |
|---|---|---|---|---|---|
| 100 | 10 | 18 | 24 | 31 | 40 |
| 150 | 14 | 22 | 30 | 39 | 48 |
| 200 | 19 | 28 | 38 | 49 | 60 |
| 250 | 24 | 35 | 45 | 56 | 70 |
| 300 | 28 | 40 | 52 | 64 | 80 |
| 350 | 35 | 45 | 58 | 70 | 90 |

Table 27. Comparative analysis of different consensus mechanisms with the proposed method

| Metrics | Traditional PoW | Pure PBFT | DPoS | PoA | Raft | Proposed Hybrid PoW-PBFT |
|---|---|---|---|---|---|---|
| Average gas computation (Gwei) | 45 000 | 38 000 | 32 500 | 30 000 | 29 500 | 28 500 |
| Average transaction fee (ETH) | 0.000050 | 0.000045 | 0.000032 | 0.000030 | 0.000028 | 0.000022 |
| Processing speed (Tx/second) | 5 | 15 | 25 | 30 | 28 | 35 |
| Error rate (%) | 2.5 | 1.2 | 0.6 | 0.4 | 0.3 | 0.5 |
| Scalability (max Tx/block) | 100 | 200 | 350 | 400 | 320 | 450 |
| Consensus finality (seconds) | 120 | 50 | 30 | 25 | 20 | 15 |

In terms of processing speed, the proposed model handles 35 transactions per second, surpassing both Traditional PoW (5 Tx/s) and Pure PBFT (15 Tx/s) while remaining competitive with more centralized systems like PoA (30 Tx/s). Additionally, the error rate is reduced to 0.5%, highlighting the robustness of the Hybrid PoW-PBFT model, which offers better reliability than Pure PBFT and other consensus mechanisms. Regarding scalability, the proposed approach supports up to 450 transactions per block, far exceeding the scalability of Traditional PoW (100 Tx/block) and Pure PBFT (200 Tx/block). Finally, consensus finality is improved dramatically to just 15 s, significantly faster than PoW (120 s) and even better than Pure PBFT (50 s), making it suitable for applications requiring quick finality and high throughput. As a result, the Hybrid PoW-PBFT consensus mechanism emerges as a highly suitable option for large-scale blockchain systems that demand high performance and efficiency. It is suitable for high-transaction-volume applications requiring robust security and efficient processing.

## 5.5. Discussion and findings

The integration of blockchain technology into IoT-based healthcare systems presents significant advancements in securing and managing healthcare data. This research paper has examined the effectiveness of a blockchain-based IoT healthcare management system, focusing on privacy-preserving data sharing and the impact of various factors on system performance. The following discussion highlights the key findings from the experimental evaluation and analysis of the proposed system:

- **Security and privacy enhancements:** The proposed blockchain-based system demonstrates a robust approach to enhancing the security and privacy of healthcare data. The system ensures data integrity and traceability by leveraging unique identification numbers for medical certificates. The hybrid consensus mechanism, which combines PoW and PBFT, provides a balanced solution for improving security and transaction speed. Additionally, homomorphic encryption for privacy-preserving computations and ZKPs for verifying certificates without revealing sensitive patient data significantly strengthen the privacy and trustworthiness of the system.
- **Performance analysis:** The analysis of operational costs reveals that the proposed system incurs varying gas costs depending on the functions performed. The cost of executing operations such as registration, generation of blocks, issuing certificates, and verifying certificates was evaluated using the Etherscan tool. Results indicate that costs increase with the complexity of transactions. For instance, the registration function required 34 219 Gas, while verifying a certificate required 37 100 Gas. This variance highlights the impact of different operations on the overall cost and emphasizes the need for efficient gas management in blockchain-based systems. When comparing Ethereum-based blockchain platforms with SQL databases, notable differences in performance metrics were observed. While the Ethereum blockchain exhibited higher latency and processing times, it provided superior security and privacy features compared to traditional SQL databases. For example, the issuing certificate operation on Ethereum had a latency of 5.32 ms and a processing time of 6.37 ms. In contrast, SQL platforms had lower latency but lacked the security benefits of blockchain. This trade-off underscores the importance of considering security and performance when evaluating blockchain-based systems.
  Scalability analysis of the proposed system uncovered several challenges associated with increasing the number of IoT nodes and transactions. Registration time, consensus execution time, block creation, and access times progressively increased with the number of nodes and transactions. For instance, registering 100 nodes took 308 s, while generating a block with 100 nodes and 350 transactions required 1433 s. These findings highlight the scalability challenges inherent in blockchain systems and underscore the need for optimized consensus mechanisms and block management strategies to handle more extensive networks efficiently. The evaluation of gas price consumption demonstrated a steady increase in

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

nodes and transactions. For example, with 100 nodes and 350 transactions, the Gas price reached 74 510, compared to 20 000 for 20 nodes and 100 transactions. This increase reflects the growing computational and storage demands as the network scales, emphasizing the need for cost-effective strategies to manage operational expenses. Regarding storage size and off-chain utilization, the analysis of IPFS-based off-chain storage indicated that storage requirements increase with the number of transactions. For instance, with 350 transactions, the storage size was 39 KB, compared to 9 KB for 100 transactions. This growing storage requirement highlights the importance of efficient data management strategies to accommodate larger transaction volumes without compromising system performance.

- **Latency and throughput:** The evaluation of latency and throughput revealed that the Ethereum blockchain-based system generally had higher latency and lower throughput than SQL databases. For example, the latency for verifying certificates on Ethereum was 7.12 ms, compared to 2.63 ms on SQL platforms. Throughput was also lower on Ethereum, with 6.38 Kbps compared to 10.57 Kbps on SQL platforms. These performance differences highlight the trade-offs between blockchain's security benefits and traditional database performance metrics.

In conclusion, integrating blockchain technology into IoT-based healthcare systems offers significant security, privacy, and data integrity advantages. The proposed system demonstrates a promising approach to addressing privacy concerns and managing healthcare data effectively. The findings from this study provide valuable insights for developing more efficient and secure blockchain-based healthcare solutions in the future.

# 6. CONCLUSION AND FUTURE WORK

## 6.1. Conclusion

This paper presents a novel blockchain-based IoT application to improve medical certificate management's security, privacy, and efficiency in healthcare systems. The proposed solution uses unique identification numbers for medical certificates. It incorporates a hybrid consensus mechanism combining PoW and PBFT for enhanced security and transaction speed. Privacy is safeguarded through homomorphic encryption and NIZKPs, which enable secure medical certificate verification without revealing sensitive data. Integrating the IPFS also ensures scalable and efficient data storage. An IDS has been added to monitor IoT traffic and detect security threats. Experimental results demonstrate the system's robustness, achieving an accuracy of 98.1% for the IDS. At the same time, blockchain evaluation metrics show low latency, high throughput, and enhanced security, positioning this solution as a transformative approach to secure healthcare management.

## 6.2. Limitation

The application faces high energy consumption and computational costs due to the PoW component. Advanced cryptography adds overhead, and IPFS may struggle with data retrieval and reliability issues.

## 6.3. Potential industrial applications

The proposed system offers significant industrial benefits across various sectors, particularly in healthcare:

- **Healthcare providers**: Simplifies the management and verification of medical certificates, improving operational efficiency by reducing manual processes and paperwork.
- **Insurance companies**: Enhances the validation of medical claims and records, minimizing fraud and reducing administrative errors using secure blockchain-backed certificates.
- **Medical institutions**: Provides an efficient platform to manage large volumes of medical records, ensuring the security and privacy of sensitive patient information through advanced encryption and IDS monitoring.

## 6.4. Future work

Future work will explore more efficient consensus algorithms, optimize smart contracts, improve IPFS integration, and conduct real-world testing to enhance scalability and performance.

## 6.5. Societal applications

This system holds great potential for improving the security, privacy, and trust in healthcare operations:

- **Data security and privacy**: Ensures patient data remains confidential, addressing data breaches and unauthorized access concerns. Patients can have peace of mind knowing their medical records are secure.
- **Trust in healthcare**: By leveraging transparent and secure certificate management, this system reduces fraudulent activities in medical certificate handling and increases trust among patients, healthcare providers, and insurers.
- **Wider technological adoption**: The adoption of blockchain in healthcare can inspire its broader integration across other sectors, enhancing overall data security and setting a standard for secure information sharing in various industries.

# AUTHOR CONTRIBUTIONS

All authors contributed equally in data curation, methodology, visualization, and writing—original draft.

# SUPPLEMENTARY DATA

Supplementary data are available at *The Computer Journal* online.

Conflict of interest: None declared.

# FUNDING

No funding recieved.

# DATA AVAILABILITY

The dataset is publicly available on the repository.

**Himanshu Nandanwar** is a research scholar at Delhi Technological University. He received his M. Tech and B. Tech degrees in Information Technology from Delhi Technological University. His research interests include the Internet of Things, Networking, Cloud computing, IoT security, and 5G/6G.

**Rahul Katarya** is currently working as a Professor at Delhi Technological University, Delhi, India, in the Department of Computer Science and Engineering. He has published more than 150 papers in refereed conferences and journals. His research interests include Data Science, Big Data analytics, Web Mining, Social

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

Networks Analysis, Recommender Systems, Machine Learning, Data Mining, and Human Behavior Evolutionary Computing.

## REFERENCES


1. Da-Yu KAO, Hsiao SC, Raylin TSO. Analyzing WannaCry ransomware considering the weapons and exploits. In: *2019 21st International Conference on Advanced Communication Technology (ICACT)*. PyeongChang, Korea (South): IEEE, 2019, 1098–1107. https://doi.org/10.23919/ICACT.2019.8702049.
2. Sharma P, Namasudra S, Crespo RG. *et al.* EHDHE: enhancing security of healthcare documents in IoT-enabled digital healthcare ecosystems using blockchain *Inf Sci*. 2023;**629**:703–718. https://doi.org/10.1016/j.ins.2023.01.148.
3. Sun J, Zhu X, Zhang C. *et al.* HCPP: Cryptography based secure EHR system for patient privacy and emergency healthcare. In: *2011 31st International Conference on Distributed Computing Systems*. Minneapolis, MN, USA: IEEE, 2011, 373–382. https://doi.org/10.1109/ICDCS.2011.83.
4. Daraghmi EY, Daraghmi YA, Yuan SM. MedChain: a design of blockchain-based system for medical records access and permissions management *IEEE Access*. 2019;**7**:164595–164613. https://doi.org/10.1109/ACCESS.2019.2952942.
5. Newaz AI, Haque NI, Sikder AK. *et al.* Adversarial attacks to machine learning-based smart healthcare systems. In: *GLOBECOM 2020–2020 IEEE Global Communications Conference*. Taipei, Taiwan: IEEE, 2020, 1–6. https://doi.org/10.1109/GLOBECOM42002.2020.9322472.
6. Nakamoto S, Bitcoin A. A peer-to-peer electronic cash system. *Bitcoin* 2008;**4**:15. https://bitcoin.org/bitcoin.pdf.
7. Salankar PS, Avasthi V, Pasricha A. Lightweight authentication scheme based on modified EAP security for CoAP protocol-based IoMT applications *Int J Inf Comput Secur*. 2023;**20**:176–198. https://doi.org/10.1504/IJICS.2023.128026.
8. Miao J, Wang Z, Wu Z. *et al.* A blockchain-enabled privacy-preserving authentication management protocol for internet of medical things *Expert Syst Appl*. 2024;**237**:121329. https://doi.org/10.1016/j.eswa.2023.121329.
9. Vatambeti R, Krishna EP, Karthik MG. *et al.* Securing the medical data using enhanced privacy preserving based blockchain technology in internet of things *Clust Comput*. 2024;**27**:1625–1637. https://doi.org/10.1007/s10586-023-04056-0.
10. Kumari D, Parmar AS, Goyal HS. *et al.* Healthrec-chain: patient-centric blockchain enabled ipfs for privacy preserving scalable health data *Comput Netw*. 2024;**241**:110223. https://doi.org/10.1016/j.comnet.2024.110223.
11. Rhayem A, Mhiri MBA, Drira K. *et al.* A semantic-enabled and context-aware monitoring system for the internet of medical things *Expert Syst*. 2021;**38**:e12629. https://doi.org/10.1111/exsy.12629.
12. Sharma A, Sarishma, Tomar R. *et al.* Blockchain based smart contracts for internet of medical things in e-healthcare *Electronics*. 2020;**9**:1609. https://doi.org/10.3390/electronics9101609.
13. Masood I, Daud A, Wang Y. *et al.* A blockchain-based system for patient data privacy and security *Multimed Tools Appl*. 2024;**83**:60443–60467. https://doi.org/10.1007/s11042-023-17941-y.
14. Ahmed SF, Alam MSB, Afrin S. *et al.* Insights into internet of medical things (IoMT): data fusion, security issues and potential solutions *Inf Fusion*. 2024;**102**:102060. https://doi.org/10.1016/j.inffus.2023.102060.
15. Soleymani SA, Goudarzi S, Anisi MH. *et al.* A privacy-preserving authentication scheme for real-time medical monitoring systems *IEEE J Biomed Health Inform*. 2022;**27**:2314–2322. https://doi.org/10.1109/JBHI.2022.3143207.
16. Nagarajan SM, Deverajan GG, Kumaran U. *et al.* Secure data transmission in internet of medical things using RES-256 algorithm *IEEE Trans Industr Inform*. 2021;**18**:8876–8884. https://doi.org/10.1109/TII.2021.3126119.
17. Chen X, He D, Khan MK. *et al.* A secure certificateless signcryption scheme without pairing for internet of medical things *IEEE Internet Things J*. 2022;**10**:9136–9147. https://doi.org/10.1109/JIOT.2022.3233180.
18. Wu G, Wang S, Ning Z. *et al.* Blockchain-enabled privacy-preserving access control for data publishing and sharing in the internet of medical things *IEEE Internet Things J*. 2021;**9**:8091–8104. https://doi.org/10.1109/JIOT.2021.3138104.
19. Saidi H, Labraoui N, Ari AAA. *et al.* DSMAC: privacy-aware decentralized self-management of data access control based on blockchain for health data *IEEE Access*. 2022;**10**:101011–101028. https://doi.org/10.1109/ACCESS.2022.3207803.
20. Barnawi A, Chhikara P, Tekchandani R. *et al.* A differentially privacy assisted federated learning scheme to preserve data privacy for IoMT applications *IEEE Trans Netw Serv Manag*. 2024;**21**:4686–4700. https://doi.org/10.1109/TNSM.2024.3393969.
21. Onesimu JA, Karthikeyan J, Sei Y. An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services *Peer Peer Netw Appl*. 2021;**14**:1629–1649. https://doi.org/10.1007/s12083-021-01077-7.
22. Sarier ND. Privacy preserving biometric authentication on the blockchain for smart healthcare *Pervasive Mob Comput*. 2022;**86**:101683. https://doi.org/10.1016/j.pmcj.2022.101683.
23. Younis M, Lalouani W, Lasla N. *et al.* Blockchain-enabled and data-driven smart healthcare solution for secure and privacy-preserving data access *IEEE Syst J*. 2021;**16**:3746–3757. https://doi.org/10.1109/JSYST.2021.3092519.
24. Nie X, Zhang A, Chen J. *et al.* Blockchain-empowered secure and privacy-preserving health data sharing in edge-based IoMT *Secur Commun Netw*. 2022;**2022**:1–16. https://doi.org/10.1155/2022/8293716.
25. Wu G, Wang S, Ning Z. *et al.* Privacy-preserved electronic medical record exchanging and sharing: a blockchain-based smart healthcare system *IEEE J Biomed Health Inform*. 2021;**26**:1917–1927. https://doi.org/10.1109/JBHI.2021.3123643.
26. Alzubi OA, Alzubi JA, Shankar K. *et al.* Blockchain and artificial intelligence enabled privacy-preserving medical data transmission in internet of things *Trans Emerg Telecommun Technol*. 2021;**32**:e4360. https://doi.org/10.1002/ett.4360.
27. Xu G, Qi C, Dong W. *et al.* A privacy-preserving medical data sharing scheme based on blockchain *IEEE J Biomed Health Inform*. 2022;**27**:698–709. https://doi.org/10.1109/JBHI.2022.3203577.
28. Tan L, Yu K, Shi N. *et al.* Towards secure and privacy-preserving data sharing for COVID-19 medical records: a blockchain-empowered approach *IEEE Trans Netw Sci Eng*. 2021;**9**:271–281. https://doi.org/10.1109/TNSE.2021.3101842.
29. Hossein KM, Esmaeili ME, Dargahi T. *et al.* BCHealth: a novel blockchain-based privacy-preserving architecture for IoT healthcare applications *Comput Commun*. 2021;**180**:31–47. https://doi.org/10.1016/j.comcom.2021.08.011.
30. Singh S, Rathore S, Alfarraj O. *et al.* A framework for privacy-preservation of IoT healthcare data using federated learning and blockchain technology *Futur Gener Comput Syst*. 2022;**129**:380–388. https://doi.org/10.1016/j.future.2021.11.028.
31. Zhang G, Yang Z, Liu W. Blockchain-based privacy preserving e-health system for healthcare data in cloud *Comput Netw*. 2022;**203**:108586. https://doi.org/10.1016/j.comnet.2021.108586.


Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

32. Alzubi JA, Alzubi OA, Singh A. *et al.* Cloud-IIoT-based electronic health record privacy-preserving by CNN and blockchain-enabled federated learning *IEEE Trans Industr Inform*. 2022;**19**:1080–1087. https://doi.org/10.1109/TII.2022.3189170.
33. Huang H, Zhu P, Xiao F. *et al.* A blockchain-based scheme for privacy-preserving and secure sharing of medical data *Comput Secur*. 2020;**99**:102010. https://doi.org/10.1016/j.cose.2020.102010.
34. Wang Y, Zhang A, Zhang P. *et al.* Security-aware and privacy-preserving personal health record sharing using consortium blockchain *IEEE Internet Things J*. 2021;**9**:12014–12028. https://doi.org/10.1109/JIOT.2021.3132780.
35. Kumar P, Kumar R, Gupta GP. *et al.* A blockchain-orchestrated deep learning approach for secure data transmission in IoT-enabled healthcare system *J Parallel Distrib Comput*. 2023;**172**:69–83. https://doi.org/10.1016/j.jpdc.2022.10.002.
36. Xie Q, Zhu F, Feng X. Blockchain-enabled data sharing for IoT: a lightweight, secure and searchable scheme *J Syst Archit*. 2024;**154**:103230. https://doi.org/10.1016/j.sysarc.2024.103230.
37. Li C, Dong M, Xin X. *et al.* Efficient privacy-preserving in IoMT with blockchain and lightweight secret sharing *IEEE Internet Things J*. 2023;**10**:22051–22064. https://doi.org/10.1109/JIOT.2023.3296595.
38. Wei X, Yan Y, Guo S. *et al.* Secure data sharing: Blockchain-enabled data access control framework for IoT *IEEE Internet Things J*. 2021;**9**:8143–8153. https://doi.org/10.1109/JIOT.2021.3111012.


Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf065/8139855 by Delhi Technological University user on 27 May 2025

© The British Computer Society 2025. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com
*The Computer Journal*, 2025, 1–25
https://doi.org/10.1093/comjnl/bxaf065
Original Article

Software Quality Journal (2025) 33:25
https://doi.org/10.1007/s11219-025-09724-5

RESEARCH

# QNet: exploring deep learning for quantum code smell detection

**Ruchika Malhotra**[1] · **Bhawna Jain**[1,2] · **Marouane Kessentini**[3]

Accepted: 19 May 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

**Abstract**
Quantum computing (QC) has surged as a burgeoning domain, driving the evolution of novel programming paradigms. Despite extensive exploration, a critical aspect remains underexplored: *detecting code smells in quantum programs (QPs)*. Code smells, indicative of potential maintenance challenges, have been extensively studied in classical programming but pose unique hurdles in the quantum realm due to inherent disparities caused by unstable states in QC. This paper proposes a novel approach leveraging deep learning (DL) techniques for detecting quantum code smells (QCS). A comprehensive ablation study compares DL methodologies, including Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM), and proposes a hybrid model combining CNN, LSTM, and GRU layers. Critical research questions are posed regarding QCS influence on error rates in quantum computers, the relative impact of different QCS on quantum system performance, and the complementary behaviour of distinct DL models in QCS detection. Through manual curation of a labelled dataset comprising 136 open-source projects, quantum circuits are extracted and analyzed. When evaluated on the quantum dataset, this hybrid model outperforms single-layer models. Furthermore, a comparative analysis with a transfer learning (TL) approach employing a pre-trained Bidirectional Encoder Representation Transformers (BERT) model underscores the superiority of the proposed DL-based solution. The proposed model achieves an impressive accuracy rate of **92.86%**, surpassing existing DL and TL approaches. In conclusion, this research demonstrates the potential of DL to identify QCS, with the hybrid model offering avenues for further discoveries in QCS detection using DL.

**Keywords** Code smells · Deep learning · Quantum-specific code smell · Smell detection · Quantum computing

✉ Bhawna Jain
dr.bhawnajain.e@gmail.com

Ruchika Malhotra
ruchikamalhotra@dtu.ac.in

Marouane Kessentini
kessentm@gvsu.edu

1 Delhi Technological University, Delhi, India

2 Indira Gandhi Delhi Technical University for Women, Delhi, India

3 Grand Valley State University, Michigan, USA

Published online: 27 May 2025

## 1 Introduction

Quantum computing (QC) is the discipline of computing in which quantum physics is used to address complicated real-world problems in various fields, including machine learning (ML) (Benedetti et al., 2016), computational physics (Barends et al., 2014), and cryptography (Bennett amb Brassard, 2020) in much faster and more efficient ways than classical computers. It uses unique features such as superposition and entanglement, allowing for exponential speedups in activities such as factorisation and searching. With this potential, private businesses, government labs, and educational institutions worldwide have significantly invested in QC in recent years (Zhao, 2020). This collaborative effort has led to rapid technological and scientific advancements in the field, propelling QC into the forefront of cutting-edge research. This surge of interest has, in turn, sparked the creation of novel programming languages and frameworks tailored specifically to adopt QC (Shaydulin et al., 2020).

The availability of dedicated quantum programming languages and frameworks, such as Cirq by Google, Q# by Microsoft, and Qiskit by IBM, has been instrumental in accelerating the development of more complex Quantum Programs (QPs) (Ball et al., 2021). These tools offer an abstraction layer that simplifies the implementation of quantum algorithms and allows practitioners to focus on the core logic of their applications. With the rise of Quantum-as-a-Service (QaaS) platforms providing access to universal quantum devices, QP has become more accessible to a broader community of researchers and developers (Leymann et al., 2020).

The continuous availability of QPs is crucial for the seamless functioning of quantum applications. However, constant monitoring and maintenance have become challenging as QC technology advances due to several factors (Li et al., 2020; Huang amb Martonosi, 2019).

In classical software development, it has been observed that poor design and implementation choices (Gupta amb Kumar Singh, 2022; Barbez et al., 2019), often stemming from developers' lack of experience and skills, can hinder effective maintenance (Shaydulin et al., 2020; Ahmed et al., 2017). Similarly, as the field of QC is emerging, programming practices are still evolving rapidly, and related libraries and frameworks frequently undergo significant revisions (Ali et al., 2021). Consequently, QPs often consist of code pieced together to accommodate these changes. This code-patching approach can lead to potential issues and complexities as QPs grow in size and complexity (Zhao et al., 2021).

Recognizing these challenges, it is crucial for the software engineering research community to proactively investigate techniques to aid in the maintenance of QPs before widespread problems arise (Fowler and Beck, 1997; Ahmed et al., 2017). Despite the growing importance of QC, the software engineering research community still needs to fully explore the maintenance of QPs from a purely quantum perspective.

In the realm of software engineering, prior research has extensively explored maintenance issues, mainly focusing on indicators known as "code smells", and methodologies to identify and measure their impact on classical software (Jebnoun et al., 2020; Tufano et al., 2017; Yamashita amb Moonen, 2013). Quantum code smell (QCS) detection in the context of QC is an emerging and relatively unexplored area of research. One such noteworthy contribution in this area is the work by Chen et al. (2023), where they introduced a tool designed for detecting QCS in QPs. Additionally, Openja et al. (2022) employed ML algorithms for a similar purpose. Nevertheless, only a few studies have been dedicated to identifying QCS in QPs.

Considering the gaps mentioned above, this study investigates the following research questions (RQs):

***RQ1: How do QCS influence the frequency of errors in quantum computers?***

This RQ explores the complex relationship between QCS and the errors they amplify within quantum computers. It seeks to understand how QCS affects the reliability and performance of quantum systems.

***RQ2: What is the relative impact of different QCS on the performance of quantum systems?***

Each QCS carries a severity value that indicates its potential to escalate error likelihood. This RQ aims to rank QCS by analyzing their severity values and provides insights into identifying QCS that pose higher risks of error amplification compared to those with lesser consequences.

***RQ3: Do the outcomes of DL models, developed with distinct hyper-parameters, exhibit complementary behaviour?***

This RQ focuses on conducting an in-depth analysis of the hyper-parameters in CNN, GRU, and LSTM networks that can assist in selecting the ideal layers for constructing a hybrid model. This RQ explores whether the chosen layers and hyper-parameters for each specific model align effectively with their respective capabilities in detecting various QCS.

***RQ4: What are the predictive capabilities of different DL models developed for QCS detection?***

This RQ evaluates diverse DL models across various performance metrics to assess their effectiveness in detecting QCS. Comparison among the Receiver Operating Characteristic Curve (ROCC) of different models employed for QCS detection is featured to categorize and quantify the results. Each model is assessed independently and individually based on these specified performance parameters.

***RQ5: What is the effectiveness of the proposed model when compared with state-of-the-art models?***

This RQ assesses the predictive power of the proposed model by comparing it to various models. Additionally, it contrasts hybrid models with large language models (LLMs), transformers and transfer learning (TL), using a pre-trained BERT model trained on classical computers code smell detection for QCS detection. The study aims to determine if the proposed model surpasses the state-of-the-art models, offering insights into both approaches' strengths and weaknesses in QCS detection.

***RQ6: How does the proposed hybrid QCS detection model outperform existing QCS detection tools?***

This RQ aims to evaluate the proposed hybrid model's efficiency and adaptability compared to existing tools such as QChecker (Zhao et al., 2023), QPAC (Nayak et al., 2023), and QSmell (Chen et al., 2023). The focus lies on runtime complexity, the format of inputs required, and compatibility with prominent quantum computing frameworks, namely Qiskit, Q#, and Cirq.

Thus, the primary contributions of this research study are threefold. Firstly, it investigates the domain of QCS and its potential impact on errors within QPs, recognizing the significance of addressing these errors to enhance quantum computer performance. Secondly, it explores the feasibility of employing deep learning (DL) techniques to detect QCSs within the code, aiming to improve code quality. Lastly, the study delves into the potential of hybrid DL

algorithms, merging multiple DL techniques to ascertain whether they can surpass current methods and elevate QCS detection to new heights.

To the authors' knowledge, this study is the first to apply DL models to detect QCS. The study's results confirm the superiority of the proposed model as it outperforms the individual DL and pre-trained model. Furthermore, the proposed model was found to be comparable with baseline techniques. Software practitioners can use this framework to detect QCS. Therefore, focusing on these smells and their early detection would guarantee a maintainable and better-quality product.

The remainder of the paper is organized as follows: Section 2 describes the background of QC and QCS. Section 3 briefs about the related work done in this field and how this paper deals with previous research gaps. Section 4 outlines the proposed methodology, presenting a detailed account of the steps undertaken in this study. Section 5 discusses the results. Section 6 points out the threats to validity in this study. Finally, Section 7 concludes the paper and highlights the guidelines for future works.

## 2 Background

This section provides an introductory understanding of QC, detailing its key concepts and principles. Furthermore, it delves into the realm of QCS, which uniquely addresses issues around QP. Examining these distinct code anomalies reveals quantum developers' challenges and intricacies while ensuring code quality and reliability in QC.

### 2.1 Quantum computing

QC is an emerging field of computational science that uses the concepts of quantum physics to carry out complex calculations at unimaginably fast rates for conventional computers. Quantum computers use quantum bits $|0\rangle$ and $|1\rangle$, also known as qubits, as the basic unit of information, as opposed to classical computers that use either a 0 or a 1. Due to a phenomenon known as superposition, qubits can exist in numerous states simultaneously. This enables quantum computers to explore a vast range of possibilities concurrently. Qubits can also be entangled, which means their states are inextricably linked even when a great distance separates them. Because of this characteristic, quantum computers can resolve currently computationally intractable issues. Table 1 summarises the basics of quantum programming used to write QPs.

In 1981, physicist (Feynman, 1982) suggested the idea of quantum programming as a revolutionary method for effectively simulating quantum systems. Since then, substantial developments in quantum physics and information theory have propelled the growth at a breakneck pace. The creation of Shor's algorithm by Shor (1997) demonstrated the quantum computer's potential to handle some problems exponentially faster than conventional com-

**Table 1** Quantum Programming Basic Terminology

| Parameter | Notation | Description |
|---|---|---|
| Qubit | $\|\psi\rangle = \alpha\|0\rangle + \beta\|1\rangle$ | $\|0\rangle$ or $\|1\rangle$ or linear combination of both |
| Quantum Registers | q() | Performs quantum computation |
| Quantum Gates | h(), cx(), x(), etc. | Modifies qubit's probability magnitude and phase component |
| Measurement | measure() | Measure the results |
| Classical Registers | c() | Stores the measured result |

puters and was a critical turning point in 1994. This method sparked a lot of curiosity and changed how computing was possible.

Novel coding techniques were developed as technology advanced to exploit quantum systems efficiently. But much like conventional computers, quantum computers are prone to mistakes, leading to the idea of "bugs" or "QCS" in QP. Several things, including structure, noise, decoherence, and erroneous control mechanisms in quantum gear, can cause these technological debts to build up. Technical deficits in QC must be addressed to ensure quantum algorithms' and applications' efficiency and dependability. Quantum calculations can be severely hampered by unpaid technical debts, which results in errors and poor performance. To fully realise the potential of and open the door to valuable and significant quantum applications, in-depth research and approaches to pinpointing and addressing these problems are essential.

### 2.2 Quantum code smells

As technology advances and QPs become more complex, researchers have identified QCS in these programs. This study aims to demonstrate the effectiveness of applying DL algorithms to detect five specific QCS that have emerged in QPs. Furthermore, this research can be extended to include the detection of other QCS as they arise. This section will provide definitions and parameters for the five QCS introduced in the study by Chen et al. (2023). These parameters are critical in establishing a metric value derived from the code's execution matrix. If this metric value exceeds a predetermined threshold, the code is labelled as containing QCS. Conversely, if the metric value remains below the threshold, the code is considered free from detected QCS. This approach serves as an effective method to identify and categorise potential issues within QP.

**Repeated set of operations on circuit (ROC)** When a set of operations is iteratively applied to a consistent number of qubits within a circuit, a Repeated set of Operations on Circuit (ROC) phenomenon emerges. This concept is illustrated in Fig. 1, which showcases the "Test Amplitude Estimator" circuit. Notably, the critical observation in the circuit is that the initial operation sequence is replicated four times in an unvarying order. As a result, the ROC value for this scenario is 4.

**Long circuit (LC)** A circuit that involves an extensive sequence of gates operating on a single qubit but can be more efficiently achieved using a significantly smaller number of gates is referred to as a Long Circuit. These circuits exhibit program and operational lengths that exceed a predefined threshold value. When a circuit displays considerable depth, i.e., a high number of operations, the error rate on the backend of the quantum circuit tends to rise. Simplifying these lengthy circuits into more concise ones can lead to substantial

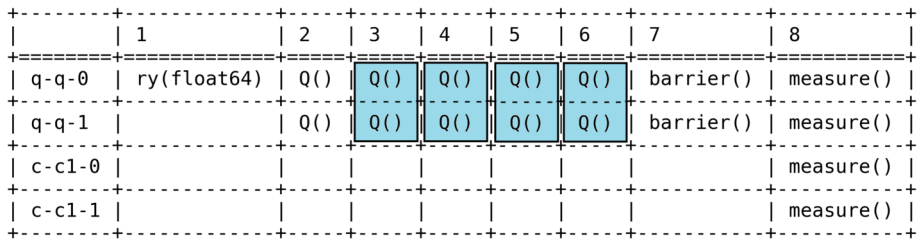



**Fig. 1** ROC- Test Amplitude Estimator circuit

computational time savings. The figure illustrates a circuit labelled "Load Qasm," which contains a long circuit. As seen in Fig. 2, the quantum circuit's length is notably extended. This is due to the concurrent utilisation of four quantum registers and four classical registers in three distinct operations, resulting in a combined circuit length of 28, which aligns with the LC metric value.

**Idle qubit(IdQ)** In the context of quantum coding, it is a common practice to initialise all qubits at the beginning. However, some qubits are intentionally left without active operations for a certain period. During this period of inactivity, the qubits tend to lose their stability. To restore their strength, a separate set of functions is applied. This state of the quantum circuit is referred to as Idle Qubit (IdQ). The circuit depicted in Fig. 3 exemplifies this phenomenon. Notably, qubit $|q4\rangle$ is initialised initially but remains inactive for four subsequent operations. Consequently, the IdQ value for this circuit configuration is 4.

**Initialization of qubits (IQ)** The most stable state of a qubit is typically observed as $|0\rangle$. However, during initialisation and subsequent gate operations, a qubit's state might transiently shift to an unstable state. Notably, this transition occurs when actions on other qubits separate the qubit's initialisation and gate operations. This scenario is termed the Initialization of Qubits (IQ). IQ is primarily different from IdQ as in the former, the difference between the

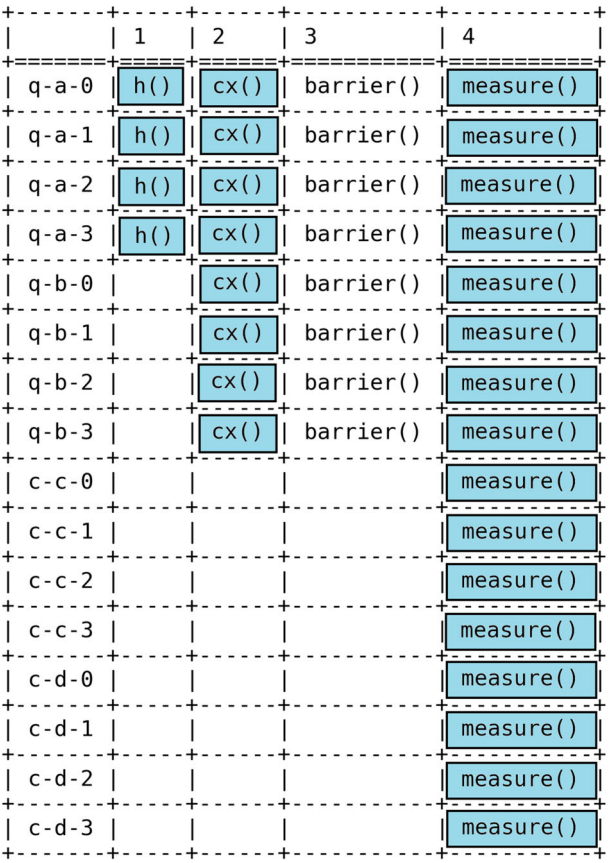



**Fig. 2** LC- Load Qasm circuit

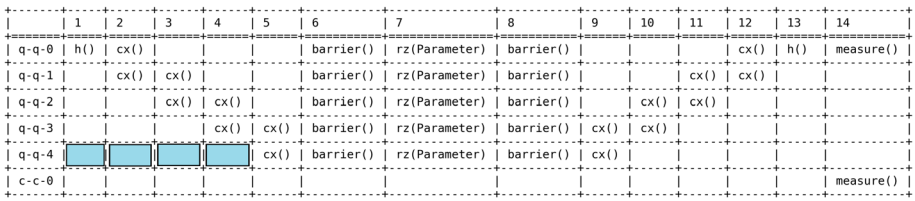



**Fig. 3** IdQ- Parametrized circuit

first and the second operation matters, while in the latter, the maximum time a qubit is left alone is significant.

The Fig. 4 illustrates the execution matrix of the Bernstein Vazirani Circuit. This circuit efficiently discovers a secret binary number, outperforming classical computers. While a classical computer requires O(n) time to uncover the n-bit private number, this algorithm accomplishes the task in O(1) time. In the depicted figure, qubit $|q2\rangle$ is initialised at the circuit's commencement, and a Hadamard gate(x()) gate is applied to it. Yet, the subsequent gate application to this qubit follows three operations. Consequently, this particular circuit exhibits an IQ value of 3, indicating a potential QCS.

**Intermediate measurements (IM)** The measure operation in QP extracts classical information from quantum bits. However, it's important to note that the quantum bit collapses after this operation and becomes unstable. The measure operation is typically strategically placed as the final step within a quantum circuit. This precaution is taken because performing measurements at intermediate stages can introduce errors to the circuit. However, this scenario is called Intermediate Measurement (IM) when the measurement function is inserted within the circuit. Therefore, adhering to the principle of deferred measurement, it is advisable to postpone all measurements until the final section of the quantum circuit.

As shown in Fig. 5, the circuit "Stochastic Swap" features a measure operation occurring in the middle of the sequence. This departure from the deferred measurement principle results in an IM value of 1 for this circuit configuration.

## 3 Related work

The concept of code smells was initially introduced by Fowler in his work (Fowler, 2002). These code smells were defined as defects contributing to the software code's complexity and difficulty. Over time, diverse techniques have been developed to identify code smells.

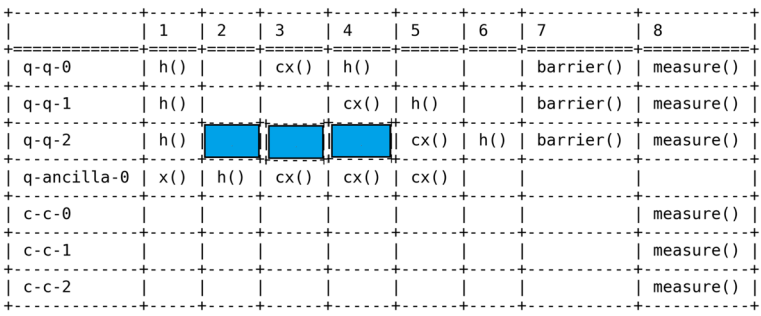



**Fig. 4** IQ- Bernstein Vazirani circuit

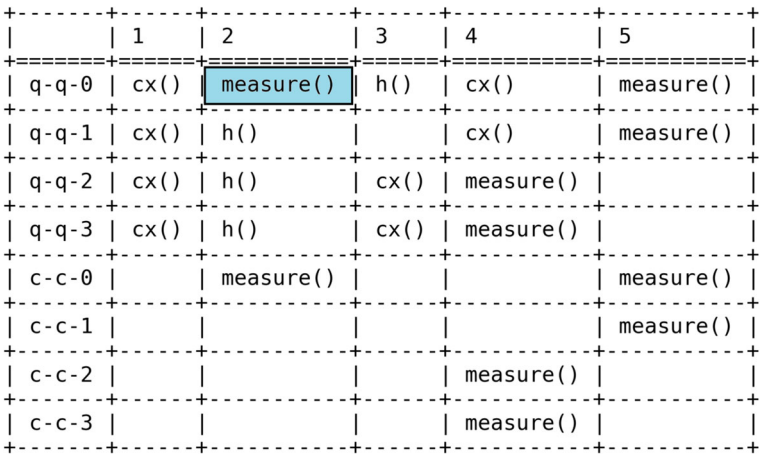



**Fig. 5** IM- Stochastic Swap circuit

Various approaches for detecting code smells have been utilised for traditional computers. Initially, code was manually checked for code smells (Dewangan et al., 2022; Van Oort et al., 2021). Later, researchers began employing heuristic-based techniques (Pecorelli et al., 2019; Lewowski amb Madeyski, 2022), generic principles or guidelines that help detect potential code smells. As the length and complexity of programs increased, newer methods, such as metric-based detection, were used to produce faster and more reliable results (Di Nucci et al., 2018).

Subsequently, ML algorithms were utilised for the identical objective. A range of ML algorithms, such as Random Forest, Naive Bayes, Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and ensemble ML, were frequently employed for the task of code smell detection (Muralidhar et al., 2021; Dewangan et al., 2021; Saheb-Nassagh et al., 2022; Alkharabsheh et al., 2022; Gesi et al., 2022). However, these traditional ML approaches often struggled with capturing complex relationships and dependencies within the code, particularly in large and intricate projects.

Recent studies began to shift towards more sophisticated methods, notably DL algorithms, which have proven more effective in capturing intricate patterns and dependencies in code structure and semantics. DL models, such as CNN, LSTM, and GRU, have been employed to detect code smells with remarkable accuracy, highlighting the shift from heuristic-based and machine-learning techniques to more advanced DL-based solutions. Bhave et al., as outlined in Bhave amb Sinha (2022), proposed innovative CNN and bi-LSTM models to ascertain whether code exhibits a code smell. This novel approach yielded an impressive accuracy rate of 91.2%.

Furthermore, specific code smells have been targeted in specialised studies, creating DL models tailored for particular instances such as feature envy (Liu et al., 2018; Guo et al., 2019), brain class, and brain method (Tarwani amb Chug, 2022). These models often leverage structural characteristics for detection, while alternate methods incorporating semantic features, as depicted in Zhang et al. (2022), have also been explored.

In addition to these DL approaches, hybrid architectures that combine multiple DL layers, such as CNN and LSTM, have also been developed (Das et al., 2019; Gupta et al., 2021). These hybrid models have shown superior performance by capturing both spatial and temporal dependencies within the code. Notably, advancements in DL have also introduced TL (Sharma et al., 2021), which involves training in one programming language and applying

the knowledge to a smaller dataset of another language. Incorporating pre-trained models like CodeBERT and codeT5 was also explored in Ma et al. (2023). These approaches have primarily been employed in classical computer environments for code smell detection.

Moreover, recent research has attempted to bridge TL across programming languages, which allows leveraging knowledge from well-established languages like Java or C++ to improve code smell detection in Python and other languages. This adaptability has shown the potential to enhance detection accuracy across diverse software projects.

However, since this paper focuses on code smell detection within the Python programming language, it's essential to highlight prior research endeavours in this context. Sandouka amb Aljamaan (2023) used ML algorithms to detect Python code smells. Bafatakis et al. (2019) and Omari amb Martinez (2020) utilised the Python code smell detection tool, Pylint. Additionally, Chen et al. (2016) developed Pysmell for similar purposes, tailored explicitly to Python.

In parallel, recent research efforts have focused on more extensive datasets and complex project repositories to improve the robustness of detection algorithms, ensuring better generalisation and scalability in real-world scenarios.

In the realm of QC, the field of code smell detection is still evolving, and ongoing research is dedicated to its development. It's worth noting that the detection of QCS remains a relatively unexplored area, with limited existing studies. Recently, Openja et al. (2022) conducted a survey focusing on generic QCS in QPs, and Chen et al. (2023) presented a categorisation of eight types of QCS specific to quantum computers. They also introduced *QSmell*, a detection tool to detect smells in quantum systems.

Both studies contributed significantly by categorising and analysing QCS in quantum environments but lacked the integration of advanced DL detection algorithms capable of handling complex dependencies and quantum-specific challenges. While these studies have made significant contributions to the field, it is essential to note that neither of these works specifically addressed the detection algorithms that can be employed to detect QCS. This research presents a unique technique of using DL to address the issue of QCS in quantum computers and checks the feasibility of applying DL algorithms to detect QCS.

By incorporating hybrid models, TL, and advanced DL techniques, our work aims to fill the gaps left by previous research, offering a more scalable and robust approach to QCS detection in real-world quantum software environments.

## 4 Experimental framework

This section concisely explains the systematic approach employed for detecting QCS. The overview encompasses a sequence of essential steps undertaken to achieve this objective.

### 4.1 Overview

The initial phase involved the collection of 136 QPs from GitHub repositories. These programs were primarily built using the Qiskit libraries and were authored in the Python programming language. The next step included separating quantum circuits from the codes to create execution matrices from quantum codes.

Once the execution matrices were established, the QSmell detection tool was engaged. Chen et al. (2023) developed this tool to identify and detect eight distinct categories of QCS in QPs. This tool harnessed the matrices to determine the presence of QCS within the quantum circuits. This analysis created a manually annotated dataset, where code was categorised as

either smelly or non-smelly. Further, DL techniques are utilised by conducting an ablation study to determine the optimal hyperparameters and layers for the model, and a hybrid model is proposed. A detailed explanation of all the steps is in the subsequent subsections.

### 4.2 Dataset preparation

In the rapidly evolving landscape of technology, QC stands as a burgeoning field with limited project availability. To conduct this research effectively, high-quality quantum projects containing codes focused on constructing functional quantum circuits for real-world quantum computers were searched. Github repositories were explored, particularly those housing Qiskit, a prominent open-source software development kit (SDK) for QC.

Since quantum code is predominantly scripted in Python, the focus naturally turned toward GitHub repositories hosting Python-centric content. After conducting a comprehensive manual search of around 200 GitHub projects, the process involved eliminating duplicate and irrelevant projects, resulting in a curated dataset of 136 relevant projects for this study. Among these projects were essential components of quantum code related to significant gate operations and quantum measures, laying the groundwork for subsequent processes. The summary of the proposed methodology used in this study is shown in Fig. 6. Quantum source codes are manually selected and then processed by the qc2matrix algorithm to generate an execution matrix. This matrix serves as the input data for QSmell used in this study, from which code smell labels are extracted. The combination of the source code and these labels constitutes the dataset, which is further pre-processed and provided as input data to the DL models.

**Converting quantum code to execution matrix** QPs are relatively different from regular programs coded in languages like Java and C++. QPs use quantum gates to build quantum circuits. These circuits process data and give the results through qubit phases. Detecting problems in quantum code requires understanding how these gates are used in the source code.

QSmell, which works on the dynamic approach of QCS detection, is used to help find the QCS. For the input to the tool, an execution matrix is required to visualise the processes undergoing a quantum qubit. An example of conversion of a python file to an execution matrix using qc2matrix is shown in Fig. 7. In this figure, each row represents a quantum bit (qubit) or a classical bit (bit), and each column represents a step in the process. Each cell in the chart shows a quantum operation that happened at a specific time, giving a detailed picture of how the circuit works. All the selected repositories were first converted into an execution matrix, and their output was saved in processable format to be passed to the QSmell tool.

**Creating labeled dataset** Creating a labelled dataset began with acquiring QSmell from its primary GitHub repository. Subsequently, a step-by-step analysis of each quantum code's execution matrix was conducted, utilizing QSmell to compute the values of these five parameters. This meticulous analysis generated parameter values and facilitated code classification and the creation of the manual dataset, which would serve as input data for the DL models applied in the study. For this study, distinct threshold values for each of the five parameters were established. These thresholds served as decisive benchmarks, guiding code classification into either smelly or non-smelly code. The threshold values taken for these parameters are mentioned in (1) to (5) below:

$$LC(longCircuit) > 10 \tag{1}$$

$$ROC(RepeatedSetOfOperationsOnCircuit) > 0 \tag{2}$$

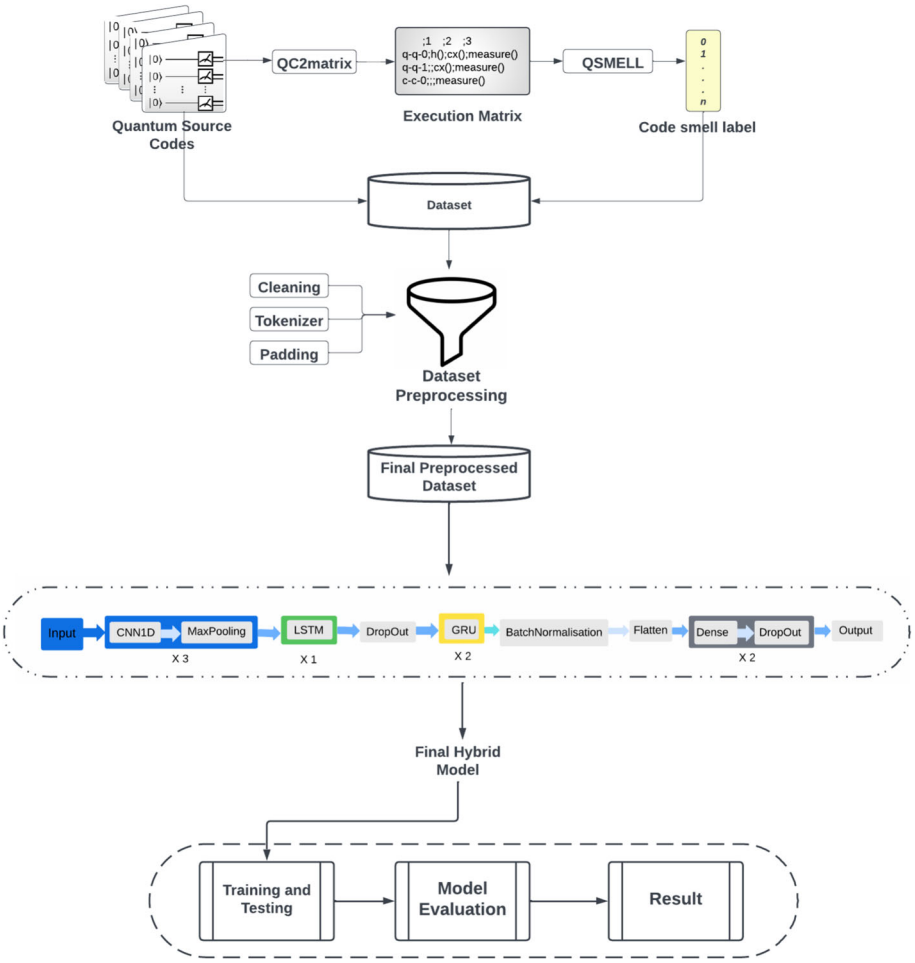



**Fig. 6** Overview of the Proposed Approach

$$IM(IntermediateMeasurements) > 0 \tag{3}$$

$$IdQ(IdleQubits) > 0 \tag{4}$$

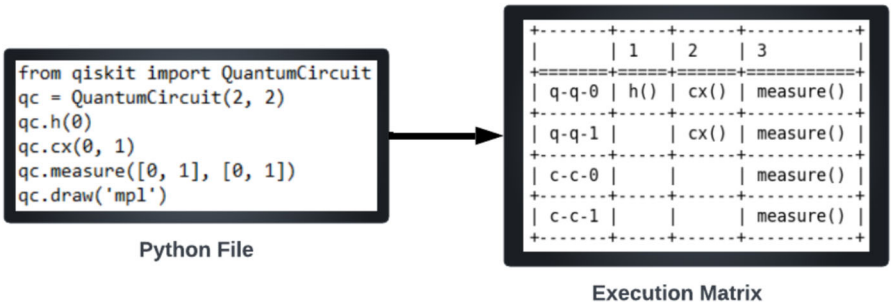



**Fig. 7** Converting Quantum Code to Execution Matrix

$$IQ(InitializationOfQubitsDifferentlyFrom|0\rangle) > 0 \tag{5}$$

If a quantum code satisfies one or more of the outlined criteria, it is classified as having a QCS. Conversely, a quantum code is considered non-smelly if it meets none of the provided criteria. To show the outcomes after QSmell detection, 13 sample projects are selected, and their outcomes are demonstrated in Fig. 8. This figure displays various values observed ranging across specific parameters within each QCS.

### 4.3 Dataset pre-processing

The dataset, formed in the preceding phase, undergoes pre-processing before being fed into the model for training. So, data pre-processing represents a pivotal stage in the overall process. The un-processed dataset has two columns: one with the raw Python scripts and the other with the binary rating of each code as smelly or not. Below are the steps needed to pre-process this dataset.

**Cleaning code** Pre-processing this dataset is essential, and the initial step involves code cleaning. This process begins by eliminating all comments embedded within the QPs as shown in (6) and (7). By doing so, the model exclusively assesses the unadulterated functioning code, enabling a focused evaluation.

$$\text{cleanedCode} = \text{remove}(\text{start}(\text{`\#'}), \text{end}(\text{`}endline\text{'})) \tag{6}$$

$$\text{cleanedCode} = \text{remove}(\text{start}(\text{```}), \text{end}(\text{```})) \tag{7}$$

After removing the comments, a fundamental cleansing process ensued, including eliminating stop words and punctuation. This pruning ensured that only pertinent code segments remained for further analysis.

**Tokenizer** Following the data cleansing process, the source code underwent tokenization, a pivotal step involving converting Python program text into numerical data, which is essential for model training. This approach is consistent with established practices in tokenization, ensuring that the same tokens consistently represent analogous words. This streamlined approach enhances the effectiveness of feature extraction from the data, aligning with the methodologies employed in prior research on tokenization techniques. This process ensures forming an accurate matrix that acts as input to the neural networks for classification purposes.

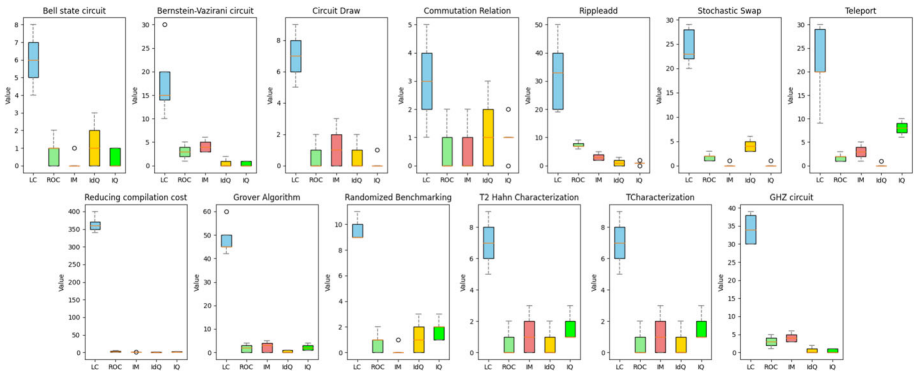



**Fig. 8** Values of QCS metrics for sample projects

**Padding** Given the diversity in code lengths across the Python files, the necessity arose for sequence padding to standardize dataset dimensions. This padding process was carried out seamlessly by padding to *max_length* following tokenization.

### 4.4 Analysis of DL models

In the previous sections of this study, the dataset has been carefully prepared and pre-processed for input into a DL model. The final dataset visual is shown in Fig. 9. This figure concisely represents the distribution of source code sizes and their corresponding frequencies. Now, the focus is on creating an efficient DL model for accurate predictions. To achieve this, an ablation study is conducted to find the best DL models and their optimal settings. This study considers three neural network architectures: 1-D CNN, LSTM, and GRU. CNN is taken to be 1-D CNN, which suits the structure of the code dataset. These models are chosen for their unique abilities in feature extraction and prediction. Additionally, these models exhibit coordinated features when combined. The goal is to combine these models to create a hybrid model using the best parameters from each analysis.

**Single unit ablation study** Individual CNN, LSTM, and GRU models are constructed during the ablation study. It started with a baseline configuration, systematically increased the number of layers, and changed hyper-parameters while carefully monitoring performance. If the model produces satisfactory results initially, layers are incrementally added until a decrease in performance is observed. Hyperparameters were fine-tuned at each step, exploring combinations to find the best settings.

**Hybrid layer ablation study** This thorough ablation study helped identify the most effective CNN, LSTM, and GRU models. These models demonstrated their effectiveness individually on the curated dataset, paving the way for the next step: constructing a hybrid model. In this

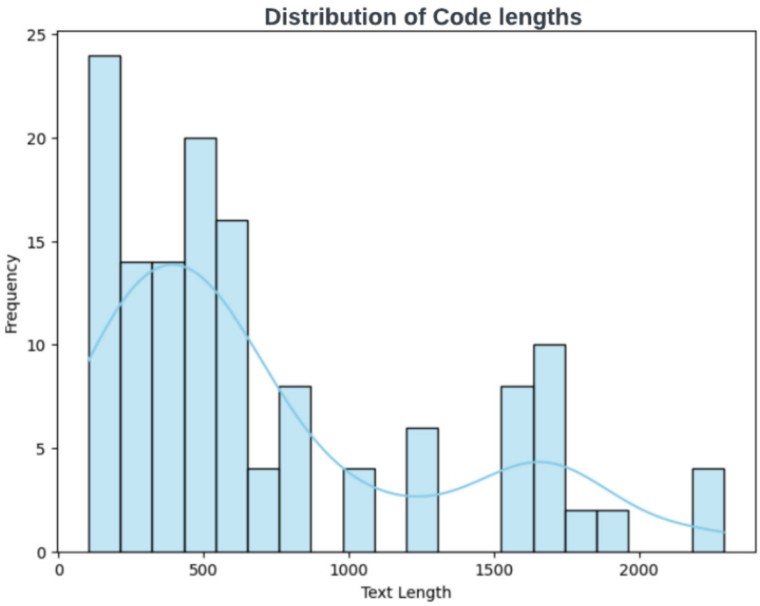



**Fig. 9** Analysis of Source Code Lengths

phase, two approaches were explored. First, the models were connected sequentially, one after the other, creating the "sequential hybrid model." Second, all individual models were run in parallel on the same dataset, and their outputs were merged to generate a single prediction, referred to as a "parallel hybrid model." Both hybrid models were evaluated independently on the curated dataset, and a detailed comparison of their results is presented in the Section 5. After conducting the analysis, the final "Proposed model" is formulated.

## 5 Analysis and results

This section provides concise responses to all the research inquiries posed in this study, focusing on illuminating the significant outcomes and findings highlighted as the major contributions of this study.

### 5.1 RQ1: how do QCS influence the frequency of errors in quantum computers?

In QC, noise emerges as a disruptive factor that compromises the stability of qubits, the fundamental units of quantum information. These disturbances lead to decoherence, wherein the qubit loses its delicate information over time. A variety of noise sources causes this effect; for example, IdQ introduces oscillations that cause qubit decoherence, compromising the integrity of encoded data. Even IQ contributes to decoherence by making qubits vulnerable to instability. Furthermore, the complexity of quantum algorithms with several gates can exacerbate mistakes via LC, resulting in deviations from the required quantum states. IM increases the measurement error, making the qubit unstable and collapse, and ROC can induce readout errors throughout the measurement process. These errors accumulate, leading to decoherence and changing the final results as shown in Fig. 10. This diagram shows that a qubit should remain consistent without any changes in an ideal scenario.

This is ideal as qubits are unstable and possess more than one or partial state at a time. When gates are worked upon the qubits, it tends to change the qubit's state. When gates are

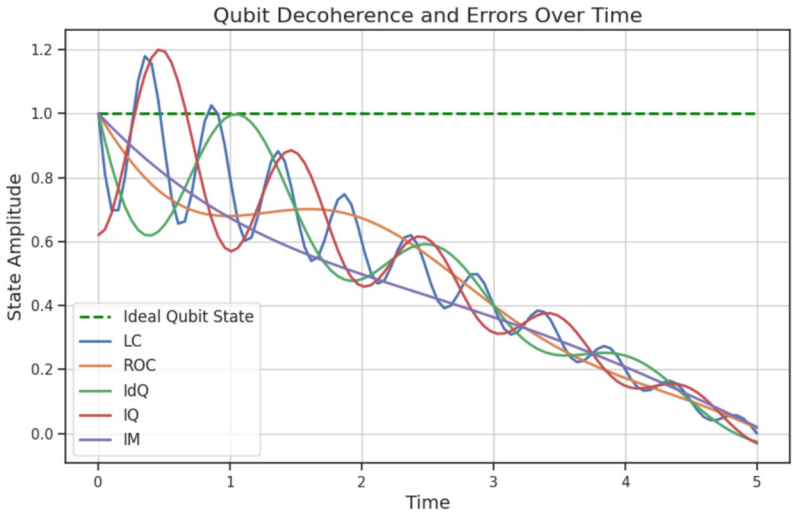



**Fig. 10** Error Produced due to Decoherence of Qubits

not applied to these qubits, they still change their state, which can hinder the calculation done in that QP. This adds to the errors in the calculation, which are impacted or increased by the induction of QCS in these qubits. When operating within a quantum environment, various errors and QCS can lead to qubit decoherence, causing it to revert to its stable state $|0\rangle$ gradually. The figure illustrates how different types of QCS impact the qubit's behaviour. As depicted in the diagram, the presence of various types of QCS affects the stability of qubits. It's important to note that each QCS has a distinct impact. For instance, IM leads to a gradual change in qubit behaviour with fewer frequent errors, whereas LC and IQ introduce more pronounced variations characterised by frequent changes in qubit states. 10

### 5.2 RQ2: what is the relative impact of different QCS on the performance of QC systems?

In QC, researchers are discovering how QCS in the code can make computer programs run slower and less reliably. Like classical computers, some code smells are more problematic than others, and their impact on the qubits varies with the threshold values. This is akin to how experts study computer problems in classical computers to improve programs.

In QC, it's crucial to identify which QCS have a more significant impact and which have a more minor impact. This is similar to understanding which issues in classical computers are more serious. By doing this, a focus can be placed on fixing the significant issues first, making QP work better and making them easier to manage. It is also important to note that various quantum computer types have unique characteristics. So, what may be a significant issue in one kind of quantum computer may be less effective in another.

By better understanding these issues, researchers can develop improved methods for detecting and correcting errors in quantum code. This will help quantum computers perform their tasks effectively and bring the research closer to using them for many exciting applications.

This research examined 136 projects and thoroughly recorded the occurrences of QCS within them. To provide insight into this section, a subset of 50 sample projects has been selected from the larger pool of 136 projects, and their data is presented for better comprehension. Table 2 describes the values of these 50 sample projects. The table provides the established threshold values for each QCS.

Among these QCS, the most critical one in QP is IdQ. It stands out as the most severe because it triggers decoherence, introducing a substantial number of errors. Conversely, IM is considered the least impactful QCS due to its comparatively minor influence on the quantum system's stability and accuracy.

### 5.3 RQ3: do the outcomes of DL models, developed with distinct hyper-parameters, exhibit complementary behavior?

In this RQ, various combinations of DL models have been evaluated by adjusting their architecture and hyperparameters. Specifically, three distinct types of individual models, CNN, LSTM, and GRU, have been considered. Each model has been assessed to determine the optimal number of layers that yield the best performance on the dataset.

**Single layer ablation study** Initially, each of the chosen models, namely CNN, LSTM, and GRU, has been evaluated independently to assess its efficacy in identifying QCS. The outcomes of the different CNN, LSTM, and GRU architectures are compared and illustrated in Table 3. It is evident from the results that the 1D CNN model with three layers consis-

**Table 2** Project-Wise Code Smell values

| Projects | QSmell | | | | |
|---|---|---|---|---|---|
| | LC | IdQ | IQ | IM | ROC |
| Bell State Circuit | 6 | 0 | 0 | 0 | 0 |
| Bernstein-Vazirani Circuit | 20 | 3 | 3 | 0 | 0 |
| Circuit Draw | 6 | 0 | 0 | 0 | 0 |
| Commutation Relation | 4 | 0 | 0 | 0 | 0 |
| Five-Qubit GHZ Circuit | 30 | 3 | 3 | 0 | 0 |
| GHZ | 4 | 0 | 0 | 0 | 0 |
| Initialize | 4 | 0 | 0 | 0 | 0 |
| Grover | 42 | 2 | 1 | 0 | 1 |
| Load Qasm | 24 | 0 | 0 | 0 | 0 |
| Rippleadd | 40 | 6 | 3 | 0 | 0 |
| Stochastic Swap | 20 | 1 | 0 | 1 | 0 |
| Teleport | 20 | 1 | 0 | 1 | 0 |
| Three-Qubit GHZ | 6 | 0 | 0 | 0 | 0 |
| Using Qiskit Terra Level | 6 | 0 | 0 | 0 | 0 |
| Quantum Circuit.draw | 2 | 0 | 0 | 0 | 0 |
| Deutsch-Jozsa Circuit | 12 | 1 | 1 | 0 | 0 |
| Quantum Fourier Transform | 18 | 2 | 2 | 0 | 0 |
| Superdense Coding | 10 | 1 | 0 | 0 | 0 |
| Swap Test | 6 | 0 | 0 | 0 | 0 |
| MLAE | 1 | 0 | 0 | 0 | 0 |
| Adapt VQE | 18 | 7 | 6 | 0 | 0 |
| AE | 4 | 0 | 0 | 1 | 0 |
| FAE | 4 | 1 | 1 | 0 | 0 |
| IAE | 3 | 0 | 0 | 0 | 0 |
| IPE | 24 | 0 | 0 | 0 | 3 |
| QAOA | 18 | 1 | 1 | 0 | 0 |
| QEOM | 30 | 3 | 0 | 0 | 0 |
| Multi-Controlled CNOT | 5 | 0 | 0 | 0 | 0 |
| Composite Gates | 6 | 0 | 0 | 0 | 0 |
| Parameterized Circuits | 30 | 3 | 0 | 0 | 0 |
| Reducing Compilation Cost | 360 | 1 | 1 | 0 | 6 |
| Composition | 8 | 0 | 0 | 0 | 0 |
| Mitigation | 3 | 0 | 0 | 0 | 0 |
| Mitigation1 | 4 | 0 | 0 | 0 | 0 |
| Mitigation2 | 4 | 0 | 0 | 0 | 0 |
| Mitigation3 | 11 | 0 | 0 | 0 | 0 |
| HHL | 2 | 0 | 0 | 0 | 0 |
| QGAN | 4 | 0 | 0 | 0 | 0 |
| QSVC | 4 | 0 | 0 | 0 | 0 |
| Shor | 19852 | 15650 | 15650 | 0 | 0 |
| $T\Phi$ Characterization | 3 | 0 | 0 | 0 | 0 |

**Table 2** continued

| Projects | QSmell | | | | |
|---|---|---|---|---|---|
| | LC | IdQ | IQ | IM | ROC |
| $T\Phi$ Characterization1 | 6 | 0 | 0 | 0 | 1 |
| $T\Phi$ Characterization2 | 5 | 0 | 0 | 0 | 0 |
| T2 Hahn Characterization | 6 | 0 | 0 | 0 | 1 |
| T2 Hahn Characterization1 | 4 | 0 | 0 | 0 | 0 |
| T2 Hahn Characterization2 | 6 | 0 | 0 | 0 | 1 |
| Randomized Benchmarking | 4 | 0 | 0 | 0 | 0 |
| Randomized Benchmarking1 | 7 | 0 | 0 | 0 | 0 |
| Randomized Benchmarking2 | 9 | 0 | 0 | 0 | 1 |
| T2 Ramsey Characterization | 5 | 0 | 0 | 0 | 0 |

tently outperforms the other configurations. This observation can be attributed to the doubled number of filters applied in each layer of this CNN architecture, resulting in highly efficient feature extraction. Consequently, among the CNN models evaluated, the 3-layer 1D CNN emerges as the most promising choice for achieving accurate results.

Additionally, the performance of various GRU configurations is considered. It is evident from the findings that the GRU model with a single layer demonstrated promising results. Consequently, an experiment was conducted by increasing the number of layers to two, each containing 128 units. The results indicated that as the layers increased to two, the performance remained comparable to that of the single-layer GRU, with only a marginal improvement.

However, upon further exploration with three layers, a notable decline in all evaluation metrics was observed. This resulted in the conclusion of the experiment after this stage. Therefore, the GRU model with two layers and 128 units in each layer emerged as the most suitable configuration among those tested.

Lastly, the outcomes for various LSTM configurations are measured. In the case of LSTM, the initial experiment with a single layer and 128 hidden units yielded impressive results. However, as the number of layers was increased, the performance began to deteriorate. An experiment combining LSTM with Bidirectional LSTM (Bi-LSTM) layers was also conducted to explore alternative architectures. Regrettably, the results obtained from the Bi-LSTM model did not meet the expectations. Consequently, based on this evaluation, it was concluded that the LSTM model with a single layer of 128 hidden units represents the most effective choice among the tested configurations.

**Hybrid layer ablation study** From the single-layer ablation study, the best layers in different DL algorithms were evaluated. This section will further test the best layers from the above results based on their combination working to detect QCS. For this evaluation, each configuration from CNN, LSTM, and GRU was coupled with one another to form different hybrid models. Keeping their feature extraction strength and model working enhancement with each other, three different combinations were selected for this evaluation. CNN+GRU, CNN+LSTM, and LSTM+Bi-LSTM were tested for the detection of QCS.

Table 4 describes the results of the following evalution. The CNN+LSTM combination achieved the highest accuracy at 89.84%, outperforming the other models. Additionally, it exhibited superior precision (90.12%), an F1-Score of 85%, and a Matthews Correlation Coefficient (MCC) value of 88%, making it a stable and compatible choice. The recall value, while slightly lower at 90%, was competitive and only marginally less than the highest value achieved by CNN+GRU at 91%.

**Table 3** Evaluation of Individual models

| Metric | CNN 1-D | | | LSTM | | | GRU | | |
|---|---|---|---|---|---|---|---|---|---|
| | One Layer | Two Layers | Three Layers | One Layer | Two- Layers | Two-Layer (Bi-LSTM) | One Layer | Two Layers | Three Layers |
| Accuracy | 82.1 | 78.0 | 85.7 | 91.9 | 86.0 | 61.0 | 89.0 | 89.3 | 86.0 |
| Precision | 68.0 | 69.0 | 76.9 | 93.0 | 90.0 | 37.0 | 92.0 | 92.0 | 90.0 |
| Recall | 100.0 | 81.0 | 90.9 | 93.0 | 86.0 | 61.0 | 89.0 | 89.0 | 86.0 |
| F1-Score | 81.5 | 75.0 | 83.0 | 93.0 | 87.0 | 46.0 | 92.0 | 92.0 | 90.0 |
| MCC | 69.7 | 71.7 | 73.8 | 85.0 | 74.0 | 19.0 | 80.4 | 80.4 | 74.9 |

**Table 4** Evaluation of Combination of Models

| Evaluation Metric | CNN+GRU | CNN+LSTM | LSTM+Bi-LSTM |
|---|---|---|---|
| Accuracy | 86.87 | 89.84 | 87.09 |
| Precision | 88.0 | 90.12 | 89.0 |
| Recall | 91.0 | 90.0 | 86.0 |
| F1-Score | 80.0 | 85.0 | 82.0 |
| MCC | 79.0 | 88.0 | 75.6 |

These evaluations highlight the superiority of the CNN+LSTM hybrid model combination. These results can be attributed to the strengths of CNNs in capturing local patterns and spatial features and their ability to extract hierarchical features through convolutional and pooling layers.

LSTM models excel at modelling sequential dependencies in data, which is particularly important for time series and sequential data. Combining these two types of information proved advantageous, allowing the model to consider spatial relationships and temporal dependencies simultaneously in QCS detection.

Considering the features mentioned above and the results from training the QCS dataset with different model combinations, it becomes clear that CNN+LSTM is the optimal choice for creating an improved hybrid model for QCS detection. The findings from this section will be extended to further analysis by incorporating additional compatible layers, enhancing the model's complexity in pursuit of higher accuracy.

**Proposed model** Based on the results from the previous section, it was evident that the combination of CNN and LSTM (CNN+LSTM) significantly improved QCS detection, achieving an accuracy of 89.84%. To build on this success, the next step involved sequentially adding a GRU layer to the CNN+LSTM hybrid model. This increased the model's complexity and enhanced its ability to capture long-range dependencies present in the dataset.

By combining CNN, LSTM, and GRU, the model exhibited an ensemble effect, with each component contributing its unique strengths and compensating for the weaknesses of the others. The goal was to explore different layer combinations to ultimately propose a model yielding the highest accuracy for QCS detection.

The detailed architecture of each layer in this proposed model is described in Table 5. It commences with a CNN as the initial network, featuring three layers for effective feature

**Table 5** Architectural Details of the Proposed Model

| Model | CNN 1D | GRU | LSTM |
|---|---|---|---|
| Layers | 3 | 2 | 1 |
| Architecture | Filters: 64, 128, 256<br>Kernel Size: 3 | Units: 128 | Hidden Units: 128 |
| Learning Rate | 0.001 | 0.001 | 0.0001 |
| Dropout | 0.5 | 0.2 | 0.5 |
| Accuracy | 85.71 | 89.28 | 91.86 |
| F1-Score | 83 | 86 | 93 |
| Recall | 90.9 | 89 | 93 |
| Precision | 76.9 | 92 | 93 |
| MCC | 73.76 | 80.43 | 85 |

extraction. This is followed by an LSTM layer for modelling and integrating the extracted features. Finally, a GRU layer is added to enhance the dependency factor. The model includes max-pooling, dense, flattened, and dropout layers to ensure input data consistency at each layer. This model achieved an impressive accuracy of 92.86%, precision of 100%, recall of 81.81%, F1-Score of 90%, and MCC value of 85.56%, which is the highest among all the tested combinations.

### 5.4 RQ4: what are the predictive capabilities of different DL models developed for QCS detection?

From the ablation study, the parameters and layers showing the best accuracies were selected for each DL model, including CNN, LSTM, and GRU. These models were then used to construct the "Proposed Model." The final configuration of the proposed model is shown in Table 5. Remarkably, these models demonstrated strong performance during their ablation studies. In this RQ, all the models are compared to determine the best-performing model in predictive capabilities.

To achieve this, all of these models underwent testing on the test split dataset. Their performance was assessed, and their respective ROCC was generated, a critical tool for this Comparison. ROCCs provide valuable insights into the models' discriminative power across various classification thresholds. The Comparison based on ROCs is shown in Fig. 11.

By examining these ROC curves, how well each model can distinguish between positive and negative samples can be visualized, revealing their varying degrees of discriminative ability. Additionally, the Area Under the ROCC (AUC) is a numerical metric for overall

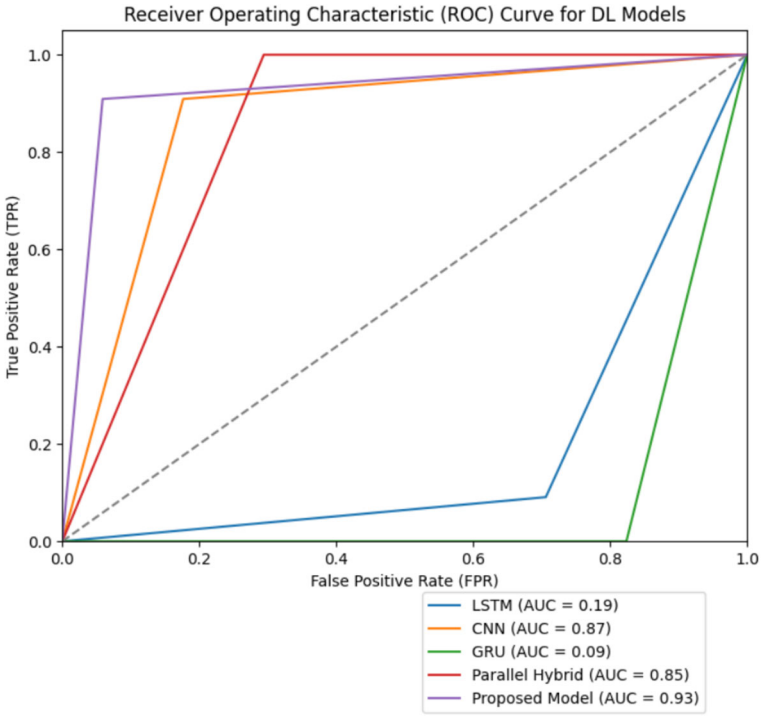



**Fig. 11** ROC curve comparison among DL models

model performance, with higher AUC values signifying better discriminatory power. It is observed from the figure that the proposed model shows the best AUC value as **0.93**.

### 5.5 RQ5: what is the effectiveness of the proposed model when compared with state-of-the-art models?

This RQ focuses on validating the novelty and credibility of the proposed model compared to other comparable and pre-trained models, including TL models, transformers, and LLMs. To accomplish this, a comparative analysis of each model's performance was undertaken, leveraging diverse performance metrics. Each of these metrics plays a crucial role in the evaluation process, providing a comprehensive assessment of the models under examination.

This analysis compared independent CNN, LSTM, and GRU models to various hybrid models. The hybrid models included a parallel hybrid model and a pre-trained BERT model. Additionally, the proposed model was evaluated alongside transformers and LLMs. These model results were compared with the proposed model introduced in this study.

The results are shown in Fig. 12. All the models were evaluated using five metrics - accuracy, precision, recall, F1-score, and MCC. The values from each model were assessed for an overall comparison.

Observations indicated that neither of the single-model structures, hybrid models, transformers, nor LLMs outperformed the proposed model. When comparing the proposed model to the single-layer models and other hybrid models, it became apparent that the complexity and enhancements in the proposed model significantly surpassed those of the individual models, LLMs, and transformers.

Similarly, a hybrid parallel model was created based on combining individual layers. In this setup, each layer was granted the autonomy to extract features from the data independently, and the results were subsequently integrated and trained on the dataset. Notably, unlike the proposed model combination, where the models are connected one after the other, the parallel

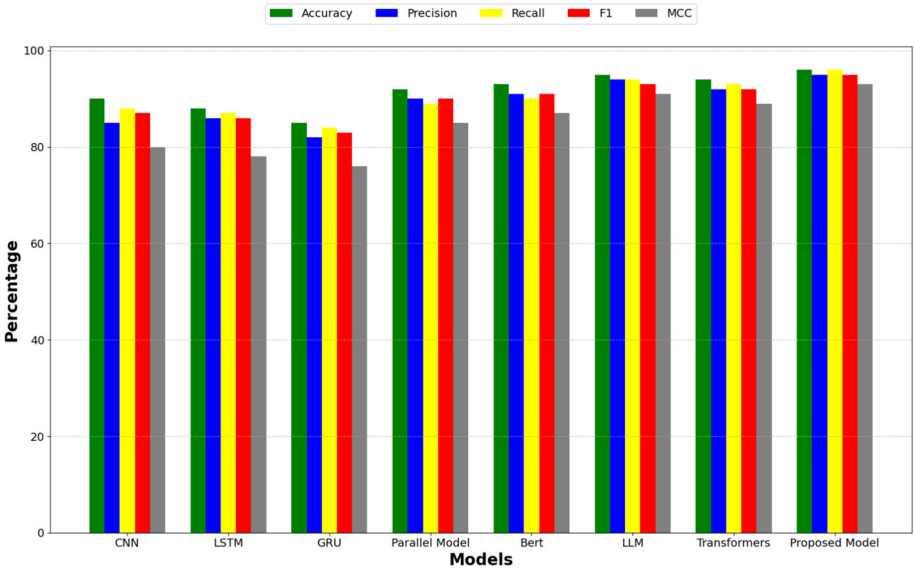



**Fig. 12** A Comparative Analysis of Results Between the Proposed Model and Other Models

architecture exhibited a distinct behaviour. In this case, the overall accuracy demonstrated a noticeable decrease, aligning with the findings observed across other result metrics.

To delve further into the details, it's worth noting that while the proposed model emphasizes a sequential flow of data, the parallel model enables each layer to work independently and parallelly. This shift in approach may have led to differences in feature extraction and how the models collectively interpret the data, ultimately impacting their overall performance.

These findings emphasize the significance of the architectural design choice, and they underline that the sequential and parallel model configurations can have varying impacts on the model's ability to understand and generalize from the dataset. The choice between these approaches should consider the data's specific characteristics and the analysis's goals.

A comparative analysis using TL was conducted to enhance the diversity of testing for the proposed model. Specifically, this study utilized a specialized pre-trained BERT model that had already undergone training for classical computer code smell detection, along with TL approaches involving LLMs and transformers. The pre-trained BERT model, along with these TL models, was fine-tuned and evaluated on a quantum computer dataset. The results obtained from these models were then compared to the proposed model. As depicted in the figure, it is evident that the BERT model outperformed the parallel model. Still, interestingly, the proposed model outshined all of these models, including LLMs and transformers, achieving an impressive accuracy rate of 92.86%. This suggests that the combination of feature extraction, sequential architecture, and specific input feature analysis within the proposed model contributed to significantly improved detection of QCS.

### 5.6 RQ6: how does the proposed hybrid QCS detection model outperform existing QCS detection tools?

The comparison between the proposed hybrid QCS detection model and existing tools (QChecker, QPAC, and QSmell) demonstrates significant advantages, particularly in computational cost, compatibility, and usability. These insights are visually represented in Fig. 13, which highlights the runtime performance of each technique across increasing problem sizes.

To evaluate and answer this research question, each of the techniques QChecker, QPAC, QSmell, and the proposed hybrid model was implemented on the same system under identical conditions. The models were tested using a large dataset of QP to measure runtime, assess compatibility with quantum frameworks, and determine computational cost. The figure presents the runtime performance of each tool, illustrating the superior efficiency of the proposed hybrid model.

The proposed hybrid model consistently outperformed the existing tools in terms of computational cost. While QChecker exhibited a fast linear complexity, its limited functionality and framework support restrict its applicability in diverse quantum development scenarios. The hybrid model, in contrast, maintained an efficient complexity, ensuring scalability even with large quantum codebases. This makes it more suitable for real-world scenarios where developers often deal with large-scale projects. QSmell, despite its robust detection capabilities, showed significantly higher computational costs due to its complex runtime operations, rendering it less practical for large datasets.

The figure also highlights the runtime differences between the tools as the problem size increases. The hybrid model demonstrated balanced performance, effectively combining the strengths of existing approaches. It processed larger quantum codebases with lower computational overhead compared to QPAC and QSmell. QPAC, while reasonably efficient, struggled with input format restrictions, leading to additional preprocessing time. The hybrid model's

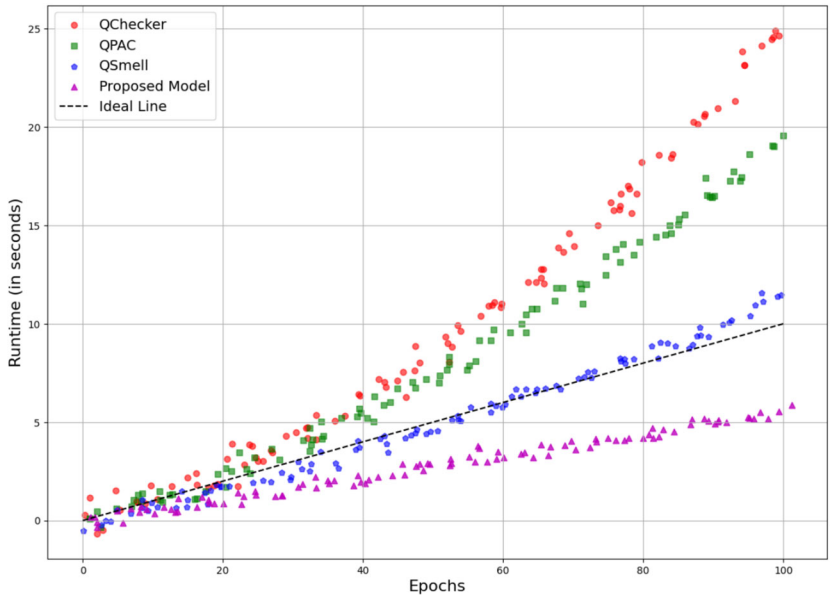



**Fig. 13** Runtime Comparison for different QCS Detecting Techniques

compatibility with widely accepted Python files minimized conversion overhead, streamlining the detection process and contributing to its lower runtime.

In terms of compatibility, the figure underscores the hybrid model's support for major quantum frameworks, including Qiskit, Q#, and Cirq. This comprehensive support contrasts sharply with QPAC, which lacks Q# integration, and QSmell, which does not work with Cirq. Such broad framework compatibility enhances the hybrid model's flexibility and usability, positioning it as the most practical tool for diverse quantum development environments.

Overall, the figure illustrates how the proposed hybrid model effectively balances computational efficiency, runtime scalability, and cross-framework compatibility. This combination ensures improved developer productivity and reduced computational costs while maintaining broad applicability. By addressing the limitations of existing tools, the hybrid model establishes itself as an advanced solution for detecting QCSs, tailored to meet the demands of modern quantum development workflows.

## 6 Threats to validity

This section delves into potential factors that could affect the validity of this study.

### 6.1 Internal threats to validity

One potential internal validity threat in this study is the selection bias that arises from the manual assessment and project selection process. Based on their judgment, the authors selected 136 datasets from Qiskit documentation and GitHub repositories, introducing subjectivity into the selection process. If this subjectivity affects the choice of datasets and projects, it could undermine the study's internal validity. Additionally, the use of specific models

and their parameters for dynamic analysis must be carefully justified and validated to avoid introducing bias and threats to internal validity.

### 6.2 External threats to validity

The external validity of this study faces several challenges. Firstly, the study's exclusive focus on the Qiskit ecosystem for project selection limits its domain generalizability. Findings may not extend well to other software development domains, diminishing external validity. Secondly, manual datasets specifically selected for Qiskit-related projects may not represent the broader universe of software projects, affecting the study's dataset generalizability. Moreover, the exclusive concentration on Python programming languages restricts external validity, as the findings may not apply to projects developed in different programming languages. Finally, the context in which the chosen projects exist, being exclusive to Qiskit, could lead to issues of context dependence, limiting the external validity.

### 6.3 Construct threats to validity

The study's construct validity may be threatened by the limited selection of smells described in the QSmell paper for detecting QCS. If these selected smells do not comprehensively represent all relevant QCS or have limitations in their measurement, a construct validity threat could be introduced. Additionally, the choice of specific models for dynamic analysis and their parameterisation may not fully capture the essence of QCS, potentially compromising construct validity.

In conclusion, this study is mindful of various factors that might impact its findings, such as selection bias in project choice, the specificity of models, limited domain and dataset generalizability, and potential construct validity issues related to smell selection and model suitability. By addressing these potential limitations and emphasising transparency in the research process, the study aims to establish a foundation for robust research outcomes while acknowledging the scope of its application within the Qiskit ecosystem and Python programming language. Researchers should be cautious when applying these findings to broader contexts, considering these validity threats.

## 7 Conclusion and future work

In alignment with conventional software development practices, QC faces challenges related to the emergence of code smells, posing threats to long-term software maintainability. This research leverages advanced DL techniques to identify these QCS. Employing QSmell, a dedicated tool for quantum software analysis, and assembling a manual dataset comprising 136 QPs, this study establishes the potential of DL in the context of QC.

Moreover, this paper introduces a thorough ablation study to craft the most precise DL model for QCS detection. Specifically, a sequential hybrid model is proposed, incorporating an individual CNN 1D model with three layers, an LSTM model with one layer, and a GRU model with two layers.

The proposed model exhibited an impressive accuracy of 92.86%, underscoring its efficacy as an innovative approach. Additionally, the research uncovered intricate relationships between specific QCS and the errors they induce, providing insights into how these errors can compromise the maintainability of quantum code.

Future work aims to enhance QCS detection by expanding the dataset to include a broader variety of quantum programming frameworks and paradigms, ensuring greater generalizability. The exploration of Generative AI for generating synthetic datasets presents a promising avenue to address dataset limitations. Additionally, integrating QSmell into quantum Integrated Development Environments (IDEs) and adapting it for real-time detection could significantly enhance its practical utility. Extending QSmell to classify QCS as a multi-label classifier instead of a binary classifier is another potential direction.

Another critical focus is optimizing the model for scalability to analyze large-scale QP effectively. Developing a user-friendly interface that provides actionable insights on QCS will further bridge the gap between theoretical advancements and practical implementation. By demonstrating the feasibility of DL for QCS detection in the QC domain, this study paves the way for further exploration while refining software quality assessment in this evolving field.

**Author Contributions** Marouane Kessentini contributed to the inception of the paper by proposing the initial idea. Ruchika Malhotra and Marouane Kessentini collaborated on the planning and conceptualization of the paper. Bhawna Jain played a significant role in crafting the written content of the paper. All authors actively participated in the paper review process.

**Funding** No funding was obtained for this study

**Data Availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing Interests** The authors declare no competing interests.

## References

Ali, S., Arcaini, P., Wang, X., Yue, T. (2021). Assessing the effectiveness of input and output coverage criteria for testing quantum programs, In: 2021 14th IEEE Conference on Software Testing, Verification and Validation (ICST), IEEE, pp. 13–23.

Alkharabsheh, K., Alawadi, S., Kebande, V. R., Crespo, Y., Fernández-Delgado, M., & Taboada, J. A. (2022). A comparison of machine learning algorithms on design smell detection using balanced and imbalanced dataset: A study of god class. *Information and Software Technology,143*, Article 106736.

Bafatakis, N., Boecker, N., Boon, W., Cabello Salazar, M., Krinke, J., Oznacar, G., White, R. (2019). Python coding style compliance on stack overflow, In: 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), pp. 210–214. https://doi.org/10.1109/MSR.2019.00042

Ball, H., Biercuk, M. J., Carvalho, A. R., Chen, J., Hush, M., De Castro, L. A., Li, L., Liebermann, P. J., Slatyer, H. J., Edmunds, C., et al. (2021). Software tools for quantum control: Improving quantum computer performance through noise and error suppression. *Quantum Science and Technology,6*(4), Article 044011.

Barbez, A., Khomh, F., Guéhéneuc, Y.-G. (2019). Deep learning anti-patterns from code metrics history, In: 2019 IEEE International Conference on Software Maintenance and Evolution (ICSME), IEEE, pp. 114–124.

Barends, R., Kelly, J., Megrant, A., Veitia, A., Sank, D., Jeffrey, E., White, T. C., Mutus, J., Fowler, A. G., Campbell, B., et al. (2014). Superconducting quantum circuits at the surface code threshold for fault tolerance. *Nature, 508*(7497), 500–503.

Benedetti, M., Realpe-Gómez, J., Biswas, R., & Perdomo-Ortiz, A. (2016). Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applicatio2011–2021p learning. *Physical Review A,94*(2), 022308

Bennett, C. H., Brassard, G. (2020). Quantum cryptography: Public key distribution and coin tossing, arXiv:2003.06557.

Bhave, A., Sinha, R. (2022). Deep multimodal architecture for detection of long parameter list and switch statements using distilbert, In: 2022 IEEE 22nd International Working Conference on Source Code Analysis and Manipulation (SCAM), pp. 116–120. https://doi.org/10.1109/SCAM55253.2022.00018

Chen, Q., Câmara, R., Campos, J., Souto, A., Ahmed, I. (2023). The smelly eight: An empirical study on the prevalence of code smells in quantum computing, In: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), IEEE, pp. 358–370.

Chen, Z., Chen, L., Ma, W., Xu, B. (2016). Detecting code smells in python programs, In: 2016 International Conference on Software Analysis, Testing and Evolution (SATE), pp. 18–23. https://doi.org/10.1109/SATE.2016.10

Das, A. K., Yadav, S., Dhal, S. (2019) Detecting code smells using deep learning, in: TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), pp. 2081–2086. https://doi.org/10.1109/TENCON.2019.8929628

Dewangan, S., Rao, R. S., Mishra, A., & Gupta, M. (2022). Code smell detection using ensemble machine learning algorithms. *Applied Sciences,12*(20), 10321. https://doi.org/10.3390/app122010321

Dewangan, S., Rao, R. S., Mishra, A., & Gupta, M. (2021). A novel approach for code smell detection: An empirical study. *IEEE Access, 9*, 162869–162883. https://doi.org/10.1109/ACCESS.2021.3133810

Di Nucci, D., Palomba, F., Tamburri, D. A., Serebrenik, A., De Lucia, A. (2018). Detecting code smells using machine learning techniques: Are we there yet?, In: 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER), pp. 612–621. https://doi.org/10.1109/SANER.2018.8330266

Feynman, R. P. (1982). Simulating physics with computers. *International Journal of Theoretical Physics, 21*(6/7), 467–488.

Fowler, M. (2002). Refactoring: Improving the design of existing code. In D. Wells & L. Williams (Eds.), *Extreme Programming and Agile Methods – XP/Agile Universe 2002* (pp. 256–256). Berlin Heidelberg, Berlin, Heidelberg: Springer.

Fowler, M., & Beck, K. (1997). Refactoring: Improving the design of existing code. *11th European Conference* (p. 6). Jyväskylä: Finland.

Gesi, J., Liu, S., Li, J., Ahmed, I., Nagappan, N., Lo, D., de Almeida, E. S., Kochhar, P. S., Bao, L. (2022). Code smells in machine learning systems. arXiv:2203.00803.

Guo, X., Shi, C., Jiang, H. (2019). Deep semantic-based feature envy identification, In: Proceedings of the 11th Asia-Pacific Symposium on Internetware, Internetware '19, Association for Computing Machinery, New York, NY, USA, p. 6. https://doi.org/10.1145/3361242.3361257

Gupta, R., Kumar Singh, S. (2022). A novel metric based detection of temporary field code smell and its empirical analysis, *Journal of King Saud University - Computer and Information Sciences34* (10, Part B) 9478–9500. https://doi.org/10.1016/j.jksuci.2021.11.005https://www.sciencedirect.com/science/article/pii/S1319157821003050

Gupta, H., Kulkarni, T. G., Kumar, L., Neti, L. B. M., Krishna, A. (2021). An empirical study on predictability of software code smell using deep learning models, In: Advanced Information Networking and Applications, Springer International Publishing, pp. 120–132. https://doi.org/10.1007/978-3-030-75075-6_10.

Huang, Y., Martonosi, M. (2019). Statistical assertions for validating patterns and finding bugs in quantum programs, In: Proceedings of the 46th International Symposium on Computer Architecture, pp. 541–553.

I. Ahmed, C. Brindescu, U. A. Mannan, C. Jensen, A. Sarma, An empirical examination of the relationship between code smells and merge conflicts, in. (2017). ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM). *IEEE, 2017*, 58–67.

Jebnoun, H., Ben Braiek, H., Rahman, M. M., Khomh, F. (2020). The scent of deep learning code: An empirical study, In: Proceedings of the 17th International Conference on Mining Software Repositories, pp. 420–430.

Lewowski, T., Madeyski, L. (2022). Code Smells Detection Using Artificial Intelligence Techniques: A Business-Driven Systematic Review, *Springer International Publishing*, Cham, 2022, Ch. SSDC, *377*, pp. 285–319. https://doi.org/10.1007/978-3-030-77916-0_12

Leymann, F., Barzen, J., Falkenthal, M., Vietz, D., Weder, B., Wild, K. (2020). Quantum in the cloud: application potentials and research opportunities, arXiv:2003.06256.

Li, G., Zhou, L., Yu, N., Ding, Y., Ying, M., Xie, Y. (2020). Projection-based runtime assertions for testing and debugging quantum programs, *Proceedings of the ACM on Programming Languages4* (OOPSLA) 1–29.

Liu, H., Xu, Z., Zou, Y. (2018). Deep learning based feature envy detection, In: 2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 385–396. https://doi.org/10.1145/3238147.3238166

Ma, W., Yu, Y., Ruan, X., Cai, B. (2023). Pre-trained model based feature envy detection, In: 2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR), pp. 430–440. https://doi.org/10.1109/MSR59073.2023.00065

Muralidhar, N., Muthiah, S., Butler, P., Jain, M., Yu, Y., Burne, K., Li, W., Jones, D., Arunachalam, P., McCormick, H. S., Ramakrishnan, N. (2021). Using antipatterns to avoid mlops mistakes. arXiv:2107.00079.

Nayak, P. K., Kher, K. V., Chandra, M. B., Rao, M. V. P., Zhang, L. (2023). Q-pac: Automated detection of quantum bug-fix patterns. arXiv:2311.17705 .

Omari, S., Martinez, G. (2020). Enabling empirical research: A corpus of large-scale python systems, In: K. Arai, R. Bhatia, S. Kapoor (Eds.), Proceedings of the Future Technologies Conference (FTC) 2019, Springer International Publishing, Cham, pp. 661–669.

Openja, M., Morovati, M. M., An, L., Khomh, F., & Abidi, M. (2022). Technical debts and faults in open-source quantum software systems: An empirical study. *Journal of Systems and Software,193*, Article 111458.

Pecorelli, F., Palomba, F., Di Nucci, D., De Lucia, A. (2019). Comparing heuristic and machine learning approaches for metric-based code smell detection, In: 2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC), IEEE, pp. 93–104.

Saheb-Nassagh, R., Ashtiani, M., Minaei-Bidgoli, B. (2022). A probabilistic-based approach for automatic identification and refactoring of software code smells, *Applied Soft Computing130* 109658. https://doi.org/10.1016/j.asoc.2022.109658https://www.sciencedirect.com/science/article/pii/S1568494622007074

Sandouka, R., & Aljamaan, H. (2023). Python code smells detection using conventional machine learning models. *PeerJ Computer Science,9*, Article e1370.

Sharma, T., Efstathiou, V., Louridas, P., Spinellis, D. (2021) Code smell detection by deep direct-learning and transfer-learning, *Journal of Systems and Software176* 110936. https://doi.org/10.1016/j.jss.2021.110936https://www.sciencedirect.com/science/article/pii/S0164121221000339

Shaydulin, R., Thomas, C., Rodeghero, P. (2020). Making quantum computing open: Lessons from open source projects, In: Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops, pp. 451–455.

Shor, P. W. (1997). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing, 26*(5), 1484–1509. https://doi.org/10.1137/s0097539795293172

Tarwani, S., Chug, A. (2022). Application of deep learning models for code smell prediction, In: 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1–5. https://doi.org/10.1109/ICRITO56286.2022.9965048

Tufano, M., Palomba, F., Bavota, G., Oliveto, R., Di Penta, M., De Lucia, A., & Poshyvanyk, D. (2017). When and why your code starts to smell bad (and whether the smells go away). *IEEE Transactions on Software Engineering, 43*(11), 1063–1088.

Van Oort, B., Cruz, L., Aniche, M., Van Deursen, A. (2021). The prevalence of code smells in machine learning projects, In,. IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN). *IEEE, 2021*, 1–8.

Yamashita, A., Moonen, L. (2013). Exploring the impact of inter-smell relations on software maintainability: An empirical study, In: 2013 35th International Conference on Software Engineering (ICSE), IEEE, pp. 682–691.

Zhang, Y., Ge, C., Hong, S., Tian, R., Dong, C., Liu, J. (2022) Delesmell: Code smell detection based on deep learning and latent semantic analysis, *Knowledge-Based Systems255* 109737. https://doi.org/10.1016/j.knosys.2022.109737https://www.sciencedirect.com/science/article/pii/S0950705122008796

Zhao, J. (2020). Quantum software engineering: Landscapes and horizons, arXiv:2007.07047.

Zhao, P., Wu, X., Li, Z., Zhao, J. (2023). Qchecker: Detecting bugs in quantum programs via static analysis. arXiv:2304.04387 .

Zhao, P., Zhao, J., Ma, L. (2021). Identifying bug patterns in quantum programs, In,. IEEE/ACM 2nd International Workshop on Quantum Software Engineering (Q-SE). *IEEE, 2021*, 16–21.

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

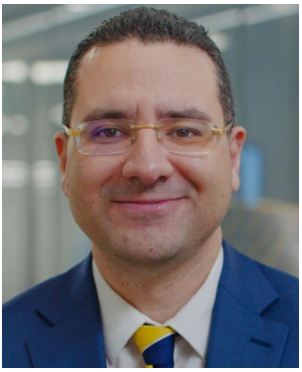
Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing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



**Dr. Ruchika Malhotra** is a Professor and Head of Department at the Department of Software Engineering, Delhi Technological University, Delhi, India. She was awarded the prestigious Raman Fellowship for pursuing Postdoctoral research at Indiana University Purdue University Indianapolis, USA. She received the IBM Faculty Award in 2013. Her h-index is 40, as reported by Google Scholar. She is the author of a book titled "Empirical Research in Software Engineering" published by CRC Press and co-author of a book on "Object-Oriented Software Engineering" published by PHI Learning. Her research interests are software testing, improving software quality, statistical and adaptive prediction models, software metrics, and the definition and validation of software metrics. She has published more than 220 research papers in international journals and conferences and is amongst the world's top 2% scientists as per reports from Stanford University.



**Ms. Bhawna Jain** is pursuing a Ph.D. in Computer Science under joint supervision from Delhi Technological University (DTU)- University of Michigan Flint, USA. She is an Assistant Professor in the Department of Computer Science and Engineering at Indira Gandhi Delhi Technical University for Women (IGDTUW). She has over three years of experience working as a Software Developer in the industry. She has won awards like All Rounder Performer and ACE Excellence award for being Growth Catalyst and received appreciation for her remarkable efforts in the R&D department during her industry experience. She also contributed to the Guinness Book of World Records for the "Largest online video chain of people handing over an object." She has research publications with IEEE and Springer. Her areas of interest are Artificial Intelligence, Deep Learning, Software Maintenance, and Generative AI.



**Dr. Marouane Kessentini** is working as Dean of the College of Comp- Her h-index is 40, as reported by Google Scholar. She is the author of uting at the Grand Valley State University. Earlier, he worked as Myron and Margaret Winegarden Professor, AssociateDean at the College of Innovation and Technology, The University of Michigan Flint. He is the recipient of the prestigious 2018 President of Tunisia Distinguished Research Award, the University Distinguished Teaching Award, the University Distinguished Digital Education Award, the College of Engineering and Computer Science Distinguished Research Award, four Best Paper awards, including an IEEE 10-Year Most Influential Paper award (2011-2021), and his AI-based software refactoring invention, licensed and deployed by several Fortune 500 companies, and selected as one of the Top 8 inventions at the University of Michigan for 2018 (including the three campuses) by the UM Technology Transfer Office. He received grants from industry and federal agencies and published over 175 papers in top journals and conferences. His h-index is 50, as reported by Google Scholar.

# Quantum Edge Cloud Computing: Revolutionizing IoT

Aanya Tiwari
*dept. of CSE*
*Delhi Technological University*
Delhi, India
aanyat2003@gmail.com

Ahsaan Ahmad Ahanger
*dept. of CSE*
*Delhi Technological University*
Delhi, India
ahsaanahmad2020@gmail.com

Anjali Meena
*dept. of CSE*
*Delhi Technological University*
Delhi, India
anjalimeen.meena@gmail.com

Pawan Singh Mehra
*dept. of CSE*
*Delhi Technological University*
Delhi, India
pawansinghmehra@gmail.com

2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT) | 979-8-3315-4335-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/DICCT64131.2025.10986380

***Abstract*— The augmented growth of the Internet of things (IoT) has brought an enormous surge in data generation requiring highly efficient, secure and scalable computational architectures. While conventional cloud computing is a widely used solution, it struggles with various issues like bandwidth constraints, latency and security vulnerabilities. Edge computing alleviates some of these concerns by processing the data closer to its source but time and again it is limited by restricted computational power. Quantum Edge Cloud Computing (QECC) presents a groundbreaking approach by assimilating the power of quantum computing with edge and cloud infrastructure to improve IoT performance. This study caverns the synergy between quantum computing, cloud computing and edge computing, foregrounding how this synergy can strengthen the IoT ecosystems. It examines the advantage of quantum assisted edge-cloud frameworks, investigates the advancements in cyber security through cryptographic methods and discusses plausible real-world applications. Additionally, the paper also features existing technological challenges, traces future research directions and suggests strategies required to realize a truly secure, intelligent and adaptive computing framework for IoT.**

***Keywords*— *Edge Computing, Quantum Computing, Cloud Technology, Quantum Edge Cloud Computing, Quantum Processors***

## I. INTRODUCTION

The galloping acceptance of IoT devices has led to a remarkable surge in data generation, prompting numerous sectors such as smart cities, industrial automation, healthcare, and autonomous technologies. However, traditional cloud computing infrastructures struggle to handle the enormous processing and security requirements of IoT, often encountering problems related to network congestion, high latency and cyber vulnerabilities. Edge computing has arisen as a viable alternative to surmount these limitations, enabling computation closer to its point of origin. While this approach reduces the delays and alleviates network congestion Edge computing has arisen as a viable alternative to tackle these limitations, enabling computation closer to its point of origin. Nonetheless, edge computing devices frequently contended with limited processing capabilities, restricting their ability to handle intricate operations and large-scale IoT implementations.

By assimilating quantum computing with edge and cloud architecture, a framework called Quantum Edge Cloud Computing (QECC) which extends a solution that substantially enhances data processing speed, strengthens cybersecurity measures and improves real-time decision making for IoT applications. This explores the transformative impact of QECC on IoT, discussing its key features, potential applications, existing challenges and potential directions for future research and implementation.

### *A. Motivation*

As the number of machines connected to IoT are soaring, facts are created aggressively, real time processing is needed, low latency and high computational abilities are created. The ordinary cloud computing is not able to cater the request effectively, mainly for the processes which require high security, optimization and large-scale data analytics. Quantum computing helps to solve the problem by processing vast data in parallel, but is not always feasible due to its difficult physical implementation.

Combining the Quantum Computing with edge computing increases the calculation power, reduces the time to respond and makes resource management efficient in IoT networks. QECC brings the computation closer to the source enabling secure and efficient data processing for even otherwise complex problems. Despite the benefits, QECC also has limitations of its own kind like scalability and hardware stability. This paper is motivated by the need to explore how Quantum Edge Cloud Computing overcomes these barriers to revolutionize IoT and unlock new possibilities.

### *B. Contribution*

This research paper lays down the theoretical foundations of Quantum Edge Cloud Computing (QECC), while critically exploring its potential to revolutionize IoT. Specifically, it contributes the following:

It explains the basic concepts of quantum cloud computing & edge computing, thus laying down the groundwork for Quantum Edge Cloud Computing.

It reviews the strengths & weaknesses of current developments & advancements made by researchers in this field.

It critically examines the role of edge and cloud computing in IoT applications, while highlighting their key differences.

It investigates the key components of the architecture, illustrates typical IoT applications benefitting from quantum edge cloud and evaluates their advantages & limitations.

Finally, it discusses various ethical & privacy concerns & outlines future possible research directions in this field.

979-8-3315-4335-8/25/$31.00 ©2025 IEEE



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:57:49 UTC from IEEE Xplore. Restrictions apply.

### C. Organization

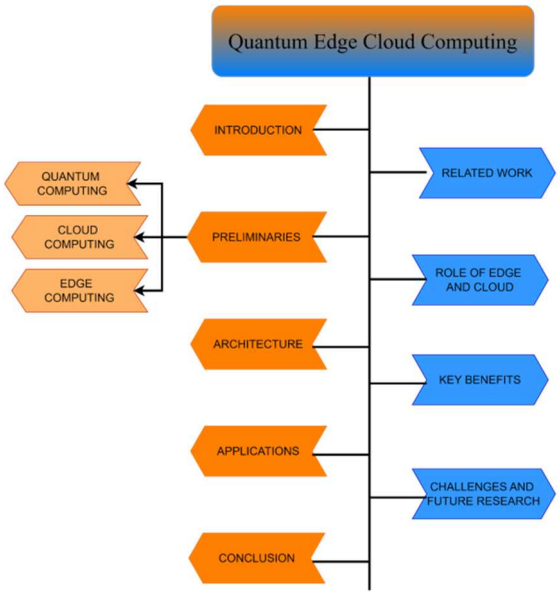



Fig. 1. Graphical Structure of Organization

## II. Related Work

The rapidly developing aspect of IoT urges calculational patterns that are able to maneuver large scale information, sustain upgraded safety, and cater contemporary proceeding competence [1]. QECC consolidates cloud, edge, and quantum computing, tackling the matter by merging the pros of each one to assure the evolving complicacy of up-to-date IoT arrangements [2]. Flourishing elastic, safe, and fruitful IoT executions which calibrates according to the expanding requirements of facts supervision and electronic affinity requisites this coalescing. A paramount prototype switch in how the cloud domains maintain the facts refining and safety [3]. The Quantum-Cloud concept was put forward by Singh and Sachdev [4], they reviewed the fusing of cloud technology with quantum mechanics. The noteworthy endowment by Quantum computing to boost proceeding rate and safety programme has been additionally explained by fresh reconstructions [5]. In contrary to regular calculating procedures, a survey suggests that the new computing schema conserved the time to 40% for complex algorithms [6].

Edge Computing is vitally important to decrease the response time and enhance the functioning of IoT arrangement [7]. An authentic edge computing survey about smart city traffic flow was accomplished in 2022, allowing it to leverage the power of artificial intelligence (AI) algorithms [8]. It was realised that edge computing declined the response time and improved the productivity of IoT organisations in municipal settings. As cloud computing provides vast repository space and enumerative assets, it presents as the bedrock for innovative and flexible IoT implementations [9]. Rao et al. sketched how by operating scalable results cloud platforms reinforce the comprehensive IoT machines [10]. The potential of cloud to vigorously expand at the time of pinnacle loads in an IoT stationing for a domestic retail chain was observed in a new pilot program to decode flexibility [11].

Cloud computing's flexibility solutions, as explored by Rao et al. [10], and edge computing's constructive principles, as stressed by Wu [12], are crucial for minimizing latency and enhancing the resilience of IoT web. In the opinion of Ren et al. [13], the exploration of internet hypothesis, explicitly graph theory and control theory, recommends pivotal intuition for self-operating and perfecting IoT web topologies, promising productivity, and enhanced balance.

TABLE I. Studies in Edge & Cloud Computing

| Research References | Strengths | Weakness |
|---|---|---|
| Wang et al [14] | Edge Computing | Identity authentication, data security |
| Khan et al [15] | Edge computing, Identity Authentication, | Data security |
| Ghosal et al [16] | Identity Authentication | Edge Computing, data security |
| Xiao et al [17] | Edge Computing, security | Applicability discussion, identity verification |
| Zhang et al [18] | Edge computing, focus on overall security and data security | Identity authentication |
| Li et al [19] | Edge computing, identity authentication | Overall security and data security |

## III. Preliminaries

### A. Quantum Computing

Qubits, the rudimentary unit of quantum facts, are the base of the whole concept of quantum computing. Quantum bits can remain in a situation of superposition, meaning they can coexist as a fusion of the situations 0 and one, unlike ordinary bits, which can be in 1 situation only [20]. In contrast to classical bits, this quality of quantum bits allows the illustration and operation of a larger set of knowledge opportunities. Entanglement a quantum circumstance in which the behavior of one quantum bit becomes related to the behavior of other quantum bit, even when took distances apart, is one more important aspect of quantum bits.

### B. Quantum Cloud Computing

End-users' plea to retrieve computing facilities by cloud computing. Taking advantage of quantum mechanics, QCC came out as an approach that can work out complicated issues that classical computers cannot manipulate in a sensible time. QCC provides users easy way in to assets for quantum calculation [21]. Remarkable struggles appear as a physical quantum computer requires specified surroundings like complicated control structures and endmost cold degrees. Thus, Using the quantum computers held within cloud computing systems is a feasible solution to this issue.

### C. Quantum Edge Cloud Computing

Edge Computing is the process in which the computing takes place at the network's edge. As mentioned by the U.S. professor Satyanarayanan of Carnegie Mellon University, it is a latest template that arranges storage assets and processing at the edge of the network near to end users or



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:57:49 UTC from IEEE Xplore. Restrictions apply.

sensors" [22]. Zha et al. [23] advocated that closer to the the origin of the facts, an open platform that intermixes intrinsic skills like networking, computing, storage, and applications and gives edge intelligent services closer to cater the requirements of industry, real-time business, data optimization, application intelligence, security, and privacy.

## IV. Role of Edge and Cloud Computing in IoT

Before the advancement of edge computing, ordinary cloud computing addressed the underlying issues by simply dispatching the whole information to the cloud computing centre. Cloud computing has evolved steadily in the modern era. The technologies like network storage, virtualisation, load balancing, parallel computing and distributed computing are incorporated in this network service platform. However, IoT acquired grip and became further ubiquitous in individual's lives, to a great degree machines are connecting to the network, and a vast number of facts is generated. Cloud computing's network capacity hasn't been capable to maintain itself with real-time performance and time-sensitive systems [24]. Therefore, cloud computing has drawbacks in data security and privacy protection, real-time, load, transmission bandwidth, energy consumption [25].

Cloud computing will follow a weighty load if the gigantic amount of information created by the associated machines are shared to the cloud. In the present age, edge computing is required to lessen the load of the cloud and supervise chores that come under its domain. various Internet services such as data mining and sharing, need the amalgamation of cloud computing with edge computing as several facts should be shared to the cloud for more proceedings after being processed by edge computing. The IoT network's connected machines become extra stable by these improvements.

Edge computing leverages concurrent aspects and, when fused with cloud computing and synergy, improves manufacturing productivity as well as rapidly figures out instrument defects. To empower the control between nodes with cloud and vice versa, the nodes compute heterogeneous information from different machines and share with the cloud for proper processing. Cloud computing and edge computing both have edge in meeting the demands of IoT machines and only by functioning together can the two continue to advance the network.

As edge computing is an add-on of cloud computing, it matches various qualities with cloud computing. Cloud computing's main attributes are its skill to comprehend the whole, process vast amounts of facts, perform rigorous analysis, and play a significant part in normal facts proceeding. With its local emphasis, edge computing can be more useful for modest, real-time intelligent analysis, including addressing local firms' urgent demands.

TABLE II. Key Differences between Edge and Cloud Computing

| Characteristics | Cloud Computing | Edge Computing |
|---|---|---|
| Application | Global | Local |
| Data Processing Location | Close to centre | Takes place at the edge |
| Latency | High | Low |
| Distance from user | More | Less |
| Storage | Large centralised storage | Decentralised storage |
| Bandwidth Requirement | High | Low |
| Security | Data encryption and access controls | Better security by encryption |
| Geo Distribution | Centralised | Decentralised |
| Scalability | Limited due to server size | High scalability options |
| Efficiency | Less | More |

## V. Quantum Architecture

Leveraging the power of Quantum Computing, QECC enhances security, computational speed, decision making capabilities; thereby optimizing IoT networks. The architecture is divided into three inter-related layers as discussed:

*Quantum-Enabled Edge Nodes*: By incorporating Quantum Key Distribution (QKD), these nodes prevent unauthorized data access by establishing highly secure communication channels. For time sensitive IoT applications like industrial and autonomous vehicles, these nodes are responsible. Decision making, traffic management, resource allocation are enhanced through the integration of advanced quantum algorithms like Quantum Approximate Optimization Algorithms (QAOA).

*Quantum-Assisted Cloud Servers*: These servers effectively collect and process massive amounts of data generated by IoT devices across domains such as healthcare, industrial manufacturing, etc. These servers support predictive analytics using Quantum Machine Learning (QML), finding applications in medical diagnostics, energy management, etc. To defend against cyber threats from quantum capable adversaries, advanced encryption protocols resistant to quantum attacks are implemented which ensures the security of data stored in the cloud.

*Secure Quantum Communication Channels*: Safe and reliable data transmission between IoT devices, edge nodes and cloud servers is ensured through a dedicated quantum-secure network backbone which uses quantum cryptographic techniques including Quantum-secure authentication and encryption, Quantum Key Distribution (QKD)[26].

## VI. Key Benefits of QECC in IoT

*Real-time analysis & reduced latency*: QECC mitigates the limitations associated with traditional cloud computing (high data volume & high pressure of network bandwidth) by storing and processing the data close to the source, ensuring real-time processing and reducing end-to-end latency. This feature is instrumental for applications including autonomous driving, multimedia gaming, industrial monitoring where rapid feedback is critical.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:57:49 UTC from IEEE Xplore. Restrictions apply.

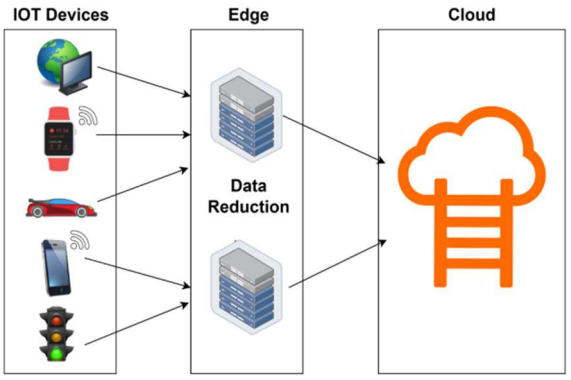



Fig. 2. Architecture of QECC

## VII. Key Benefits of QECC in IoT

*Real-time analysis & reduced latency*: QECC mitigates the limitations associated with traditional cloud computing (high data volume & high pressure of network bandwidth) by storing and processing the data close to the source, ensuring real-time processing and reducing end-to-end latency. This feature is instrumental for applications including autonomous driving, multimedia gaming, industrial monitoring where rapid feedback is critical.

*Security*: Traditional cloud computing relies on a centralized processing model, where all data must be uploaded to the cloud prior to processing. This approach increases the risk of exposing sensitive information such as user passwords and search records, thus compromising privacy and security. In contrast, QECC employs a distributed computing paradigm and leverages quantum encryption, eliminating the need for cloud uploads and significantly reducing the risks associated with data exposure.

*Low bandwidth cost & energy consumption*: Through computing offloading & quantum speed- up, the massive volume of data generated by IoT devices can be stored and pre-processed by the local edge cloud and only a small volume of processed data is required to be sent back to central data-centres, reducing bandwidth usage & energy usage.

*Scalability*: As QECC supports decentralized processing at the edge, it enables more devices to be connected without overburdening centralized servers.

## VII. Typical IoT Application Benefitting from Quantum Edge Cloud

*Smart Grid*: Combining modern information technology with conventional electricity grid, smart grid transmits the data of advanced monitoring applications like smart meters, sensors, transformers deployed within the system along with current, in real-time. By integrating flows of information with power and business, it facilitates secure, reliable, and efficient power supply. While cloud computing, through data processing and storage, guarantees efficiency and edge computing devices within the grid (IoT sensors & smart meters) process information locally closer to the source, instead of sending large volumes of data to the central server, facilitate immediate decision making. Real-time power data analysis at the edge, plays an instrumental role in load balancing through identification of peak demand periods and reducing power usage, power distribution, prediction of potential failures and integration of renewable sources of energy.

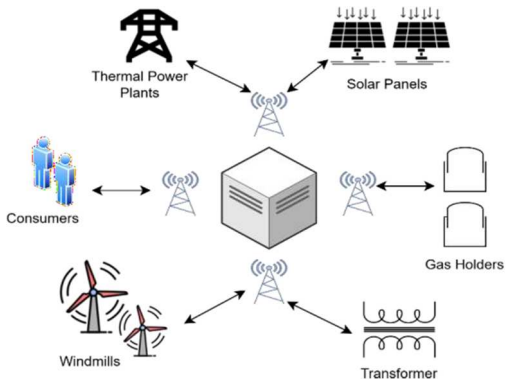



Fig.3. Smart Grid: The edge devices facilitate smooth frequency control & power distribution in the grid in real-time, mitigating power factor penalties.

*Smart Home Automation*: It is a low-cost and flexible system which accesses, manages, and remotely monitors commonly used home appliances & devices (lights, refrigerators, fans, air conditioners, ovens, washers, smoke and fire detectors, monitoring cameras, safety alarms) by establishing IP connectivity. Without changing or replacing the current household appliances, several sensors are integrated with these devices through a control panel which receives various measurements like temperature, humidity, light intensity, thus operating the devices according to the home occupants' requirements. Integrating quantum edge cloud computing significantly minimizes latency issues, optimizes energy utilization, provides location awareness and mobility support. Applications: smart cameras, smart lighting, smart windows, doors, automated safety alarms[27].

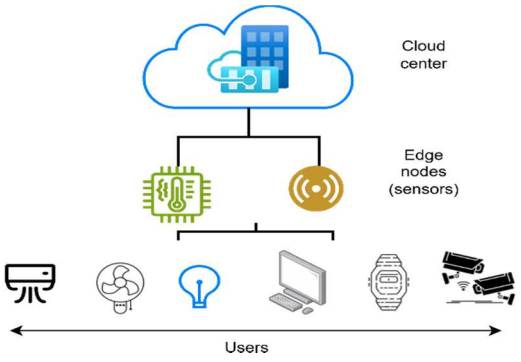



Fig.4. Smart Home Automation: Smart meters, IoT-enabled sensors track data, which is analyzed in real-time, facilitating immediate action.

*Smart Transportation:* This system is based on the application of IT and communication technology to vehicle manufacturing and service control, thereby facilitating safer and more efficient transportation. The reliability of this system depends critically on how accurately and precisely sensors detect objects and navigate under varying environmental conditions; and the analysis and timely release of this information to nearby vehicles. Quantum sensors transcend the sensitivity and accuracy limits of classical sensors, by detecting smaller and fainter signals and offering a much-refined perception. If a vehicle accident occurs, nearby monitoring systems can collect relevant data, analyze



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:57:49 UTC from IEEE Xplore. Restrictions apply.

the cause, and disseminate information to alert other vehicles and users. However, given the vast number of vehicles and monitoring devices involved, limited bandwidth and the distance between data sources and data centers can delay the release of urgent information, potentially leading to serious consequences. Thus, by deploying edge nodes—such as edge servers and sensors—close to roadside infrastructure and vehicles, critical data can be processed and analyzed near the source, ensuring faster response times in emergency situations.

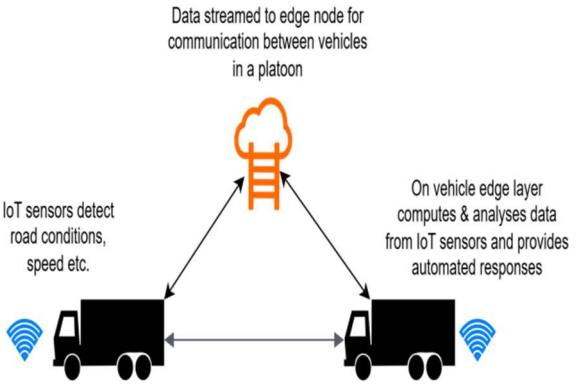



Fig.5. Smart Transportation: Facilitating safer & Efficient Driving.

QECC significantly benefits applications requiring low latency, requiring high data bandwidth, or involving large amount of IoT devices with limited computing power and better capacity.

## VIII. Challenges and Future Research Directions

### A. Challenges

While Quantum Edge Cloud Computing (QECC) offers a radical approach to address the computational demands of large-scale Internet-of-Things (IoT) networks, multiple obstacles must be overcome before it can achieve mainstream embracement. These challenges traverse across hardware limitations, security sensitivities, integration complexities and high architecture costs.

Quantum qubits are extremely fragile and reactive to extrinsic turmoil which make them unreliable and prone to errors for prolonged computations.

Quantum computers have excessive energy requirements. They demand high-power cooling systems and very-low temperatures which makes large-scale arrangement an energy ferocious and costly attempt.

Quantum qubits struggle to maintain their state for longer periods which brings disturbance in data processing and it even limits their practical usage.

To establish a seamless connection between Classical computing architectures and Quantum processors one has to face various technological bottlenecks.

Quantum cryptography offers stronger security but the ability of quantum systems to fracture classical encryption propounds serious risks to existing IoT security models.

Commutation of quantum information across distributed networks can reintroduce delays which impacts processing speed and efficiency of the system.

Due to hardware and connectivity constraints, it is a complex challenge to expand quantum architecture to accommodate large-scale IoT ecosystems.

Conventional quantum processors exhibit continual computation errors which necessitates sophisticated error correction techniques that require additional resources.

### B. Future Research Directions

Despite all these challenges there is always a continuous advancement in quantum computing, cryptography and network infrastructure which are paving the way for many efficient QECC solutions. Future development will focus on improving stability of qubits, energy-efficient architecture, hybrid commuting models and quantum-secure IoT frameworks.

Advancement in qubit design are necessary to improve stability, enhance computational accuracy and minimize errors, The research is focused on unfolding topological qubits, superconducting qubits and alternate material to make quantum qubits much more reliable and practical for IoT applications.

There must be advancements in Quantum memory and Data retention solutions as present quantum memory systems have limitations in terms of storage capacity and coherence time. Many researchers are working on establishing more stable quantum storage solutions such as spin-based memory and quantum dots in order to enable scalable and long-term quantum data retention.

Quantum computing is a rising technology, the existing cryptographic protocols face obsolescence. Many quantum-resistant cryptographic techniques such as Lattice-based encryption and hash-based signatures are being developed to ensure long term security of IoT networks.

Quantum computing requires significant energy for its operation and cooling. Future directions include research in fields like quantum photonics, energy- efficient quantum circuits and superconducting materials to develop environmentally sustainable computing solutions.

## IX. Conclusions

Quantum Edge Cloud computing (QECC) shows an incarnation in the way IoT systems process, analyse and secure immense amounts of data. By incorporating quantum computing with edge and cloud architectures, QECC presents unparalleled computational power, drastically reduces latency, optimizes real-time decision making and addresses security issues through progressive cryptographic methods. This confluence allows IoT networks to operate with high efficiency, even in complex and large-scale environments.

In spite of the fact that QECC is still in its early phase, there are numerous challenges such as integration complexities, quantum hardware instability and high implementation cost that remain hurdles to extensive adoption.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:57:49 UTC from IEEE Xplore. Restrictions apply.

However, research in many areas such as quantum-classical hardware models, error-resistant qubit infrastructures and quantum-enhanced artificial intelligence is continuously pushing the technology towards practical implementation. These obstacles are expected to mitigate sooner due to the continuous evolution of quantum computing which will make QECC a scalable and feasible solution for upcoming generation IoT applications. QECC will redefine the future of IoT by simulating adaptive, intelligent and exceedingly secure networks. QECC is set to unlock a new era of innovation with further advancements in quantum networking, sustainable quantum computing models and post-quantum cryptography which will reshape the digital landscape for years to come.

## REFERENCES


[1] J. Zhang, "Quantum healthcare analysis based on smart IoT and mobile edge computing: way into network study," *Opt. Quantum Electron.*, vol. 56, no. 4, pp. 1–16, Apr. 2024, doi: 10.1007/s11082-024-06285-y.

[2] D. Javeed, M. S. Saeed, I. Ahmad, M. Adil, P. Kumar, and A. K. M. N. Islam, "Quantum-empowered federated learning and 6G wireless networks for IoT security: Concept, challenges and future directions," *Futur. Gener. Comput. Syst.*, vol. 160, pp. 577–597, Nov. 2024, doi: 10.1016/j.future.2024.06.023.

[3] Kulkarni, Srinidhi, Rishabh Kumar Tripathi, and Meera Joshi. "A Study on Data Security in Cloud Computing: Traditional Cryptography to the Quantum Age Cryptography." *System Design Using the Internet of Things with Deep Learning Applications*. Apple Academic Press, 2023. 147-174.

[4] Jain, Sarthak, Anant Vaibhav, and Lovely Goyal. "Raspberry Pi based interactive home automation system through E-mail." In *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*, pp. 277-280. IEEE, 2014.

[5] K. Kumar, A. Pant, N. Gupta, V. Kumar, A. Kumar, and A. Jain, "Developing an Intelligent Garbage Surveillance System Using IoT and Sensor Technology," 2024. doi: 10.1109/IC3TES62412.2024.10877612.

[6] Serrano, Manuel A., et al. "Minimizing incident response time in real-world scenarios using quantum computing." *Software Quality Journal* 32.1 (2024): 163-192.

[7] Hamdan, Salam, Moussa Ayyash, and Sufyan Almajali. "Edge-computing architectures for internet of things applications: A survey." *Sensors* 20.22 (2020): 6441.

[8] Hua, Haochen, et al. "Edge computing with artificial intelligence: A machine learning perspective." *ACM Computing Surveys* 55.9 (2023): 1-35.

[9] Jeyaraj, Rathinaraja, et al. "Resource management in cloud and cloud-influenced technologies for internet of things applications." *ACM Computing Surveys* 55.12 (2023): 1-37.

[10] Rao, BB Prahlada, Paval Saluia, Neetu Sharma, Ankit Mittal, and Shivay Veer Sharma. "Cloud computing for Internet of Things & sensing based applications." In *2012 sixth international conference on sensing technology (ICST)*, pp. 374-380. IEEE, 2012.

[11] P. Yadav, K. Kumar, and S. C. Sharma, "Machine Learning Based Techniques for Node Localization in WSN: A Survey," *Proc. - IEEE Int. Conf. Device Intell. Comput. Commun. Technol. DICCT 2023*, pp. 12–17, 2023, doi: 10.1109/DICCT56244.2023.10110235.

[12] Wu, Yulei. "Cloud-edge orchestration for the Internet of Things: Architecture and AI-powered data processing." *IEEE Internet of Things Journal* 8.16 (2020): 12792-12805.

[13] Ren, Ju, et al. "Edge computing for the internet of things." *IEEE Network* 32.1 (2018): 6-7.

[14] Wang, Xiaofei, Yiwen Han, Victor CM Leung, Dusit Niyato, Xueqiang Yan, and Xu Chen. "Convergence of edge computing and deep learning: A comprehensive survey." *IEEE communications surveys & tutorials* 22, no. 2 (2020): 869-904.

[15] L. U. Khan, I. Yaqoob, N. H. Tran, S. Kazmi, T. N. Dang, and C. S. Hong, "Edge computing enabled smart cities: A comprehensive survey," IEEE Internet Things J., early access, Apr. 10, 2020

[16] A. Ghosal and M. Conti, "Key management systems for smart grid advanced metering infrastructure: A survey," IEEE Commun. Surveys Tuts., vol. 21, no. 3, pp. 2831–2848, 3rd Quart., 2019.

[17] Y. Xiao, Y. Jia, C. Liu, X. Cheng, J. Yu, and W. Lv, "Edge computing security: State of the art and challenges," Proc. IEEE, vol. 107, no. 8, pp. 1608–1631, Aug. 2019.

[18] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data security and privacy-preserving in edge computing paradigm: Survey and open issues," IEEE Access, vol. 6, pp. 18209–18237, 2018.

[19] Q. Li, P. Chen, and R. Wang, "Edge computing for intelligent transportation system: A review," in Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health. Singapore: Springer, 2019, pp. 130–137.

[20] D. Chawla and P. S. Mehra, "QSMAH: A novel quantum-based secure cryptosystem using mutual authentication for healthcare in the internet of things," Internet of Things, vol. 24, p. 100949, Dec. 2023, doi: 10.1016/j.iot.2023.100949.

[21] Leymann, Frank, Johanna Barzen, Michael Falkenthal, Daniel Vietz, Benjamin Weder, and Karoline Wild. "Quantum in the cloud: application potentials and research opportunities.",pp. 9-24, *arXiv preprint arXiv:2003.06256* (2020).

[22] M. Satyanarayanan, "The emergence of edge computing", Computer, vol. 50, no. 1, pp. 30-39, Jan. 2017.

[23] Z. M. Zha, F. Liu and Z. P. Cai, "Edge computing: Platforms; Applications and challenges", J. Comput. Res. Develop., vol. 55, no. 2, pp. 327-337, 2018.

[24] L. Jiao, R. Friedman, X. M. Fu, S. Secci, Z. Smoreda and H. Tschofenig, "Cloud based computation offloading for mobile devices: State of the art challenges and opportunities" in Future Netw. Mobile Summit, Lisbon, Portugal, pp. 1-11, Jul. 2013.

[25] Zhang, Y. C. Zhao and B. Chen, "Survey on data security and privacy-preserving for the research of edge computing", J. Commun., vol. 39, no. 3, pp. 1-21, 2018.

[26] D. Chawla and P. S. Mehra, "QAKA: A novel quantum authentication and key agreement (QAKA) protocol using quantum entanglement for secure communication among IoT devices," Trans Emerging Tel Tech, vol. 35, no. 3, p. e4957, Mar. 2024.

[27] D. Chawla and P. S. Mehra, "Performance Analysis of the Quantum Schemes in the Resource-Constrained IoT Network," in Intelligent System and Data Analysis, P. Nanda, S. Srivastava, V. K. Verma, and P. Vyas, Eds., Singapore: Springer Nature, 2025, pp. 427–443.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 16,2025 at 10:57:49 UTC from IEEE Xplore. Restrictions apply.

Check for updates

ChemistrySelect

Review
doi.org/10.1002/slct.202405247

Chemistry Europe
European Chemical Societies Publishing

www.chemistryselect.org

# Recent Advances on Catalysts Based on $MoS_2$: A Review on Diverse Applications

Priya,[a] Suman Srivastava,*[b] and Ritika Khatri*[a]

This review explores the versatile catalytic applications of molybdenum disulfide ($MoS_2$), emphasizing its significance in chemical transformations. Out of various transition metal dichalcogenides (TMDs), $MoS_2$ has garnered considerable attention due to its high surface area, tunable band gap, and unique electronic properties, which collectively contribute to its exceptional catalytic efficiency. Its layered structure, enriched with active edge sites, facilitates strong chemical interactions and enables diverse applications in electrocatalysis, photocatalysis, and cocatalysis. In electrocatalysis, $MoS_2$ demonstrates remarkable efficiency in hydrogen evolution and oxygen reduction reactions, with enhanced activity and stability achieved through nano-structuring and surface modifications. As a photocatalyst, its ability to optimize band structures and promote charge separation makes it highly effective for water splitting and pollutant degradation. $MoS_2$ acts as an excellent cocatalyst to enhance the catalytic properties of various catalysts by modifying the surface area of adsorption, the reactive oxygen species (ROS) production mechanism, and the microscopic process change through its active sites. The recent advancements in heterojunctions, nanocomposites, and hybrid materials of $MoS_2$ and their catalytic performance are also highlighted.

## 1. Introduction

Two-dimensional (2D) nanomaterials, characterized by their layered structure and substantial surface area, have become a focal point in research due to their intriguing physicochemical properties. The exploration of monolayer graphene initiated a surge in the sighting of 2D materials. The distinctive physical and chemical properties exhibited by these materials are largely influenced by the weak van der Waals forces between their basal planes.[1] Within the typical layer-to-layer distances of 0.4–0.6 nm, van der Waals forces facilitate the formation of well-structured layered materials.[2] A primary advantage of 2D materials lies in their ability to integrate or stack different thin layers, enabling the synthesis of novel heterostructures. This also leads to synergistic effects between layers, significantly enhancing their performance as heterostructural catalysts.[3] The layered structure possesses distinctive features such as expanded active surfaces, seamless interfaces between heterostructures, swift carrier transport, and band gaps that can be readily modified by gate tuning.[4] In 2D materials, the efficiency of electronic carrier transportation, such as phonons or electrons, is notably higher within the basal plane than between distinct layers. As a result, the properties and behavior of phonons and electrons in 2D catalyst materials are greatly influenced by the properties of the 2D plane. The research field has predominantly focused on TMDs among the numerous 2D materials available, owing to their diverse and distinctive electrical, optical, physical, and chemical properties.[5] TMDs are expressed by the chemical formula $MX_2$, where M belongs to the transition metal like titanium (Ti), tungsten (W), molybdenum (Mo), niobium (Nb), and so on, and X is an element from the oxygen family such as sulphur (S), tellurium (Te), selenium (Se), and so on, and these are called as chalcogens.[6] The coordination geometries of transition metal atoms can vary between octahedral and trigonal prismatic.[7] TMD typically manifests two prevalent structures, namely 1T(octahedral) and 2H(Trigonal prismatic). Each of the arrangements is made up of three levels of -M- containing a hexagonally arranged layer of metal atoms located between two layers of chalcogen atoms, with covalent bonds between the transition metal and chalcogen atoms. TMD materials, possessing thicknesses of just a few nanometres, can be stacked to form unique structures, exhibiting plate-like structures. Some examples of TMD materials are $MoS_2$, tungsten disulfide ($WS_2$), vanadium disulfide ($VS_2$), molybdenum diselenide ($MoSe_2$), and so on. These materials possess advantageous characteristics, including strong spin-orbit coupling, nanometer-scale thickness, and beneficial mechanical and electrical properties, owing to their partially filled d-band, sandwiched structure, and layered structure induced by van der Waals forces.

Out of all of them, $MoS_2$, a TMD, is emerging as a promising 2D nanomaterial. $MoS_2$ is a stable, nonhazardous, and cost-effective material that can be synthesized using various methods and has demonstrated its exceptional capabilities in catalysis,[8] biosensing,[9] supercapacitors,[10] gas sensing,[11] environmental applications,[12] and optoelectronics[13] etc. A catalyst is characterized as a substance that speeds up a chemical reaction rate by lowering its energy of activation while remaining unchanged in its chemical composition throughout the reaction. $MoS_2$ serves as a versatile catalyst across multiple domains, which is attributed to its capacity to improve both reaction rates and overall efficiency. Its unique layered structure offers

[a] Priya, R. Khatri
School of Basic and Applied Sciences, K. R. Mangalam University, Sohna, Gurugram, Haryana, 122103, India
E-mail: khatriritika1992@gmail.com
ritika.khatri@krmangalam.edu.in

[b] S. Srivastava
Department of Chemistry, M. V. College Buxar (A Constituent Unit of Veer Kunwar Singh University Ara) Bihar, Bihar 802101, India
E-mail: sumansamiran@gmail.com



© 2025 Wiley-VCH GmbH

extensive surface area and a significant concentration of active edge sites, and its high recyclability enhances the efficiency of various catalytic processes.[14] The applications of $MoS_2$ extend beyond typical catalysis; it is notably effective as a photocatalyst, harnessing light energy to drive chemical reactions, and as an electrocatalyst, enhancing the efficiency of electrochemical reactions, such as $H_2$ production. Additionally, $MoS_2$ can function as a cocatalyst in conjunction with other materials, thereby improving overall catalytic performance. Its versatility is further demonstrated in organic transformations, where it participates in reactions like hydrodesulfurization and dehydrogenation. The diverse applications of $MoS_2$ in these catalytic roles underscore its significant importance and adaptability in advancing both industrial and research-focused catalytic technologies. Previously reported reviews have tended to concentrate on specific catalytic uses of $MoS_2$.[15a–c] Our review takes a more comprehensive approach, exploring $MoS_2$ roles in diverse areas such as photocatalysis, electrocatalysis, and cocatalysis in various organic transformations collectively. By thoroughly examining these varied catalytic functions, we present a more complete understanding of $MoS_2$ capabilities and possibilities as a catalyst. This view point not only underscores the adaptability of $MoS_2$ in diverse catalytic processes but also reveals potential synergies and avenues for various catalytic conversions. Additionally, our study incorporates the most recent developments and applications of $MoS_2$ as a versatile catalyst and sets the stage for future research and progress in catalytic transformations.

## 2. $MoS_2$ as an Emerging Catalytic Material

Out of all known TMDs, $MoS_2$ is the most studied due to its distinctive composition and vast application. $MoS_2$ is becoming increasingly important as a promising catalyst with considerable potential in diverse chemical reactions, especially in sustainable energy generation.[16] Known for its layered structure and exceptional electronic properties, its active sites for catalytic activity are found in its edges and defect sites, making it highly effective. No doubt, graphene-based and hexagonal boron nitride (hBN) -based 2D nanomaterials have been broadly used in various applications because of their unique properties.[17] However, the lack of reactivity in hBN catalysts is due to strong covalent bonds, making it difficult to modify for energy conversion uses, thus restricting its catalytic capabilities.[18] Graphene as a catalyst has different limitations, such as the need for high energy input, lack of band gap, and potential negative impacts on living organisms.[19–21] On the other hand, $MoS_2$ provides several distinct advantages over both graphene-based and hBN-based nanomaterials, making it a promising choice for a range of applications. $MoS_2$ can be a viable substitute for catalysts like platinum (Pt)[22] in hydrogen evolution reaction (HER) and photocatalytic water splitting. $MoS_2$ exhibits excellent visible light performance as a photocatalyst for organic contaminant decomposition in wastewater.[23] It also acts as a cocatalyst in the Fenton reaction and organic transformations.[24]



Priya received her master's degree from Panjab University (PU), Chandigarh, India. She is a Ph.D. scholar and Innovative Research Fellow (IRF) in the School of Basic and Applied Sciences at K.R. Mangalam University, Gurugram, Haryana, India. She is dedicated to driving innovation in material science. Her research interest includes the synthesis of 2D nanomaterials with particular emphasis on transition metal dichalcogenides (TMDs) and their applications in various catalytic transformations and sensing.



Dr. Suman Srivastava is an Assistant Professor at Department of Chemistry, M. V. College Buxar (A Constituent Unit of Veer Kunwar Singh University Ara) Bihar, India. She earned her master's degree in chemistry from Banaras Hindu University (BHU), Varanasi, India, in 2006. She completed her Ph.D. in 2012 under Prof. Ranjana S. Khanna at BHU and Dr. Atul Kumar (CDRI-CSIR). Dr. Srivatava pursued postdoctoral research at the Indian Institute of Technology (IIT), Mandi, Himachal Pradesh, India. She worked as a DST Inspire Faculty at the National Institute of Technology (NIT) Delhi, India. Her research interests focus on the catalytic activity of transition metal dichalcogenides (TMDs).



Dr. Ritika Khatri is working as an Assistant Professor at the School of Basic and Applied Sciences, K.R. Mangalam University, Gurugram, Haryana, India. She obtained her graduation and post-graduation degrees in physics in 2014 and 2016, respectively, from the University of Delhi. Thereafter, she received her Ph.D. degree in physics in 2023 from Delhi Technological University (DTU), Delhi, India, under the Supervision of Prof. Nitin K. Puri. Her research is focused on the synthesis of 2D nanomaterials with a specific focus on $MoS_2$, for applications in biosensing, gas sensing, and catalysis.


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

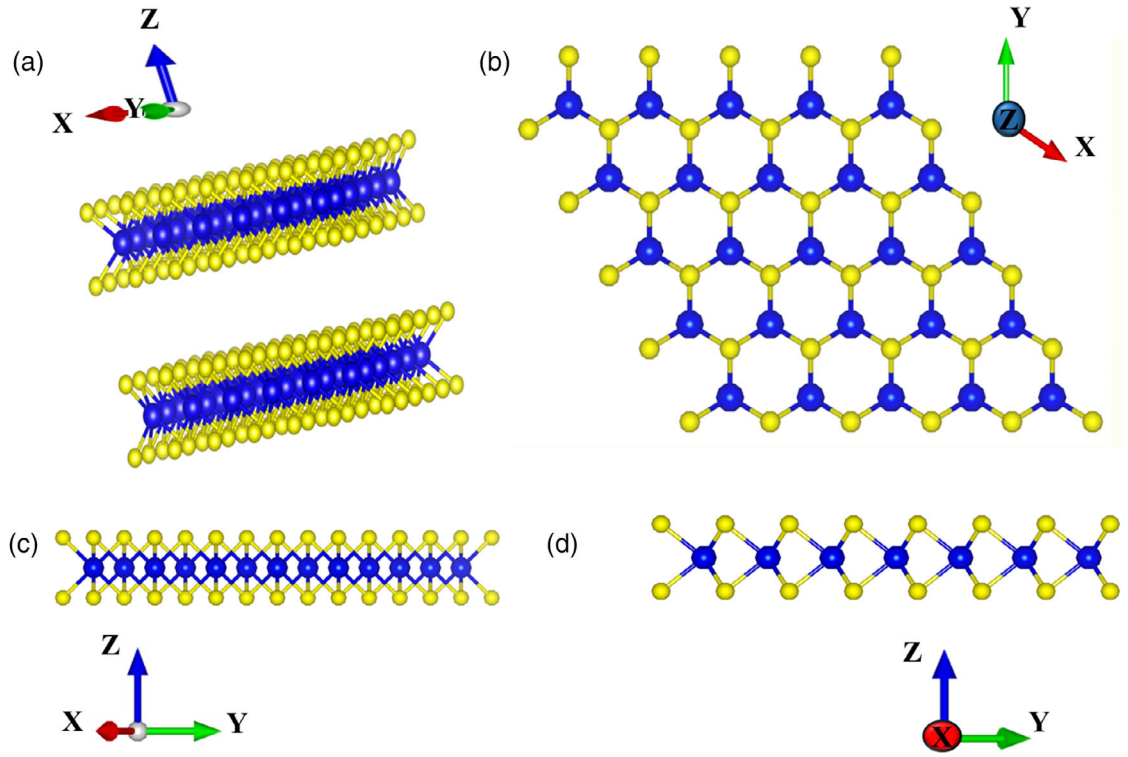



**Figure 1.** Different views [(a) 3d view, (b) top view, (c) side view along *y*-axis, and, (d) side view along *x*-axis] of the crystalline structure of monolayer $MoS_2$. Mo atoms (blue color) are positioned between the double layers of S atoms (yellow color).

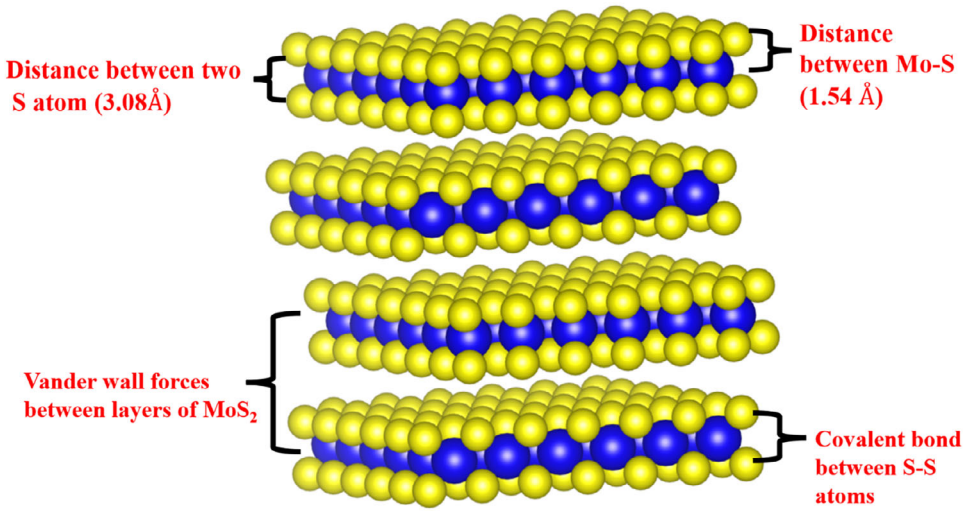



**Figure 2.** Layered structure visualization of $MoS_2$ depicting the bond length between different layers of atoms (Mo—S and S—S) and the force between the layers.

### 2.1. Structure of $MoS_2$

The layered structure of $MoS_2$ comprises molybdenum atoms sandwiched between two sulphur atoms in each layer (Figure 1). The interaction between these layers is governed by the presence of weak van der Waals forces, enabling an easy exfoliation of this material into thin sheets. The bond length between Mo and S is measured to be 1.54 Å, while the bond length between S and S is 3.08 Å (Figure 2).[25] Bulk $MoS_2$ consists of vertically packed layers bound by weak VanderWaals forces and an indirect band gap of 1.20 eV. In contrast, single-layer $MoS_2$ shows a direct band gap of 1.80 eV.[26] This specific structure presents numerous opportunities for diverse applications in energy storage, $H_2$ production, contaminant degradation, disinfection of water, and so on. There are three distinct types of $MoS_2$, which are determined by the coordination of S atoms and Mo atoms, as well as the sequence in which the layers are constructed. These classifications are denoted as 2H-$MoS_2$, 1T-$MoS_2$, and 3R-$MoS_2$. 1, 2, and 3 correspond to the respective number of layers present in each type, respectively. The letters R, T, and H


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

represent the crystal structures, which are rhombohedral, trigonal, and hexagonal, respectively.[27] Out of all, rhombohedral and hexagonal are very stout and naturally occurring. 2H-$MoS_2$ is the most frequently found poly type and is commonly utilized because it is abundant in the earth's crust. The coordination geometry of both 2H and 3R-$MoS_2$ is characterized by a trigonal prismatic arrangement, in contrast to 1T-$MoS_2$, which exhibits an octahedral coordination of Mo atoms. The structural differences between 2H and 3R-$MoS_2$ arise from the arrangement of layers. Additionally, $MoS_2$ exhibits structural defects like point defects, grain boundaries, and edges, which offer opportunities for surface modification and functionalization, thereby influencing their chemical, electrical, and optical properties.[28]

### 2.2. Various Methods for the Synthesis of $MoS_2$

$MoS_2$ nanomaterials can be categorized based on their dimensional characteristics, including zero-dimensional nanodots or quantum dots (0D),[29] one-dimensional nanotubes (1D),[30] two-dimensional nanosheets or nanoflakes (2D),[31] and three-dimensional nanocomposites (3D).[32] This classification system is primarily divided into two fundamental methodologies: the top-down approach and the bottom-up approach. The top-down methodology, alternatively called stepwise design, consists of dissecting a bulk entity into its compositional subsystems. The method involves breaking down bulk $MoS_2$ into its nanoscale form through physical or mechanical means, as well as utilizing solvents and reagents for exfoliation and intercalation. Various top–down techniques for preparing nanoscale $MoS_2$ include liquid/solvent-assisted exfoliation, mechanical exfoliation, electrochemical exfoliation, and chemical-assisted exfoliation. On the other hand, the bottom-up strategy entails combining elemental subsystems to form a complex structure. This method encompasses techniques like hydrothermal/solvothermal reactions and chemical vapor deposition (CVD), enabling the direct synthesis of nanomaterials from precursor materials.[33]

However, the advantages and disadvantages of each of these methods should be considered. For instance, it has been observed that synthesis techniques that possess characteristics such as safety, ease of execution, cost-effectiveness, and the ability to manufacture nanoparticles in reduced sizes are more advantageous and suitable for biomedical applications. In these applications, the size and toxicity of the synthesized nanomaterials play a crucial role in determining the effectiveness of the specific application. Methods such as liquid/solvent-assisted exfoliation, chemical-assisted exfoliation, intercalation, and hydrothermal/solvothermal methods become the best way to prepare $MoS_2$ nanostructures that are smaller in size[34] and exhibit better biocompatibility, particularly in the field of biomedicine.

#### *2.2.1. Chemical Exfoliation*

The process of chemical exfoliation of 2D nanomaterials entails the application of solvents or chemical agents to disaggregate layered materials into monolayers or few-layer configurations. This technique often utilizes ultrasonication or chemical treatments to improve dispersion and attain uniformity. It facilitates the scalable synthesis of high-quality nanomaterials that exhibit favorable electronic and mechanical characteristics. Solvent-based exfoliation, also called chemical exfoliation, can be divided into two primary forms: intercalation exfoliation and electrochemical exfoliation. Intercalation exfoliation involves the separation of $MoS_2$ layers through chemical reactions, making it the most traditional and widely used method for the large-scale synthesis of $MoS_2$.[35] By inserting ions into the interlayer space between $MoS_2$ layers, an expansion force is created, causing the $MoS_2$ layers to separate. The mechanical force or gentle agitation can enhance this process, with ultrasonication being a common technique used by researchers. The combination of ultrasonication and intercalation can accelerate the production of atomically thin $MoS_2$ layers and enhance the parameters of the resulting materials. It has been observed that this procedure can effectively decrease the lateral size of nanoflakes[36] and produce monolayer $MoS_2$ quantum dots after repeated exfoliation.[37] Nevertheless, expertise is required to execute this method, and challenges exist for its large-scale implementation. Research indicates that by elevating the power input and using a higher frequency of ultrasound, it is possible to improve the quality and thickness of the nanoflakes. Sonication for an extended period can lead to structural defects, diminishing its efficacy in industrial settings, especially in electronics and circuit devices. Ultrasonication produces waves of high pressure that aid in the easier infiltration of intercalating compounds among sulfur layers. The incorporation of ultrasound can decrease the intercalation duration. Moreover, ultrasound can assist in the separation of $MoS_2$ layers during mechanical exfoliation by reducing the adhesive forces among them.Thus, minimizing the amount of mechanical force required for exfoliation and lowering the possibility of causing damage to the thin layers.[38] Following intercalation, the thin $MoS_2$ layers can be isolated through transfer processes like deposition or via filtration over the base (substrate). The thin layers produced have properties dependent upon several factors, including the consistency of the intercalation process, the dimensions and morphology of the $MoS_2$ structures, characteristics of the intercalating factor, the method of application, and the circumstances in which exfoliation takes place. A range of characterization techniques, such as photoluminescence (PL), transmission electron microscopy (TEM), and Raman spectroscopy can be utilized to investigate the optical, structural, and electronic properties of the synthesized layer.[39] Intercalation processes typically involve various ions, such as potassium (K), sodium (Na), phosphorus (P), chlorine (Cl), lithium (Li), and ammonium ($NH_4^+$). Various considerations determine the selection of intercalating ions, including ion size, polarizability, intercalation kinetics, the stability of the resulting $MoS_2$ nanoflakes, and the feasibility of ion recovery. In one investigation,[40] $Na^+$ was utilized as an intercalation agent to assist in the exfoliation of $MoS_2$ nanoflakes. The exfoliation process involved the combination of liquid-phase reflux with $Na^+$ and ultrasound treatment in an *N*-methyl-2-pyrrolidone (NMP) solvent. The properties of these exfoliated nanoflakes were improved, especially in applications


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

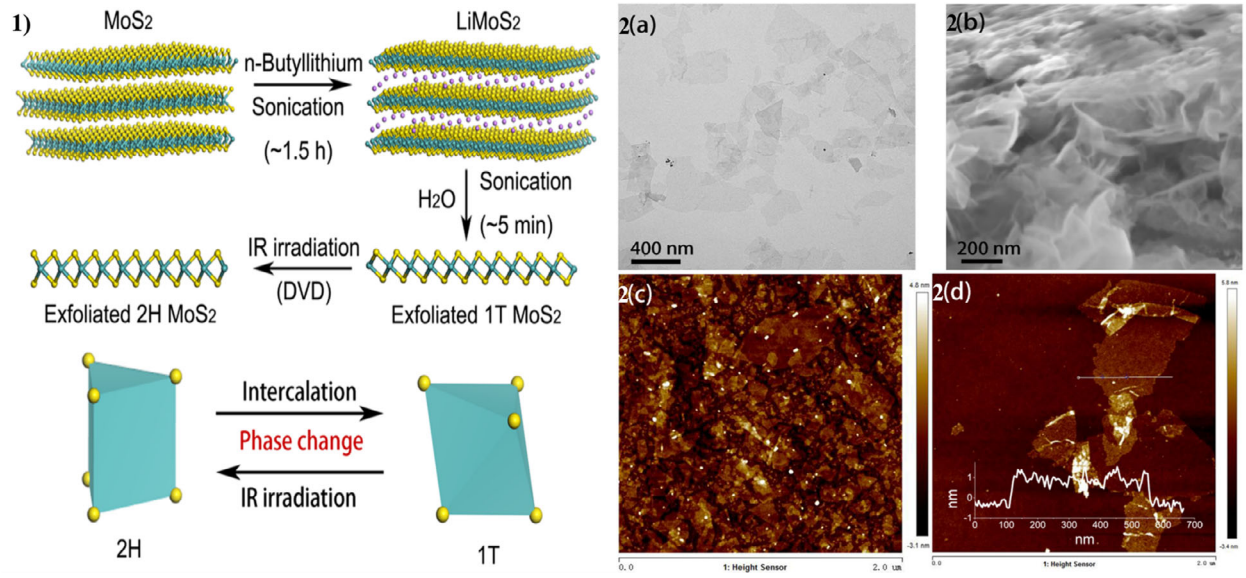



**Figure 3.** 1) Synthesis pathway for the generation of exfoliated 2H $MoS_2$, 2(a) TEM images of exfoliated $MoS_2$ at a scale of 400 nm, 2(b) SEM image of exfoliated $MoS_2$ at a scale of 200 nm, 2(c and d) AFM images of exfoliated $MoS_2$ at scan sizes of 2 μm. Reprinted with permission from ref. [36]. Copyright 2015, American Chemical Society.

involving photocatalysis. Despite this, lithium-ion ($Li^+$) remains the most used intercalating ion due to its small size, which enables reversible intercalation into lamellar-structured materials. When bulk crystals are exposed to a lithium-ion source (typically *n*-butyllithium in hexane solution), the ions penetrate between $MoS_2$ layers, resulting in uneven distribution and locally high Li content, leading to increased reducibility and defects in $MoS_2$.[41] Ultrasonication causes the layers to separate, forming monolayer or few-layer $MoS_2$, often with a metallic structure (1T) that can be converted to the desired semiconductor structure (2H) via thermal annealing. The process of synthesis of exfoliated $MoS_2$ using Li intercalation is depicted in Figure 3 below.

### *2.2.2. Liquid Phase Exfoliation (LPE)*

LPE of $MoS_2$ involves dissolving $MoS_2$ crystals in a solvent or solution and then exfoliating them into thin layers using mechanical, chemical, or thermal techniques. The simplicity, cost-effectiveness, and efficiency of this technique facilitate its wide utility for creating $MoS_2$ sheets with outstanding properties. LPE of $MoS_2$ provides additional benefits, including the ability to regulate the thickness of exfoliated flakes, easy modifications of flake surfaces chemically, and the possibility of utilizing a variety of support materials. Consequently, it is extensively employed in preparing nanomaterials for diverse applications such as sensors, catalysts, batteries, and electronic devices.[42] $MoS_2$'s exfoliation involves the use of various solvents, including water, alkaline solutions of sodium carbonate, potassium hydroxide (KOH), sodium hydroxide, and organic solvents like *N*-methyl pyrrolidone (NMP), dimethylformamide (DMF), acetone, and chloroform. Water offers specific advantages as an exfoliation solvent. A recent research study suggests that water can substitute organic solvents in spreading $MoS_2$ nanoflakes.[43] Water dispersions do not have the same level of stability as dispersions in organic solvents, but the nanoflakes that are created can vary in lateral dimensions from 50 μm to 500 nm and are maintained in a stable state due to electric charges.[44] Furthermore, Gupta et al.[45] observed that adding the right quantity of water to NMP enhances exfoliation yields, stabilizes the nanoflakes for an extended period, and decreases the defect densities of exfoliated nanoflakes, as illustrated in Figure 4. Ma and coworkers suggested that water may not be a suitable solvent for exfoliating 2D materials because of variations in surface energy.[46] The restriction is especially important when it comes to exfoliating $MoS_2$ because $MoS_2$ is highly attracted to water and could clump together again, which would decrease the level of exfoliation obtained. Conversely, alkaline solutions provide possible benefits for $MoS_2$ exfoliation, such as effective and environment-friendly synthesis and the creation of smaller nanoflakes in comparison to the larger original material. Wu et al.[47] reported a combination of lithium hydroxide/sodium hydroxide (LiOH/NaOH), where a mixed alkaline solution was utilized to produce $MoS_2$ nanoflakes cost-effectively and sustainably. The nanoflakes, which had approximately 2–9 layers, when used as a lubricating additive, exhibited antiwear and antifriction characteristics. This method is more efficient in comparison to the other techniques.[48] The exfoliation process is significantly influenced by the choice of organic solvents, with surface tension and solubility limitation, influencing the choice of solvent. It has been discovered that matching the surface tension of the solvent with that of $MoS_2$ is extremely effective for achieving LPE and large-scale synthesis of high quality. Minimizing the difference in Gibbs free energy ($\Delta G$) during exfoliation is imperative for successful exfoliation, with surface tension playing a key role in this process.[49]


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

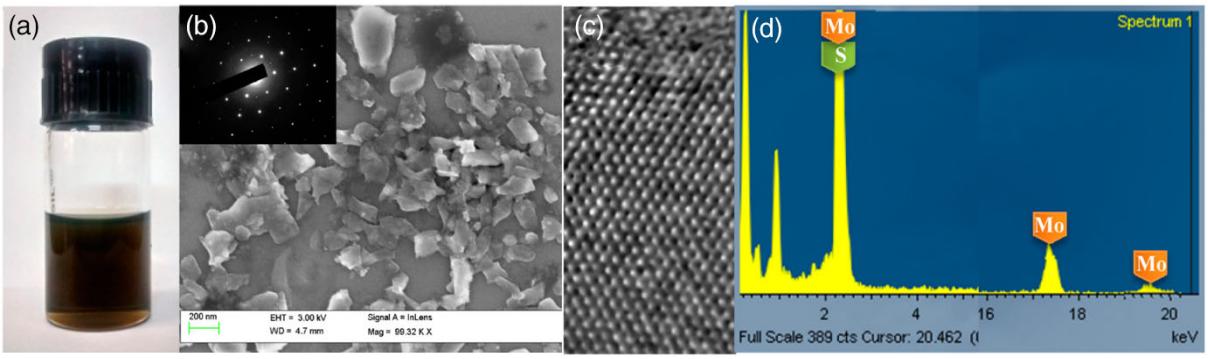



**Figure 4.** (a) The dispersion of $MoS_2$ in NMP with 0.1 mole fraction of water. (b) Scanning electron microscopy (SEM) image of $MoS_2$-NMP dispersion (with a water mol fraction of 0.1) displaying the exfoliated layers of $MoS_2$. The inset shows the electron diffraction pattern. (c) High-resolution transmission electron microscopy (TEM) image of an individual sheet and (d) the corresponding energy-dispersive X-ray spectroscopy (EDXS) data of exfoliated $MoS_2$. Reprinted with permission from ref. [45]. Copyright 2016, American Chemical Society.

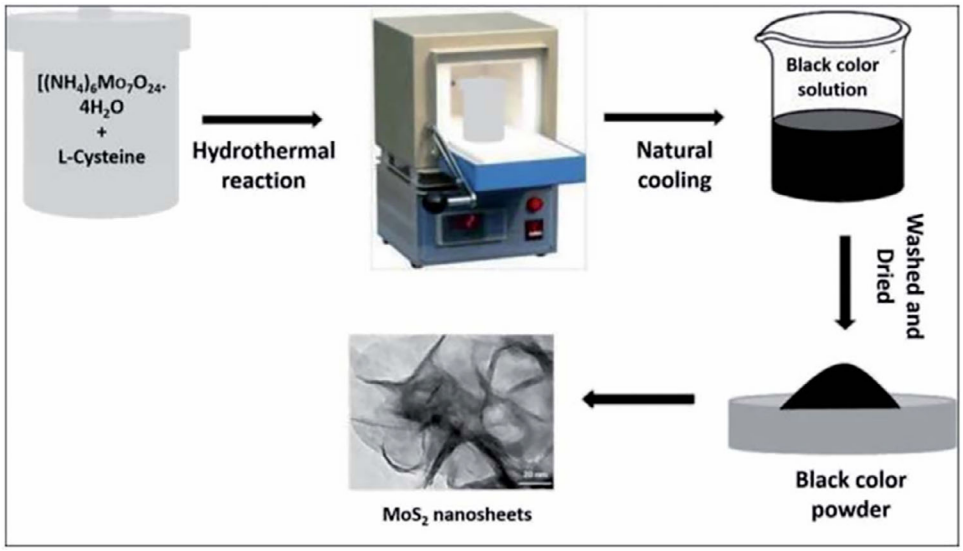



**Figure 5.** Synthesis of $MoS_2$ by hydrothermal method. Reproduced with permission from ref. [51].

### *2.2.3. Hydrothermal Method*

The hydrothermal method is a sophisticated approach for synthesizing $MoS_2$. Unlike conventional synthesis routes, hydrothermal synthesis operates under controlled aqueous environments at elevated temperatures and pressures.[50] This method harnesses the chemical reactivity of aqueous solutions to facilitate the formation of $MoS_2$ nanostructures from precursor materials. Key advantages of the hydrothermal approach include its scalability, allowing for the production of $MoS_2$ in bulk quantities suitable for industrial applications. Moreover, the method offers precise control over the morphology and composition of the synthesized $MoS_2$, ensuring uniformity and consistency in material properties.[51] These attributes make hydrothermal synthesis particularly advantageous in fields such as catalysis, where the specific surface area and structure of $MoS_2$ nanoparticles can significantly influence performance.[52]

The hydrothermal synthesis of 2D $MoS_2$ (Figure 5) can be achieved using a variety of precursors in different systems. The most commonly used Mo precursors include ammonium molybdate tetrahydrate [$(NH_4)_6Mo_7O_{24}$], molybdenum trioxide ($MoO_3$), and sodium molybdate ($Na_2MoO_4$), while S precursors consist of thiourea ($NH_2CSNH_2$), elemental sulphur (S), potassium thiocyanate (KSCN), sodium thiosulfate ($Na_2S_2O_3$), L-cysteine ($C_3H_7NO_2S$), and thioacetamide ($CH_3CSNH_2$). Typically, deionized water is used as the solvent. In a standard hydrothermal process, the Mo and S precursors are combined in deionized water in a Teflon-coated autoclave and then allowed to react at a specific temperature (e.g., 150–240 °C) for varying durations (e.g., 6–48 h). Furthermore, the use of single precursors such as ammonium tetrathiomolybdate [$(NH_4)_2MoS_4$] and molybdenum diethyl dithiocarbamate oxide [$Mo((C_2H_5)_2NCS_2)_2O_2$] have also been reported for the synthesis of $MoS_2$ nanosheets in aqueous solution through a hydrothermal reaction process.[53,54]

### *2.2.4. Chemical Vapor Deposition(CVD)*

CVD has been frequently employed for the growth of $MoS_2$ thin layers on insulating substrates, such as $SiO_2$ or sapphire.[55–57] Achieving crystallographic $MoS_2$ by controlling the number of


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

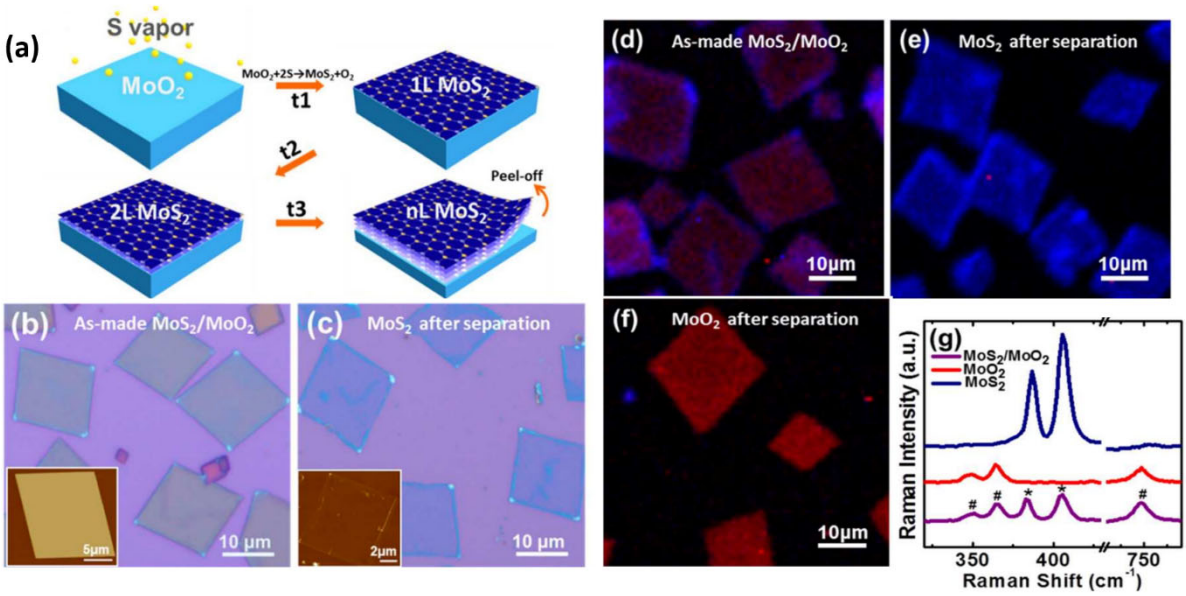



**Figure 6.** (a) The diagram shows how $MoS_2$:$MoO_2$ microplates are created and then broken down. (b) $MoO_2$ microplates are produced by reducing $MoO_3$, which acts as a template for the gradual growth of $MoS_2$ through a layer-by-layer surface sulfurization process. (c) The $MoS_2$ layer is detached from the $MoO_2$ substrate and transferred to another substrate, with the top layer positioned facing upwards for additional examination. The designations "1L", "2L", and "nL" correspond to single-layer, bilayer, and *n*-layer arrangements, respectively. (b) and (c) The $MoS_2$/$MoO_2$ plates were formed through a 3-h annealing process in sulfur vapor. The optical and atomic force microscopy (AFM) images depict these plates, as well as the transferred $MoS_2$ flakes, accompanied by Raman mapping images. (d–f) display the combined $MoS_2$ and $MoO_2$ (indicated in blue and red), as well as the separate nanosheets of $MoS_2$ (blue) and $MoO_2$ (red). Additionally, the Raman spectra obtained from the microplates comprising $MoS_2$/$MoO_2$ show characteristic peaks, with distinct signals corresponding to the individual materials $MoS_2$ and $MoO_2$. Reprinted with permission from ref. [58]. Copyright 2013, American Chemical Society.

layers through CVD is more challenging compared to graphene. At present, precursor reagents like Mo, $(NH_4)_2MoS_4$, or $MoO_3$ are used to produce $MoS_2$ layers.[58] The precursors are first placed on supports as thin layers and then undergo sulfurization or thermal decomposition at high operating temperatures to control the structure of $MoS_2$. Atomically, thin layers of $MoS_2$ can be produced using a two-step thermal evaporation–exfoliation process, which offers the benefit of needing only moderate conditions.[59] The approach mentioned can generate $MoS_2$ layers of a large size. However, these layers are polycrystalline with small crystallite dimensions because noncrystalline precursors are used, making it challenging to control the number of layers. To tackle these challenges, a method has been showcased for growing high crystallinity $MoS_2$ layers by regulating the number of layers using $MoO_2$microcrystallites as templates. In Figure 6a, one such growth process of $MoS_2$ layers is depicted. Initially, $MoO_3$ powder undergoes evaporation and reduction by sulfur vapor to generate $MoO_2$. The $MoO_2$ then forms nuclei on $SiO_2$/Si substrates and transforms into rhomboidal microsheets within a CVD furnace. Subsequently, the $MoO_2$ microsheets are annealed in sulfur vapor. The annealing process yields different numbers of $MoS_2$ layers by sulfurizing the surface of $MoO_2$ microsheets, depending on the duration of annealing. After this, a thin layer of polymethyl methacrylate (PMMA) is applied to the surface of $MoS_2$/$MoO_2$ microsheets to facilitate the separation of $MoS_2$ thin flakes from the $MoO_2$ layer. These separated flakes can then be transferred onto other supports using PMMA-mediated nanotransfer printing. The images in Figure 6b,c show the $MoS_2$/$MoO_2$ layers as they were initially prepared by reducing $MoO_3$ and then annealing, the separated $MoS_2$ layers, and the $MoO_2$ layers remaining on the supports. This demonstrates the successful creation of $MoS_2$ layers from $MoO_2$ precursors. To confirm the complete isolation of $MoS_2$ from $MoO_2$, Raman spectroscopy was employed. The produced $MoS_2$/$MoO_2$ plates displayed five Raman peaks in the 300–800 $cm^{-1}$ range under 514 nm excitation, which were attributed to $MoS_2$ ($E_{12g}$ mode ~384 $cm^{-1}$ and $A_{1g}$ mode ~405 $cm^{-1}$) and $MoO_2$ (~350 $cm^{-1}$, 365 $cm^{-1}$, and 748 $cm^{-1}$). Figure 6d–f shows Raman mapping images indicating peak intensities of approximately 405 $cm^{-1}$ (strongest peak of $MoS_2$, depicted in blue) and approximately 748 $cm^{-1}$ (strongest peak of $MoO_2$, depicted in red) obtained from the as-made $MoS_2$/$MoO_2$, separated $MoS_2$, and $MoO_2$, respectively. The mapping images of separated $MoS_2$ and $MoO_2$ display pure blue and red colors, respectively, further confirming the accurate separation of $MoS_2$ and $MoO_2$ at their interface, as shown in Figure 6g.

## 3. $MoS_2$ as a Catalyst and Its Catalytic Properties

A catalyst is defined as a substance that speeds up the chemical reaction rate by reducing the energy of activation necessary for the reaction to occur while remaining unchanged at the end of the process.[60] Catalysts can be categorized into various types according to their roles and applications. Homogeneous catalysts exist in the same phase as the reactants, typically in liquid or gaseous states, and function by directly interacting with the reactants in a solution. In contrast, heterogeneous catalysts are present in a different phase than the reactants, often as solids that facilitate reactions involving gaseous or liquid


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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.** Comparative evaluation of $MoS_2$ alongside other 2D materials for catalytic applications.

| S.No. | Nanomaterials | Composite | Application | Performance Parameters | Reference |
|---|---|---|---|---|---|
| 1 | $MoS_2$ | ZnS/$MoS_2$ composite | HER | Best hydrogen generation at Zn: Mo ratio of 1:1.30 with 3% Co-doping (1256 µmol $H_2$/hour) | Jayan et al.[98] |
| | | Porous $MoS_2$ nano islands | HER and OER | Enhanced OER and HER with overpotentials of 248 mV (OER) and 300 mV (HER) | Chen et al.[81] |
| | | $MoS_2$@$CoS_2$ heterojunction | HER | HER overpotential: 48 mV (KOH), 62 mV ($H_2SO_4$), 164 mV (PBS) | Gu et al.[118] |
| | | $MoS_2$ | $CO_2$RR | Overpotential:54 mV at −3.4 mA $cm^{-2,}$ Onset potential −0.164 V, Reduction current intensity 65 mA $cm^{-2}$ at −0.764 V, Faradaic efficiency of CO formation: ~98%. | [68] |
| | | Au-$MoS_2$ on carbon fiber. | Bacterial disinfectant. | 99.999% E. coli disinfection via ROS generation. | Chou et al.[132] |
| | | $MoS_2$/graphene. | HER | $H_2$ yield: 667.2 mmol $h^{-1}$ $g^{-1}$ (formic acid, 0.32 M). | Lee et al.[134] |
| | | $MoS_2$ | Dye degradation. | 89.9% BPB degradation after four cycles at pH 5.20. | Kisala et al.[133] |
| | | $CoFe_2O_4$/$MoS_2$ | Phenol degradation. | 100% phenol degradation in neutral pH. | Yan et al.[145] |
| | | $MoS_2$/$Ni_3S_2$/rGO | Alcohol fuel generation | Methanol OCV: 0.59 V, ethanol OCV: 0.48 V; peak power: 4.42 mW $cm^{-2}$ | Salarizadeh et al.[147] |
| | | PdNPs@$MoS_2$ | OER | Overpotential of 253 mV, current density 20 mA $cm^{-2}$. | [69] |
| 2 | Graphene | N-doped, O-functionalized and edge/ defect-rich graphene nanosheets on carbon cloth. | OER | Overpotential:351 mV at –10 mA $cm^{-2}$ | [70] |
| | | N-doped graphene | $CO_2$RR | Onset potential: –300 mV Maximum Faradaic efficiency of formate production: ~73% | [71] |
| 3 | $C_3N_4$ | Co-$C_3N_4$/carbon nanotube | OER and ORR | ORR onset potential: 0.9 V OER onset potential: 1.5 V OER potential of 1.61 V at 10 mA $cm^{-2}$. | [72] |
| | | O-g-$C_3N_4$ | Dye degradation | Degradation rate of MB is 81%. | [73] |
| 4 | h-BN | CdS@ h-BN/G | HER | Photocurrent density 2.7 µA $cm^{-2}$, HER activity 6465.33 µmol·$h^{-1}$ g $^{-1}$. | [74] |
| | | C-doped h-BN/GO | ORR | High selectivity 88%–95%, 90% Faradaic efficiency. | [75] |
| 5 | M-Xenes | $Mo_2CT_x$ | HER | Overpotential:283 mV at −10 mA $cm^{-2}$ | [76] |
| | | PtNWs/$Ti_3C_2T_x$-CNT | ORR | Onset potential- 1.06 V, half wave potential-0.896 V. | [77] |
| | | $MoS_2$QDs@ $Ti_3C_2T_x$QDs@ MWCNTs | MOR | Onset potential- 0.87 V, half wave potential-0.75 V | [78] |
| 6 | BP | Few-layer black phosphorus nanosheets | OER | Onset potential:1.45 V at 0.1 mA $cm^{-2}$ | [79] |
| | | Te-doped black phosphorus | OER | Onset potential: 1.49 V | [80] |


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

reactants.[61] Biocatalysts, commonly known as enzymes, are biological molecules that promote biochemical reactions.[62] Among these categories, heterogeneous catalysts are particularly prevalent in industrial settings due to their straightforward separation and potential for reuse. $MoS_2$ is recognized as an exceptional catalyst, attributed to its distinctive layered-like structure that offers a large surface area and numerous active sites.[63] The catalytic performance of the 2D catalyst $MoS_2$ is primarily focused on enhancing edge sites and imperfections to increase active surface area, influencing reaction kinetics. Various strategies aim to augment active sites on the catalyst's surface by modifying both edge and basal planes.[64] Research efforts have explored introducing defects to enhance catalytic efficiency[65] by stimulating local electron density, particularly benefiting processes like HER through adjustments in electronic configuration (Table 1).[66] Table 1 offers a comparative evaluation of $MoS_2$ alongside other 2D materials by comparing their properties based on respective catalytic applications.[67]

In recent years, there has been significant attention toward exploring $MoS_2$ for energy conversion as a distinct metal catalyst. Among the 2H, 1T, and 1T′ phases, the most stable phase, which is the 2H phase, shows limited catalytic activityandexhibits different properties on its basal plane as compared to its edge sites. To address this limitation, a straightforward single-step hydrothermally driven process has been introduced to create porous $MoS_2$ nanosheets. This method increases the visibility of the edges of $MoS_2$, enabling the more convenient use of the active stable sites found at the periphery of $MoS_2$.[81]

The addition of small quantities of other elements to layered $MoS_2$ significantly enhances its catalytic performance.[82] Studies have shown that doping $MoS_2$ catalysts with minute concentrations of noble metals can effectively enhance their catalytic activity.[83] Interaction with different atoms is vital in modifying $MoS_2$ electronic structure, particularly in Mo, a d-band element. $MoS_2$ can be doped using both metal and non-metal. In metal doping procedures of $MoS_2$, various metals such as iron (Fe), nickel (Ni), platinum (Pt), gold (Au), cobalt (Co), and copper (Cu) have commonly been employed as effective doping agents. On the other hand, nonmetal doping approaches have extensively studied elements like nitrogen (N), phosphorus (P), and carbon (C), utilizing a range of different methods.[84,85]

$MoS_2$ demonstrates remarkable performance in various catalytic functions.[86] $MoS_2$ serves as a heterogeneous catalyst in various organic reactions, enabling the reduction of benzene derivatives in an environmentally friendly manner.[87] Due to their physiochemical properties, including mechanical strength, tunable porosity, chemical/thermal stability, and surface chemistry, $MoS_2$ is widely used in different catalytic reactions.[88] However, the choice of solvents and reagents is done specifically in heterogeneous catalysis. For example, the reduction of nitrobenzene requires potent $H_2$ sources, usually achieved through reducing agents.[89,90] Reduction reactions (RR) using eco-friendly solvents like methanol and water, along with environmentally friendly reducing agents like hydrazine hydrate ($N_2H_4.H_2O$), which produces only $H_2$ and $N_2$ gases as by-products, are preferred over traditional methods.[91] Additionally, protic solvents aid in effectively dispersing the reactant and enhancing

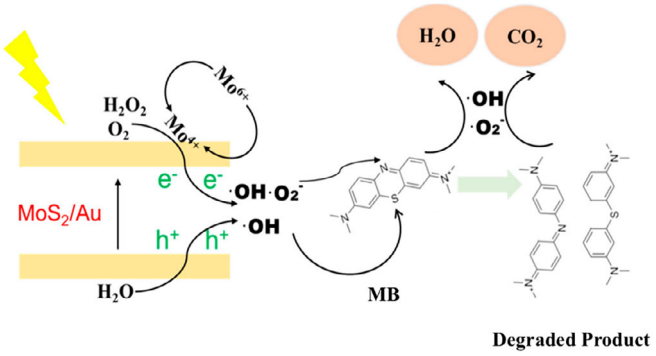



**Figure 7.** Diagrammatic representation of MB degradation using $MoS_2$/Au heterojunctions. Reproduced with permission from ref. [94].

the reaction with $H_2$.[92] $MoS_2$ being a heterogeneous catalyst has great potential in such types of reactions.

Dungey et al.[93] developed a gamma alumina ($\gamma$-$Al_2O_3$)-supported $MoS_2$ catalyst promoted by Co for hydrodesulfurization. Using the Co/Mo/S catalyst, the conversion of cyclopropyl methyl thiol to butene isomers was observed to occur through a free radical mechanism. Under 300 °C, the thiol was converted into a blend of butane isomers. Wei et al.[94] synthesized $MoS_2$/Au heterojunctions using a hydrothermal method for methylene blue (MB) degradation. Ultraviolet-visible (UV–vis) spectrum analysis revealed that optimal MB degradation reached 83% within the first 10 min of the reaction. MB degradation levels surpassed 95% at pH values of 1.40, 5.0, and 11.10 within 20 min, nearing complete degradation at 60 min, demonstrating the effective catalytic performance of $MoS_2$/Au. The C—N—C bond exhibited higher vulnerability to rupture compared to the C—S—C bond. The attachment of Au nanoparticles onto the $MoS_2$ surface enhanced the presence of interface imperfections, raising catalytic efficiency. In the absence of $MoS_2$/Au, no Fenton-like reaction occurred. However, when MB, Hydrogen peroxide ($H_2O_2$), and $MoS_2$ were combined, a reaction similar to the Fenton reaction was observed. Overall, the catalytic performance of $MoS_2$/Au surpassed that of $MoS_2$ alone. The catalytic reaction process was enhanced by the surface defects produced due to the heterojunction structure. Figure 7 represents the Fenton-like reaction mechanism of $MoS_2$/Au heterojunction.

Kaushik et al.[95] synthesized a nanocomposite aerogel of $MoS_2$ with reduced graphene oxide ($MoS_2$/rGO) using a hydrothermal-based method, comprising 3D-rGO and layered $MoS_2$ sheets. The $MoS_2$-rGO material functioned as a catalyst for different reduction reactions, including nitrobenzene, benzaldehyde, and benzonitrile-based derivatives. The fabrication process involved hydrothermal synthesis followed by freeze-drying, utilizing graphene oxide (GO) and $(NH_4)_2MoS_4$ as precursors. The reduction of the oxygen-containing entities attached to the surface of GO sheets was achieved. This resulted in the formation of an ultralight and 3D heterostructure consisting of the rGO aerogel decorated with $MoS_2$ sheets. The surface characteristics of GO were examined before and after $MoS_2$ integration through a comparative surface analysis. According to the Brunauer–Emmett–Teller (BET) isotherm, due to this composite formation, the volume of adsorption increased by approximately four-fold


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

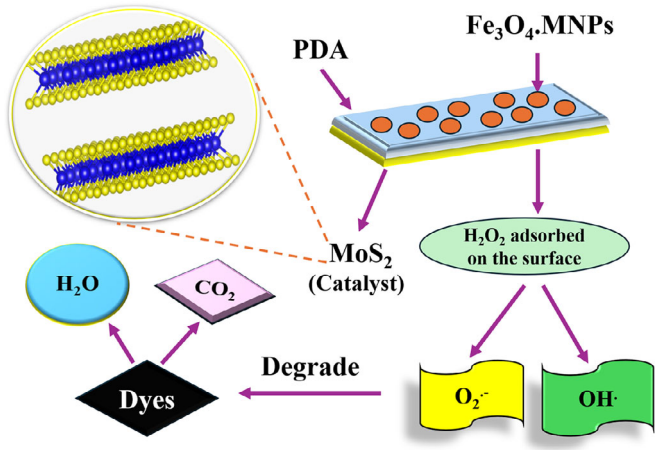



**Figure 8.** $H_2O_2$ adsorbs on the surface of $MoS_2$-$Fe_3O_4$ MNPs and degradation of dyes occurs with removing $H_2O$ and $CO_2$.

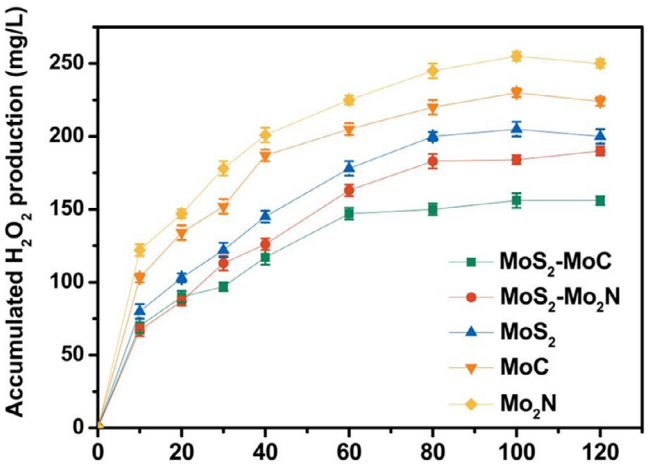



**Figure 9.** Accumulated $H_2O_2$ production on different catalytic cathodes. Reprinted with permission from ref. [97]. Copyright2022, American Chemical Society.

as the surface area of GO increased to around 748 $m^2g^{-1}$ from 148 $m^2$ $g^{-1}$. The comparison considers factors such as reaction conditions, percentage yield conversion, and the number of cycles. Shen et al.[96] successfully synthesized a magnetic composite of iron oxide ($Fe_3O_4$) and $MoS_2$ named $MoS_2$-$Fe_3O_4$ magnetic nanoparticles (MNPs), by employing a biomimetic technique. This approach depends on the self-polymerization of dopamine in a basic aqueous medium. $MoS_2$-$Fe_3O_4$ MNPs were utilized as catalysts in the degradation of organic dyes in the Fenton reaction. The effective dispersion of $Fe_3O_4$ MNPs on the $MoS_2$ sheets was facilitated by the bonding provided by polydopamine (PDA). This resulted in the prevention of $Fe_3O_4$ MNP aggregation during interaction and maintained continuous catalytic effectiveness. Catalytic experiments were conducted in a 100 mL round-bottom flask in the absence of light to avoid disruption from the spontaneous breakdown of $H_2O_2$. The reactions occurred at 30°C with magnetic stirring at a speed of 600 rpm. An aqueous solution was formulated by diluting organic dyes in water to a concentration of 1 g/L and adjusting the pH to a specified level for experimental use. The concentration of organic dye was maintained at 20 mg/L throughout the experiment, and the solution's pH was adjusted using sulphuric acid ($H_2SO_4$) or NaOH solution. A measured quantity of $MoS_2$-$Fe_3O_4$ MNPs was dissolved with a specific pH in 98 mL of distilled water. The degradation efficiency of catalysts was investigated using various concentrations ranging from 0.02 to 0.9 g/L, pH values from 3 to 11, and $H_2O_2$ concentrations from 2 to 90 mmol/L. The catalysts effectively facilitated the catalysis of $H_2O_2$, resulting in the production of superoxide free radicals ($O_2^{\cdot-}$) and hydroxyl free radicals ($OH^{\cdot}$) (Figure 8). The research work concluded that this process effectively degraded MB within a pH range of 3– 9, with the degradation efficiency approaching nearly 100%.

Li et al.[97] reported a heterogenous Fenton-like reaction, which involves the in situ electrochemical production of $H_2O_2$. This represents an environmentally friendly approach to the remediation of organic contaminants in waste water. Here, two types of catalytic cathodes, $MoS_2$-MoC and $MoS_2$-$Mo_2N$, featuring heterojunctions, were used. These catalytic cathodes exhibited an increased presence of S vacancies, enhancing the efficiency of paracetamol pollutant degradation. This enhancement is facilitated by the conversion of superoxide radicals through $Mo^{6+}$. The catalytic cathodes $MoS_2$-MoC and $MoS_2$-$Mo_2N$ displayed nearly five times greater activityin comparison to a pure $MoS_2$ catalytic cathode (Figure 9). The study provided a detailed understanding of the generation of hydroxyl radicals and singlet oxygen ($^1O_2$) in a chemical-free environment. It also sheds light on the development of Fenton-like catalysts for enhanced pollutant degradation rates. $^1O_2$ production is responsible for approximately 66% of the overall degradation.

Jayan et al.[98] synthesized a flower-like composite of zinc sulphide with $MoS_2$ (ZnS/$MoS_2$) with a highly porous structure using a one-step hydrothermal process. By adjusting the amount of Zn and Mo in the composite, the photocatalytic performance of the ZnS/$MoS_2$ composite was modified. The best $H_2$ generation activity was achieved with the composition containing a 1.0:1.30 ratio of Zn and Mo, along with 3% co-doping, producing 1256 µmol of $H_2$ in 1 h. This performance surpassed that of other compositions with different levels of co-doping (1.0%, 3.0%, and 5.0%). The mixed sulfide composite systems (ZnS/$MoS_2$ and ZnS/Co-doped $MoS_2$) enable enhanced $H_2$ generation when exposed to visible light for water-splitting. This happens without the aid of any sacrificial agents and this can be attributed to the optimized band gap and reduced electron-hole recombination rate or mixed sulfide composite systems.

Satheesh et al.[99] observed that the combination of $MoS_2$ and (silver phosphate) $Ag_3PO_4$ exhibits superior electrochemical properties for HER compared to when $MoS_2$ and $Ag_3PO_4$ are used individually. The synthesis of $Ag_3PO_4$ combined with exfoliated 2D layers of $MoS_2$ was carried out using the wet chemical hydrothermal method. The catalyst $Ag_3PO_4$ demonstrates effective degradation of pollutants when exposed to visible light due to its sensitivity to light and low bandgap energy of approximately 2.30 eV. Due to these properties, $Ag_3PO_4$ can be used for the photocatalytic production of $H_2$. By incorporating different weight percentages (1%, 3%, and 5%)


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

of $MoS_2$ into $Ag_3PO_4$, the HER activity was investigated. The combination of the 2D layered $MoS_2$ structure and nano-scaled $Ag_3PO_4$ demonstrates a synergistic effect. This combined effect leads to an increased electrochemically active surface area, resulting in faster reaction kinetics for the process of $H_2$ evolution. The electrochemical experiments took place in an acidic electrolyte solution (0.50 mol% $H_2SO_4$), indicating significant catalytic activity for $H_2$ evolution when utilizing 1.0wt% $MoS_2$ with $Ag_3PO_4$. The observed reduction in overpotential from 380 mV to 235 mV at a current density of 10 mA/cm$^2$ was particularly notable for the optimal $Ag_3PO_4/MoS_2$composite. Ibrahim et al.[100] reported a $CoFe_2O_4$@$MoS_2$ heterogeneous magnetic acid catalyst. The microwave-assisted method was used to observe its catalytic efficiency in the esterification of oleic acid. The synthesized $CoFe_2O_4$@$MoS_2$ catalyst displayed notable characteristics such as a high magnetization of 15.780 emu/g, a surface area of 6.700 m$^2$/g, and a substantial loading of acid around 2.10 mmol $H^+$/g. The optimal conditions for achieving the highest oleic acid conversion rate of 98.2% were determined (7.50 wt% catalyst amount, 15:1 molar ratio for methanol and oleic acid respectively, 2 h for reaction, and a specific temperature of 140 °C). The comparison of conventional and microwave heating methods showed that microwave heating led to a significant decrease in the reaction time, from 4 h to 2 h. Furthermore, the catalytic performance of the $CoFe_2O_4$@$MoS_2$ catalyst remained relatively stable even after eight cycles of use, indicating its robustness and recyclability.

Zheng et al.[101] created flower-shaped $MoS_2$, which is produced using a combination of the hydrothermal method and a reduction process. By adjusting the conditions of $H_2$ annealing, the amount of Lewis and Bronsted acidic sites, which come from the exposed Mo edges of $MoS_2$, can be effectively controlled. The $MoS_2$-450-R catalyst exhibits notable effectiveness in transforming 5-hydroxymethyl furfural (5-HMF) into 5-(methoxymethyl)furanal dimethyl acetal (5-MFDMA) through a tandem acetalization and etherification process in a nitrogen ($N_2$) atmosphere, achieving a conversion rate of 99% with a selectivity of 83.7% for 5-MFDMA. Plentiful acidic sites present in $MoS_2$-450-R are essential for the reaction between 5-HMF and methanol. The etherification of 5-HMF through specific catalystsis regarded as a viable method for producing biofuels from biomass resources. Huang et al.[102] reported the reduction of nitro aromatics like nitrobenzene and halogenated nitrobenzenes using $MoS_2$ and hydrazine ($N_2H_4$). They demonstrated the selective hydrogenation of halide nitrobenzenes without dehalogenation in the presence of $MoS_2$, thus exhibiting its exemplary performance as a catalyst.

Salnikov et al.[103] reported the parahydrogen-induced polarization technique (PHIP) for thiophene hydrodesulfurization by using the $MoS_2/\gamma$-$Al_2O_3$ catalyst. Bahuguna et al.[104] synthesized a $MoS_2$-RGO nanocomposite that has been identified as a strongly effective catalyst with a heterogeneous nature for the production ofindole-based alkaloids in aqueous environments. This catalyst can be reused up to six times without any prominent decrease in its catalytic strength, showing a high economy for atoms (94.70%) and a very low E-factor of 0.0890 for the reaction.

### 3.1. $MoS_2$ as an Electrocatalyst

An electrocatalyst refers to a material that enables the conversion of chemical energy into electrical energy through electrochemical processes. Electrocatalytic materials enhance and accelerate electrochemical reactions without being used in the process by interacting with reagents to modify reaction pathways and reduce the activation barrier. Electrochemical reactions are influenced by various factors like the electric field at the electrode or solution interface, temperature, pressure, medium of solution, and condition of solid surface. The reaction rate is also influenced by the composition of the electrode's double-layer structure, emphasizing the importance of modifying electrocatalysts on the electrode surface to control the electrochemical reaction rate. There are two types of electrocatalysts, one is heterogeneous electrocatalysts, like platinized electrodes, which exhibit a nonuniform composition.[105] On the other hand, we have homogeneous electrocatalysts, which are soluble in the reaction solution and aid in the transfer of electrons between the reactants and electrodes. $MoS_2$ exhibits great potential as an electrocatalyst for the HER owing to its remarkable electrochemical activity with exceptional catalytic properties.[106]

The excessive utilization of fossil fuels has resulted in various ecological issues and has also contributed to the depletion of energy resources.[107,108] The advancement of environment-friendly and non-polluting energy has emerged as a significant objective in modern scientific investigations.[22] $H_2$ energy, being a clean, effective, and sustainable energy source, holds immense potential for various applications.[109] Currently, the primary methods of $H_2$ production involve coal gasification and methane steam reforming and both of these methods rely on fossil fuels as raw materials. These methods also release $CO_2$ during the production process, thereby causing environmental contamination.[110] Extensive efforts have been undertaken to produce $H_2$ gas in a sustainable and carbon-neutral way for a wide range of industrial applications. Electrocatalytic HER plays a crucial role in the production of $H_2$ for clean energy applications. Unlike conventional $H_2$ production methods, electrocatalytic water electrolysis offers a sustainable alternative.[111] This process directly utilizes electrical energy to split water molecules into $H_2$ and $O_2$, making it an effective means of energy conversion and storage, particularly when powered by renewable energy sources (RES). The environmental impact of electrocatalytic $H_2$ production is directly linked to the source of electricity used. When coupled with RES, this approach provides a virtually zero-carbon pathway to $H_2$ generation, making it an essential technology for decarbonizing industries, energy storage, and the broader transition toward a green economy.

Although $MoS_2$ is a promising electrocatalyst, its catalytic efficiency is limited by inherent challenges. Both experimental and theoretical studies have identified several factors that hinder its performance as an electrocatalyst. The primary reasons for the low electrocatalytic efficiency of $MoS_2$ are as follows:[107]

1. Limited active edge sites: $MoS_2$ is only active at its edges, while the flat surfaces do not contribute to catalysis. In bulk form, fewer edge sites limit its efficiency. This can be improved by adding defects, doping, or making thinner layers.[112]


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

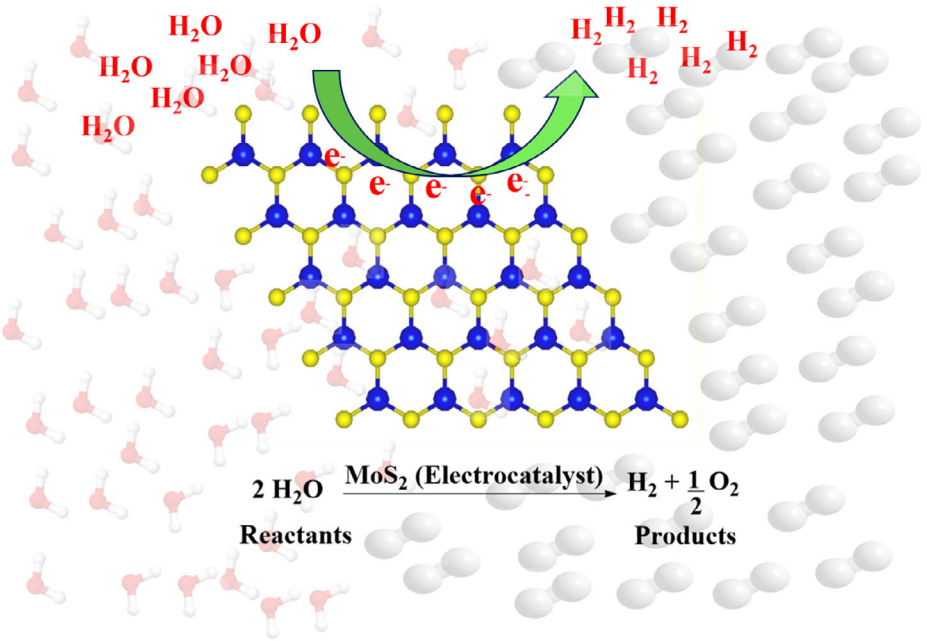



**Figure 10.** Water adsorbs on the $MoS_2$ catalyst during HER to evolve $H_2$ gas molecules.

2. Low intrinsic conductivity of the basal plane: $MoS_2$ in its 2H phase has low conductivity, limiting electron transfer and increasing overpotentials in electrocatalysis. The 1T phase offers better conductivity but is unstable and reverts to 2H over time.[113]
3. High overpotential: $MoS_2$ needs high overpotentials for electrocatalysis, especially in HER, because its $H_2$ adsorption energy is not ideal, lowering efficiency.[114]

These issues limit charge transfer and hydrogen adsorption, lowering its catalytic efficiency.[107,115]

Liu et al.[116] developed a steel mesh platform composed of $MoS_2$ functionalized with nano-sized nickel (Ni-$MoS_2$/SM). This platform showcased potential in facilitating the sulfide oxidation reaction and aiding in energy-saving production. Ni-$MoS_2$/SM demonstrated enhanced adsorption of polysulfides and facilitated the kinetics of HER by absorbing hydroxide molecule ($OH_{ad}$). HER was studied in a standard three-electrode system, a 1 M KOH aqueous solution was employed, with the self-supported electrode acting as the working electrode. The enhancement of the Ni-$MoS_2$/SM catalyst was measured by the manipulation of the electrodeposition duration. Furthermore, the $H_2$ production efficiency of the Ni/$MoS_2$-based system, combined with the sulfide oxidation reaction, reached nearly 100%, consuming 61% less power per kilowatt-hour (kW-h) compared to the oxygen evolution reaction (OER) combined with the HER system. The catalytic activity of single $MoS_2$ and Ni is limited due to the intermediates ($S_2$* and S*), excessively weak adsorption of $S_2$* onto $MoS_2$, and the excessively strong adsorption of S* and $S_2$* onto Ni. The adsorption of H and OH species on different sites of Mo and Ni is crucial in enhancing the catalytic activity. Ni, not only improved the adsorption of hydroxyl species but also prevented the recombination of $OH_{ad}$ with hydrogen atom ($H_{ad}$) synergistically increasing the HER rate. Figure 10 shows the adsorption of water molecules on the surface of $MoS_2$ to produce $H_2$ thus depicting HER reaction. In conclusion, Ni-$MoS_2$ and innovative $H_2$ generation systems hold great potential for sustainable energy generation.

Chen et al.[81] reported a one-step hydrothermal-driven process for synthesizing a series of $MoS_2$ samples for comparative study, by varying the molar ratio of Mo to S. The catalytic efficiency of the synthesized array of 2D-$MoS_2$ was confirmed through electrochemical assessments. An established three-electrode configuration was employed to study the HER of this series of catalysts in 0.5 M $H_2SO_4$, as well as the OER reaction performance in 1 M KOH. Within the range of 2D $MoS_2$ catalysts, the porous $MoS_2$ nanoislands exhibited a significant advantage in terms of the production of $H_2$. These islands can achieve a 40.00 mA $cm^{-2}$ current density at an overpotential of approximately 320.00 mV. The porous nanoislands of $MoS_2$ significantly enhanced both the OER and HER activities of $MoS_2$, because of maximum exposure of the edges of $MoS_2$ grains, which contain active stable sites suitable for various electrocatalytic reactions. Specifically, the porous $MoS_2$ nanoislands exhibited current densities of 10.0 mA $cm^{-2}$ in both OER and HER at relatively low overpotentials of 248.00 and 300.00 mV, respectively, surpassing the performance of typical nanosheet morphology. This finding suggests that the porous $MoS_2$ nanoislands can be a promising alternative to precious metal catalysts for achieving efficient and durable bifunctional HER/OER.

Zheng et al.[117] synthesized a nanofoam (NF) catalyst, using Co, selenium (Se), and $MoS_2$(Co/Se-$MoS_2$-NF). In the NF, Co atoms located in the inner layer activate neighboring S atoms, while Se atoms bound on the surface enhance the stability of the catalyst's structure. The combined effect of Co-doping within the inner layer and Se-doping on the surface leads to the elevated level of $H_2$ adsorption activity due to the presence of numer-


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

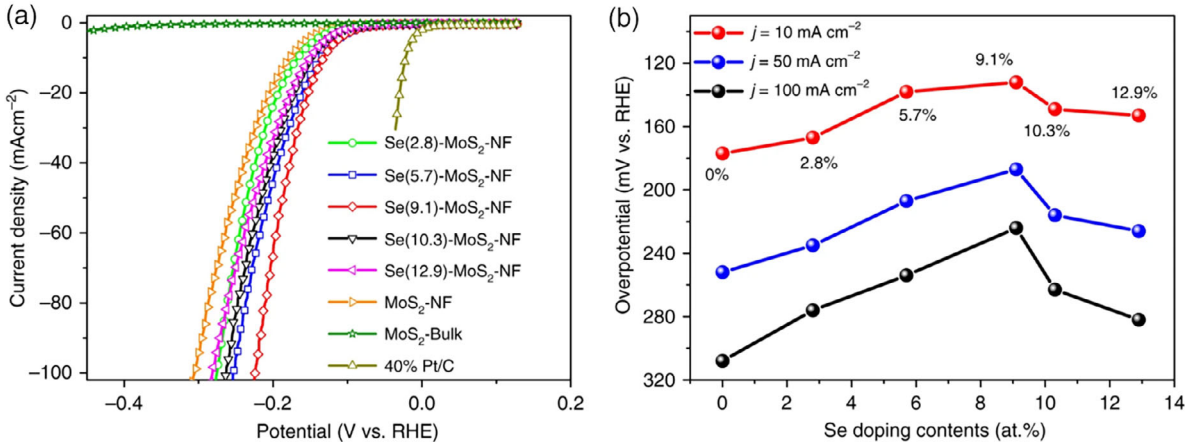



**Figure 11.** (a) The HER polarization curve for Se-$MoS_2$-NF with varying Se doping levels was compared to those of bulk $MoS_2$, $MoS_2$-NF, and 40% Pt/C. (b) The overpotentials exhibit variations based on thecontent of Se doping at 10, 50, and 100 mA cm$^{-2}$ current densities, respectively. Reproduced with permission from ref. [117].

ous active sites occurring on both the flat surfaces and the boundaries. The Se-doped sample containing 9.1% Se (Se(9.1)-$MoS_2$-NF) demonstrates superior HER performance, showcasing overpotentials of 132, 187, and 224 mV at 10, 50, and 100 mA cm$^{-2}$ current densities, respectively. This augmentation results from the increased number of active sites, which enhances the overall $H_2$ adsorption capacity by incorporating both Co-doping and Se-doping. Co/Se-$MoS_2$-NF composites demonstrate exceptional HER performance and durability even under large current densities Specifically, at 208 mA cm$^{-2}$, the overpotential is reduced compared to a 40 wt.% Pt/C catalyst (Figure 11). Moreover, at a significantly larger current density of 1000 mA cm$^{-2}$, the overpotential is notably less than that of the Pt/C catalyst. The faradic efficiency for $H_2$ production remains close to 100% across various working potentials, suggesting minimal generation of other reduction by-products. This approach represents a promising and viable pathway for the development of cost-effective and high-performance catalysts in the large-scale production of clean energy in the form of $H_2$.

Gu et al.[118] introduced a heterojunction of cobalt sulfide ($CoS_2$) with $MoS_2$ ($MoS_2$@$CoS_2$) which showed remarkable $H_2$ evolution activity across a wide pH range. Notably, it only requires an overpotential of 48, 62, and 164 mV to obtain a 10.0 mA cm$^{-2}$ current density in KOH (1 M), $H_2SO_4$(0.5 M), and phosphate-buffered saline (PBS) aqueous solutions (1 M), respectively (Figure 12). The Tafel slopes obtained for each case are 43.0, 32.0, and 62.0 mV dec$^{-1}$, respectively. This research centered on examining how the combination of $MoS_2$ and $CoS_2$ enhances electron transfer, leading to exceptional electrocatalytic performance of the $MoS_2$@$CoS_2$ heterojunction. Additionally, the observed catalytic ability is not only influenced by the combined effects of the heterojunction and sulphur vacancy but also offers a useful model and reference for investigating other efficient electrocatalysts.

Chen et al.[119] presented the detailed synthesis of structured, macroporous $MoS_2$–$Fe_2Mo_3O_8$/$MoO_2$ electrocatalysts with heterogeneous compositions. The resulting structures provide an increased number of active sites and promote ion diffusion, leading to a significant enhancement in overall performance. The catalyst exhibited remarkable efficiency in both OER and HER in a 1 M solution of KOH, with overpotentials of 213 and 60 mV, respectively, to achieve10 mA cm$^{-2}$ current density. Furthermore, the electrolyzer based on $MoS_2$-$Fe_2Mo_3O_8$/$MoO_2$ operated at an output voltage of just 1.49 V, exhibiting excellent performance and long-term stability in comparison to catalysts available in the market.

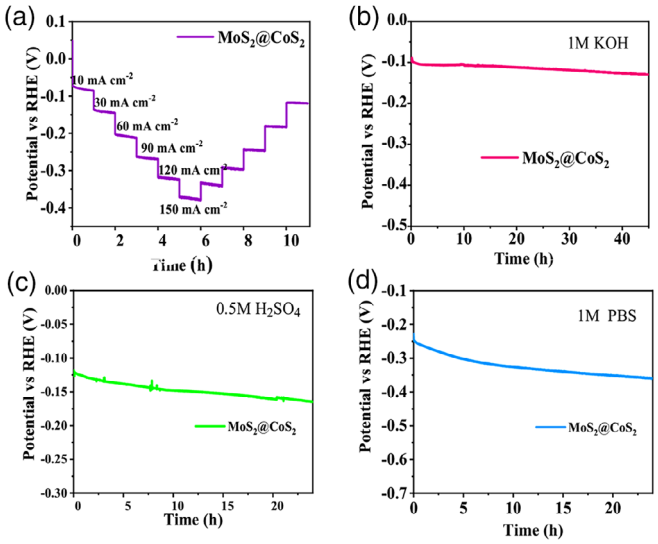



**Figure 12.** (a) Stability curves of $MoS_2$@$CoS_2$ measured at different current densities in 1 M KOH. The time-dependent current density curves of $MoS_2$@$CoS_2$ heterostructure (b) in 1 M KOH, (c) in 0.5 M $H_2SO_4$, and (d) in 1 M PBS. Reprinted with permission from ref. [118]. Copyright, 2024 American Chemical Society.

Askari et al.[120] reported the synthesis of a nano-electrocatalyst, $Fe_3O_4$@$MoS_2$/rGO, involving a two-step hydrothermal process that combined magnetite-coated $MoS_2$ with rGO. This catalyst demonstrated promising potential as a viable alternative to Pt catalysts in facilitating methanol oxidation and HER. The presence of Fe in anionic sites within the $Fe_3O_4$ structure plays a dual role in nanocatalysts by enhancing


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

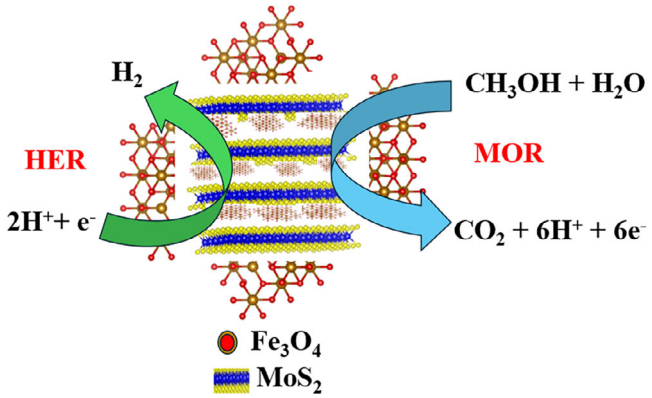



**Figure 13.** MOR reaction takes place at the surface of the electrocatalyst-$Fe_3O_4$@$MoS_2$/RGO and $CO_2$ is released.

the adsorption of methanol and promoting the desorption of oxidation by-products. Effective binding of methanol to the catalyst helps in the availability of numerous adjacent sites. The introduction of a secondary transition metal Mo facilitates the binding of $H_2O$ to methanol, resulting in $CO_2$ production. This is called a methanol oxidation reaction (MOR) and is shown in Figure 13. To assess the performance of the HER process, the electrode's surface was coated separately with $Fe_3O_4$@$MoS_2$/RGO. The findings indicated a decrease in overpotential by 108.0 mV and a Tafel slope of 48.0 mV $dec^{-1}$, facilitating the HER under acidic conditions. This phenomenon is credited to the collaborative impact of $Fe_3O_4$ and $MoS_2$, as well as the enhancement of the transfer rate of electrons by the incorporation of conductive rGO into the catalyst. These outcomes suggest that the synthesized nanocomposite shows promise for use in $H_2$ evolution and methanol oxidation reactions as well.

### 3.2. $MoS_2$ as a Photocatalyst

Photocatalyst is derived from the combination of the words, "photo" referring to light, and "catalyst". A photocatalyst is a substance that initiates chemical reactions by absorbing photons or light. Semiconducting oxide photocatalysts have garnered notable emphasis due to their potential utilization in environmental purification and solar energy conversion.[121] $MoS_2$ has sparked considerable interest due to its various advantageous photocatalytic properties. It possesses a narrow band gap, exceptional optical absorptivity, a stratified composition, high charge carrier mobility, effective transformation of photons into electrons, and low toxicity.[122] The photocatalytic properties of $MoS_2$ can further be enhanced by various methods such as changing morphology, adjusting energy band gap through doping,[123] and creating heterojunctions for band alignment (Figure 14). Collectively, these properties along with the possibility of modifications by various methods establish $MoS_2$ as an exemplary photocatalyst.[124] For example, heterojunction composites of $MoS_2$ and titanium dioxide ($TiO_2$), which have a narrow band gap, exhibited outstanding photocatalytic efficiency in the degradation of methyl orange (MO) and 4-nitrophenol (4-NP) contaminants, when exposed to visible light.[125] While, $MoS_2$ and cadmium sulfide (CdS), which are semiconductor-based photocatalysts, have garnered significant interest for purifying water due to their distinct features such as adjustable bandgap, strong absorption of visible light, and impressive electronic-charge transfer ability.[126]

The process of photocatalytic water splitting is considered a feasible approach for generating eco-friendly and renewable energy, as it can transform solar energy into $H_2$ energy. $MoS_2$ is an excellent photocatalyst in various chemical reactions due to its exceptional responsiveness to solar spectra and its impressive electrical and optical attributes.[127] The incorporation of $MoS_2$ with materials such as graphitic carbon nitride (g-$C_3N_4$), $TiO_2$, and CdS has been found to enhance the photocatalytic properties. Thus, facilitating the efficient production of $H_2$ through the splitting of water.[128]

Upon the exposure of $MoS_2$ to light, the photocatalytic redox reactions commence with the creation of photo-excited charge carriers. This simultaneous generation results in the production of electrons ($e^-$) and holes ($h^+$) in the material. Following photoexcitation, during this process, the electrons are elevated to the conduction band (CB), while the holes remain confined within the valence band (VB). These photogenerated carriers then move toward the $MoS_2$ surface, engaging in reduction and oxidation reactions. Photoexcited electrons may interact with adsorbed $O_2$, resulting in the formation of $O_2^{\cdot-}$ radicals. On the other hand, the remaining $h^+$ in the valence band can extract an $e^-$from hydroxyl ions or adsorbed $H_2O$ molecules, resulting in the production of $OH^{\cdot}$ Radicals (Figure 15). These $OH^{\cdot}$Radicals contain potent oxidizing properties and can effectively react with harmful organic, inorganic, and biological compounds.[129] So it acts as an excellent photocatalyst in decomposing organic pollutants and in wastewater treatment.[130]

Hu et al.[131] fabricated a 2D Co-doped $MoS_2$ nanojunction with 2D g-$C_3N_4$ and named it as Co-doped $MoS_2$/CN with enhanced electrical conductivity using a hydrothermal approach. The Co-doped $MoS_2$/CN nano-junction exhibited a remarkable photocatalytic performance, displaying a benzaldehyde production rate of 0.480 mmol $g^{-1}$ $h^{-1}$ through the oxidation of benzyl alcohol, and an $H_2$ production rate of 0.310 mmol $g^{-1}$ $h^{-1}$ through the reduction of $H_2O$. The photoexcited electrons from the CN material efficiently migrated through the rapid pathway of a particular bond (N—Mo—S—Co) toward the Co-doped $MoS_2$ co-catalyst, thereby facilitating the separation of charges. This significant enhancement in photocatalytic activity can be attributed to several factors. Firstly, the Co-doped $MoS_2$ possesses an edge-enriched 1T phase, which acts as a cocatalyst and contributes to its higher metallic characteristics. Additionally, compared to the pure $MoS_2$, the Co-doped $MoS_2$/CN nano-junction exhibits lower $H^*$ adsorption and provides more active sites, further enhancing its photocatalytic performance.

Chou et al.[132] developed an exceptionally effective nanocatalyst of Au-$MoS_2$, which response to both ambient mechanical vibration and light irradiation. This nanocatalyst possesses adjustable photocatalytic properties, allowing it to produce reactive oxygen species (ROS) resulting in exceptional disinfection


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

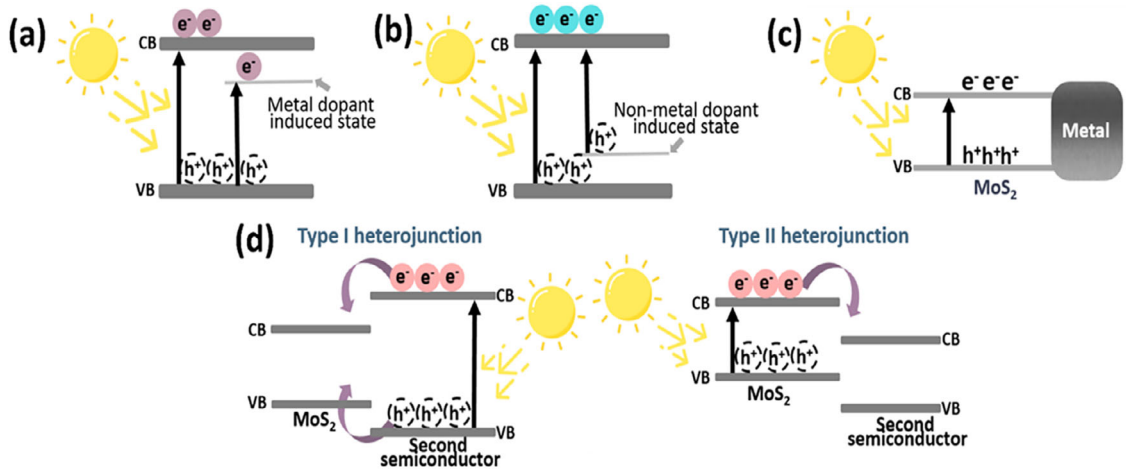



**Figure 14.** Various ways of modifying $MoS_2$ to enhance its photocatalytic behavior (a) metal-doped $MoS_2$, (b) nonmetal-doped $MoS_2$, (c) metal deposition on $MoS_2$, and (d) the establishment of a heterojunction by combining with a second semiconductor. Reproduced with permission from ref. [123].

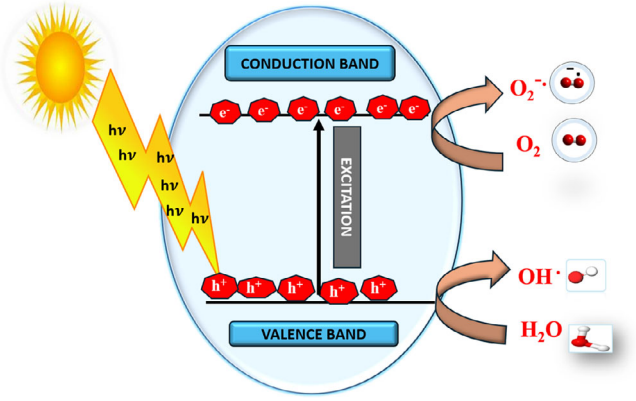



**Figure 15.** A schematic representation of the photocatalytic mechanism of $MoS_2$.

capabilities. $MoS_2$ nanoparticles (NPs) exhibit the feasibility of mass production, and they were grown uniformly on carbon fiber (CF) cloth as a substrate for catalytic reactions. Furthermore, the catalytic efficiency of $MoS_2$NPs was enhanced by the application of Au NPs, extending the lifetime of electron-hole pairs (EHP) and boosting ROS generation. ROS production was evaluated through fluorescence and absorbance measurements. The degradation of the Gram-negative bacteria (*E. coli*) colony system by 99.999% validated the disinfection performance of Au-$MoS_2$NPs. The water electrolysis process is primarily driven by charge transfer, leading to the generation of ROS. Subsequently, these ROS interact with bacteria, leading to disinfection and purification of water.

Kisala et al.[133] reported that $MoS_2$, acting as a photocatalyst, possesses a layered structure and acts as a semiconducting material. The photocatalytic performance of $MoS_2$ is accessed through the degradation of bromophenol blue (BPB) when exposed to a very low acidic aqueous solution with a pH of 5.20. In this specific environment, BPB exists in a fully dissociated state in the solution, with the I$^{st}$ and 2$^{nd}$ acid dissociation constants (pKa) being 3.00 and 4.60, respectively. At pH of 5.20, the simulated molar fraction of ionic forms indicates that dianion ($BPB^{2-}$) is present at 0.979, anion ($BPB^{-}$) at 0.021, and BPB at 0.00. The efficiency of the catalytic process is influenced by the interactions between the catalyst and the substrate. Under reaction conditions, the $BPB^{-}$ and $BPB^{2-}$ may be found near the surface of $MoS_2$, potentially near its metallic Mo-built edge. This proximity favors the $e^{-}$transfer between the surface and the substrate, leading to the reductive debromination of $BPB^{2-}$, which signifies the catalytic capability of $MoS_2$. The photocatalytic degradation of BPB dye is shown in Figure 16a. This photocatalytic degradation of BPB using $MoS_2$ was monitored by recording changes in dye concentration over time as shown. Figure 16b shows that the depletion of BPB concentration was primarily due to the photocatalytic activity of the synthesized $MoS_2$ rather than direct photolysis. The kinetic analysis, represented by a plot of $\ln(C/C_n)$ versus time, resulted in a straight line, indicating a first-order reaction where the apparent rate constant ($\mathbf{k_{app}}$) corresponds to the slope, Figure 16c. The observed debromination suggests a reductive degradation pathway, with the release of bromide over time depicted in Figure 16d. The accelerated degradation confirms highly efficient reductive debromination, with a rate constant of $k_{app} = 457.4 \times 10^{-3}$ $min^{-1}$ (Figure 16e). Additionally, the reusability of $MoS_2$ as a photocatalyst was assessed through four successive runs, demonstrating consistent catalytic performance (Figure 16f), highlighting its potential for practical applications. Even after undergoing four cycles, the efficacy of $MoS_2$ as a photocatalyst for reduction processes remained remarkably high at 89.9%.

Lee et al.[134] successfully developed a $MoS_2$/graphene hybrid (MSG) using the microwave hydrothermal technique. The photocurrent and the charge transfer resistance were adjusted by enhancing the photocatalytic nature of $MoS_2$ through graphene modification. Furthermore, optimal conditions for this process include a formic acid concentration of 0.32 M and a dosage of 0.013 g $L^{-1}$ of the $MoS_2$ along with 0.8wt% graphene ($MSG_{0.8}$). Under these conditions, the outstanding photocatalytic $H_2$ evolution is completed, resulting in the highest yield of $H_2$ being 667.2 $mmolh^{-1}g^{-1}$. Upon the introduction of graphene, the $MoS_2$ valence band maximum (VBM) experiences a shift from its initial


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

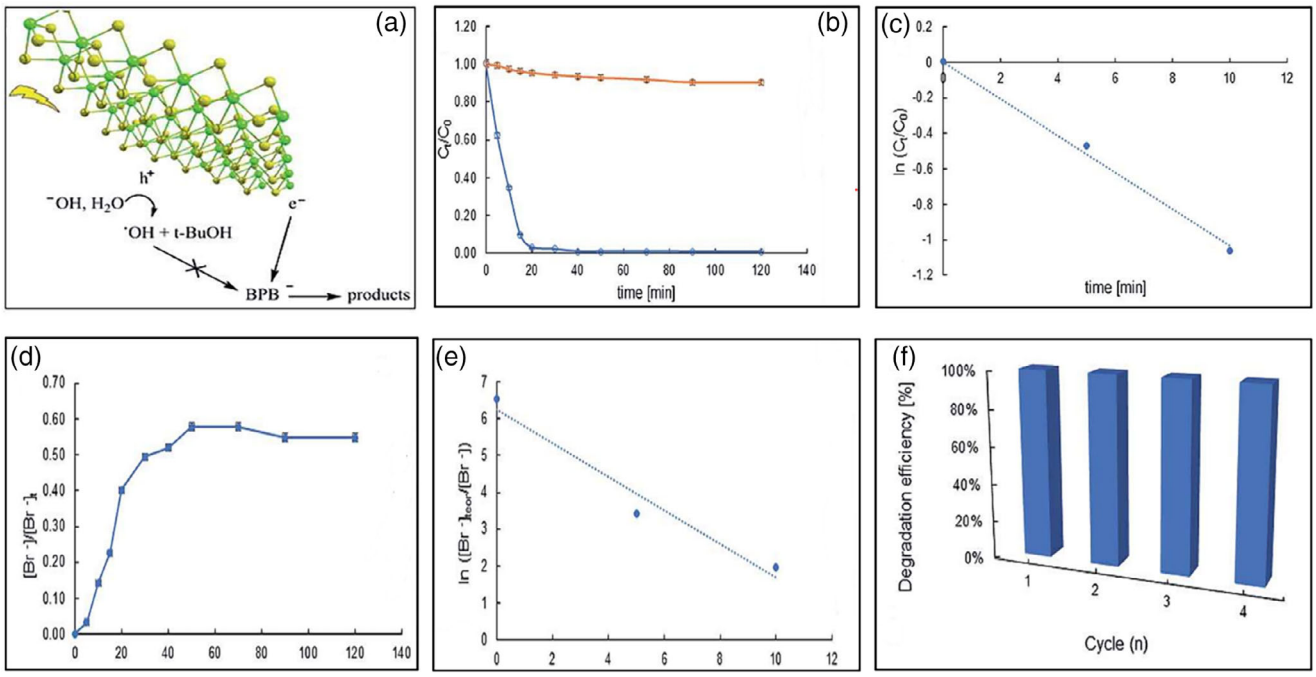



**Figure 16.** (a) Schematic description of BPB photocatalytic degradation. (b)BPB decay in time in the photocatalytic process (blue) and photolysis (orange). c)$k_{app}$ determination. (d) Bromide production in time. (e) Apparent rate constant ($k_{app}$) for bromide generation determination. (f) Reusability study of catalyst $MoS_2$. Reproduced with permission from ref. [133].

position at 0.66 eV to 0.580 eV. This shift implies the existence of S defects within the $MoS_2$/graphene hybrid. As a result, the VBM moves closer to the Fermi level, leading to a decrease in the band gap. Therefore the modification of graphene enables the MSG to achieve the highest yield of $H_2$ gas.

Furthermore, factors such as oxidation potential, the concentration of sacrificial reagents, and the dielectric constant can influence photocatalytic activity. Zhou et al.[135] used a hydrothermal method to grow $MoS_2$ directly on modified $MoO_3$, resulting in a photocatalytic film capable of eliminating bacteria and formaldehyde across the full spectrum of light. The substrate used was waste eggshells, with $MoO_3$@$MoS_2$ serving as the photocatalyst. The resulting film exhibited exceptional antibacterial properties against both Gram-negative (*E. coli*) and Gram-positive (S. *aureus*) bacteria, as well as significant efficiency in formaldehyde removal. The electron-hole pairs generated by $MoO_3$@$MoS_2$ enhanced the photocatalytic capabilities of the film. By leveraging the photothermal effect and the increased surface area of discarded eggshells, the film demonstrated nearly 99.90% antimicrobial efficacy against E. coli and S. aureus within 10 min (Figure 17a,b). Moreover, it achieved a formaldehyde removal rate of approximately 92.5% within 30 min. Therefore, this innovative photocatalytic film presented a promising approach for utilizing waste bioresources and developing environment-friendly films.

Kar et al.[136] synthesised a composite $MoS_2$-CdS (as shown in Figure 18) which displayed an outstanding photocatalytic ability to degrade potassium dichromate [Cr (VI)] by reduction (Figure 18). A high degradation efficiency of 91.9 % and increase in degradation rate constant up to 2.88-fold was obtained by using 10 wt% $MoS_2$ - CdS composite in the presence of light. The degradation efficiency of 10 wt% $MoS_2$-CdS composite was 2.09-fold more than that of controlled CdS. Additionally, the 10wt% $MoS_2$-CdS composite exhibited the highest apparent quantum yield, which was 2.15 times greater than that of controlled CdS. Furthermore, the electrical power consumption per order for Cr(VI) decreased from 25.74 kWh $m^{-3}$ for the controlled CdS to 9 kWh $m^{-3}$ for the 10 wt% $MoS_2$-CdS composite. This reduction suggests an optimal synergy between the components of the $MoS_2$-CdS composite. The transfer of photogenerated electrons from the CdS to $MoS_2$decreased the chances of photogenerated charge carriers recombining, thereby enhancing the photodegradation capability of the composite. This study has the potential to offer important knowledge about creating and synthesizing bimetallic sulfide photocatalysts rationally and straightforwardly. These catalysts could demonstrate outstanding performance and cost-effectiveness in reduction of Cr (VI) (Figure 19).

### 3.3. $MoS_2$ as a Cocatalyst

A cocatalyst refers to a substance that can modify the effectiveness or specificity of a primary catalyst in a chemical reaction. The cocatalysts find extensive application in diverse domains, including photocatalysis, electrocatalysis, etc. $MoS_2$ acts as an excellent cocatalyst to enhance the catalytic properties of various catalysts by modifying the surface area of adsorption, the ROS production mechanism, and the microscopic process change through its active sites.[137] OER consumes 90% of the driving energy due to its high kinetic barrier and complex four-electron transfer process.[138] This reaction is crucial in various


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

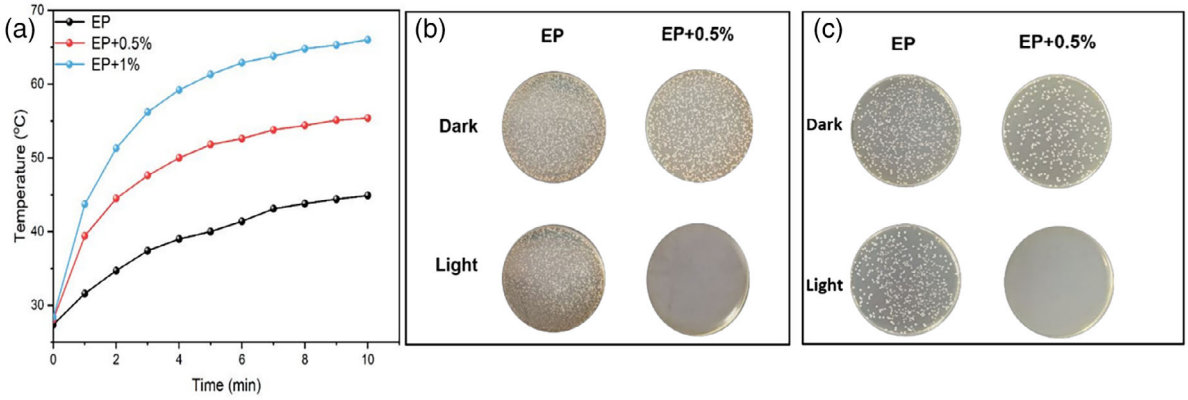



**Figure 17.** (a) Simulation of sunlight irradiation for 10 min to generate photothermal curves for various coatings with eggshell powder (EP), (b) the spread plate images of *S. aureus* of different coatings with and without simulation of sunlight irradiation for 10 min, and (c) the spread plate images of *E. coli* of different coatings with and without simulation of sunlight irradiation for 10 min. Reprinted with permission from ref. [135]. Copyright, 2024 American Chemical Society.

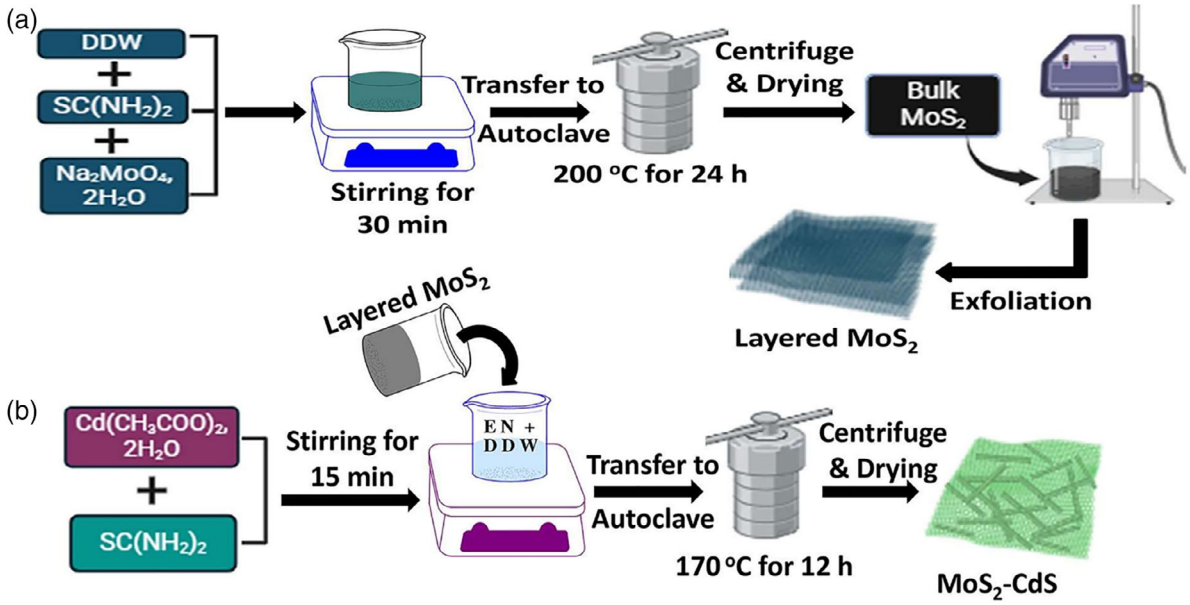



**Figure 18.** (a) Solvothermal synthesis of layered $MoS_2$ by dissolving the precursors Mo and S in double-distilled water (DDW). (b) Synthesis of $MoS_2$-CdS composite by adding layered $MoS_2$ with the precursors of Cd. Reproduced with permission from ref. [136].

systems such as $CO_2$ and $N_2$ reduction, metal-air batteries, and other energy conversion systems.[139] The unsustainable use of current noble metal oxide electrocatalysts for OER poses a significant challenge for commercial applications. Therefore, the advancement of OER catalysts, which are abundantly available on the earth and exhibit exceptional activity and stability is essential for sustainable energy conversion. Transition metals and their derivatives have garnered significant interest as OER catalysts due to the strong coupling properties arising from the *d*-orbital of the metal interacting with the 2p orbital of O-containing intermediates. Isolated single-metal sites supported by vibrant materials exhibit well-defined coordination and confinement states. These states have the potential to enhance the intrinsic activity and maximize atom utilization in catalysts.[140]

Solomon et al.[141] designed a nanostructure ($NiMoO_4/MoS_2$) in which $NiMoO_4$ nanorods were enveloped by active $MoS_2$ nanosheets on a nickel foam (NF) substrate. This nanostructure demonstrates exceptional OER performance, generating 10.00 mA $cm^{-2}$ current density at an overpotential of 90.0 mV. The superior performance of $MoS_2$ nanosheets is a result of the morphological and structural properties like surface-active sites and a strong interface with the electrolyte etc. Moreover, the unique branched network of the nanostructures enables effective charge storage, providing a readily accessible active surface area that facilitates the adsorption of intermediates. Notably, the $NiMoO_4/MoS_2$ combination exhibits 358.00 $mAhg^{-1}$ charge capacity with current of 0.500 A $g^{-1}$, compared to 230 $mAhg^{-1}$ for $NiMoO_4$ alone. This indicates a true potential for applications in devices of charge storing. Additionally, the presence of $MoS_2$-decorated nanosheet layers enables the storage of charges and facilitates a faradaic charge transfer mechanism as shown in Figure 20.


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

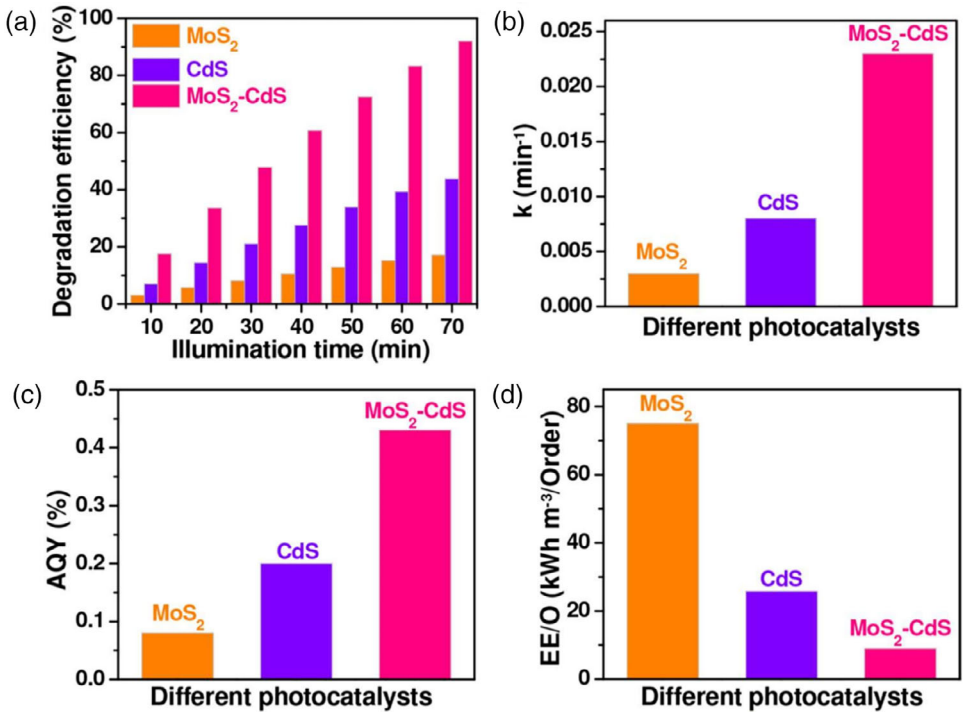



**Figure 19.** (a) Illustrates the relationship between degradation efficiency and illumination time for $MoS_2$, CdS, and $MoS_2$-CdS. (b) Presents the K values. (c) Presents the AQY values, and (d) presents EE/O values. Reproduced with permission from ref. [136].

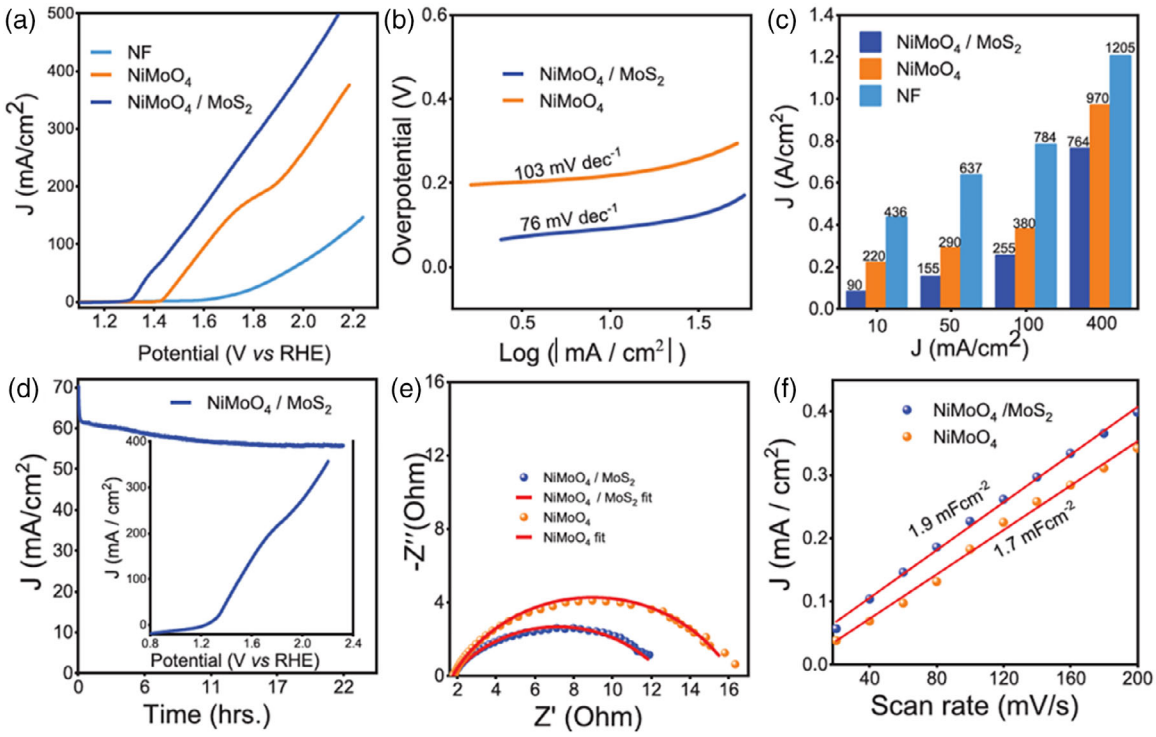



**Figure 20.** (a) The polarization curves of $NiMoO_4/MoS_2$, $NiMoO_4$, and NF. (b) Tafel plots of the catalysts ($NiMoO_4/MoS_2$, $NiMoO_4$), (c) Bar graph showing the overpotential of $NiMoO_4/MoS_2$, $NiMoO_4$, and NF at different current densities. d) Long-term stability study of $NiMoO_4/MoS_2$ using chronoamperometry, the linear sweep voltammetry (LSV) curve in the inset depicts the decreasing current density as measured in the first few hours. (e) The Nyquist plots for $NiMoO_4/MoS_2$, $NiMoO_4$ as obtained at 1.45 V. (f) Double layer capacitance ($C_{dl}$) was determined from cyclic voltammetry (CV) for $NiMoO_4/MoS_2$ and $NiMoO_4$. Reproduced with permission from ref. [141].


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

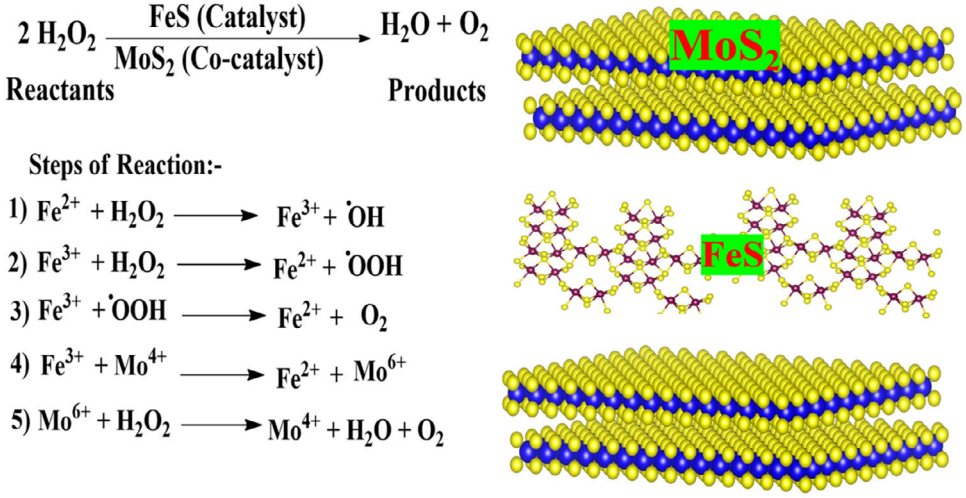



**Figure 21.** Reaction and mechanism of Fenton reaction occurring at the surface of $MoS_2$/FeS composite.

Yan et al.[142] noted that $Zn_3In_2S_6$ exhibited a reduced number of active sites and lower efficiency in separating photoproduced carriers, which hindered its further practical use. To tackle this problem, they introduced $MoS_2$ into $Zn_3In_2S_6$ to form 1T-$MoS_2$/$Zn_3In_2S_6$ (M/ZIS6), spherical catalysts (flower-shaped) with plentiful S vacancies. The 2.4% M/ZIS6 catalyst exhibited a visible-light-driven photocatalytic $H_2$ evolution rate of 8.71 $mmolg^{-1}h^{-1}$, which was significantly higher than that of pure $Zn_3In_2S_6$. Additionally, 2.4% M/ZIS6 demonstrated exceptional photocatalytic degradation capabilities toward acid orange 7 (AO7) during HER. The improved photocatalytic efficiency of 2.4 % M/ZIS6 was mainly due to themetallic phase of co-catalyst 1T-$MoS_2$, and plentiful S imperfections. A close and large contact interface between 1T-$MoS_2$ and $Zn_3In_2S_6$ also led to improved photocatalytic efficiency. Yang et al.[143] synthesized nanocomposites ($MoS_2$/FeS) with FeS as a catalyst and $MoS_2$ as a cocatalyst to achieve effective heterogeneous Fenton reactions. The inclusion of FeSNPs on $MoS_2$ nanosheet surfaces establishes a robust interface between the two components, resulting in a prominent increase in the number of exposed $Mo^{6+}$ sites and S vacancies. These structural features play a crucial role in enhancing the degradation rate of pollutants such as insecticides and organic dyes by facilitating the rapid cycling of $Fe^{3+}/Fe^{2+}$ and promoting swift electron transfer (Figure 21). Furthermore, the $MoS_2$/FeS nanocomposite catalysts exhibit remarkable efficiency when operating in near-neutral conditions at pH = 6.5 (Figure 22). This effectively tackles the obstacles faced in conventional Fenton reactions, such as metal ion leaching, iron slurry formation, and the necessity to adjust solution pH.[144] The process of creating the nanocomposite involves the development of $MoS_2$ nanosheets, the absorption of $Fe^{3+}$ on the surface, and the subsequent generation of FeSNPs through high-temperature treatment in an $N_2$ atmosphere. In neutral conditions, the RhB degradation rate of the $MoS_2$/FeS nanocomposite reaches a maximum rate of $3.3 \times 10^{-2} min^{-1}$. This degradation rate was 12,000 and 9.4 times greater than that of FeS and $MoS_2$, respectively, and 4.9 times greater than the combined mixture of $MoS_2$ and FeS. Furthermore, the $MoS_2$/FeS composite also exhibited remarkable reusability, maintaining 85% of its actual efficiency after undergoing ten cycles of advanced oxidation process (AOP) reactions as shown in Figure 22.

Yan et al.[145] developed a $CoFe_2O_4$/$MoS_2$ co-catalytic heterogeneous Fenton system that exhibited the remarkable ability to degrade organic pollutants, such as phenol, in a solution with a neutral pH. Chemical bonding with $MoS_2$ creates an acidic microenvironment within the $CoFe_2O_4$ slipping plane. This microenvironment remains stable regardless of ambient pH levels, ensuring consistent circulation of $Fe^{3+}/Fe^{2+}$ on this composite surface, even under alkaline or neutral conditions. Moreover, the $CoFe_2O_4$/$MoS_2$ system consistently provides new active sites for $H_2O_2$ decomposition and $O_2$ generation to prevent iron sludge formation and the enhanced treatment of organic pollutants. It has been shown to benefit the overall efficiency of wastewater treatment, especially when applied in practical wastewater settings. The hydrothermal treatment of $CoFe_2O_4$ enhances its phenol degradation activity within a 25-min timeframe. A physical mixture of $CoFe_2O_4$ and $MoS_2$ only slightly improves the degradation rate of phenol to 16%. In contrast, the $CoFe_2O_4$/$MoS_2$ composite exhibits a remarkable 100% degradation rate highlighting the significant role of the chemical interaction between $CoFe_2O_4$ and $MoS_2$ in the Fenton reaction.

Zhao et al.[24] discovered an enhanced Fenton-like technique utilizing calcium peroxide ($CaO_2$) in place of $H_2O_2$ and used $MoS_2$ as a cocatalyst. The resultant obtained was a Rhb removal efficiency of 98.60 % with the specified conditions of 1.0 mmol $L^{-1}$ $CaO_2$, 0.50 mmol $L^{-1}$ $Fe^{2+}$, 0.30 g $L^{-1}$ $MoS_2$, and a starting pH of 5.85 within a 5-min time frame. Here, $MoS_2$ acted as a high-efficiency cocatalyst and $CaO_2$ as an oxidant. This combination led to the successful reduction of $Fe^{3+}$ through the reactive position of $Mo^{4+}$ on $MoS_2$. Additionally, $MoS_2$ promotes the generation of ROS, enhancing the overall performance of the system. In the initial stage, the rapid dissolution of $CaO_2$ leads to the production of $H_2O_2$ within a short period. This, in turn, results in the formation of HO• radicals through the catalytic decomposition of $H_2O_2$. In the subsequent phase, the generation of $H_2O_2$ becomes slower and more stable. Also, the efficiency of its removal


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

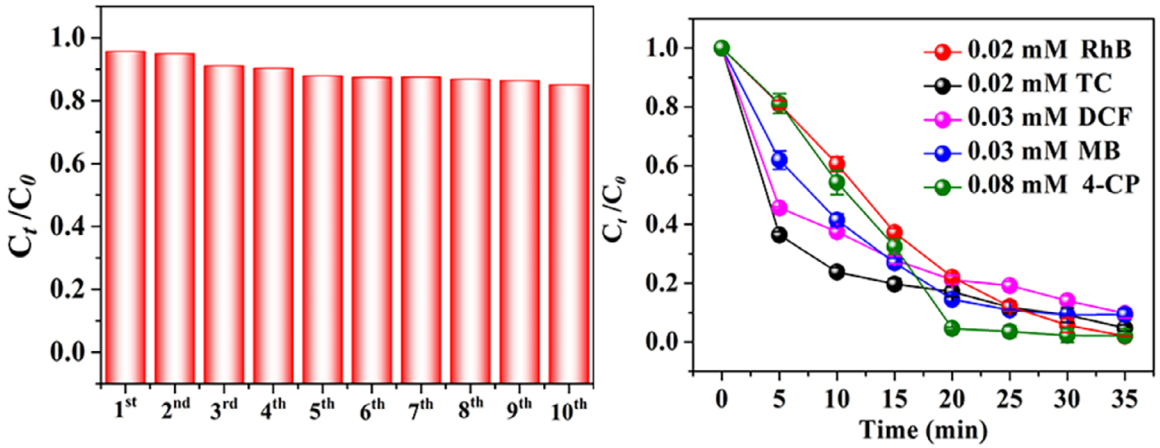



**Figure 22.** Fenton reaction catalyzed by $MoS_2$/FeS. (a) Recycling experiments of rhodamine B(RhB) degradation recorded at 30 min. (b) Various pollutants exhibit distinct degradation profiles under specific reaction conditions at pH of 6.5. Reprinted with permission from ref. [143]. Copyright, 2021 American Chemical Society.

presented in this study offers proof of the interdependent catalytic behavior of $MoS_2$ within the $CaO_2/Fe^{2+}/MoS_2$ system. The conversion between $Fe^{3+}$ and $Fe^{2+}$ would not be adequate in the absence of $MoS_2$, thereby restricting the oxidation potential of the synergistic process. However, the addition of $MoS_2$ facilitates the acceleration of electron transfer, thereby enhancing the $Fe^{3+}/Fe^{2+}$ conversion cycle. This significantly improves the Fenton-like reaction and enhances the removal of dyes. This process has also explained the markable oxidation efficiency of different common organic pollutants, dyes like MO, metronidazole (MTZ), MB, and so on, and phenol. In real wastewater treatment, the cocatalytic Fenton-like process involving $MoS_2$ has achieved a chemical oxygen demand (COD) removal rate of 64%, indicating its promising practical application prospects.

Jiang et al.[146] used a simple carbonation-hydrothermal method to modify the Co-based catalysts to overcome issues related to their leaching and agglomeration during the reactions. Therefore they prepared a composite named $Co/PCN/MoS_2$ using g-$C_3N_4$, $MoS_2$, and Co metal-organic framework (MOF). The active units were MOFs, the co-catalyst was $MoS_2$, and the carrier was g-$C_3N_4$.The unique flower-shaped nanostructure of $MoS_2$ allowed for the exposure of less active sites, leading to the effective dispersion of Co-MOF throughout the structure. Moreover, the presence of both S species and $Mo^{4+}/Mo^{6+}$ in $MoS_2$ facilitated the electron transfer process and valence style, thereby enhancing the performance of the catalyst as a composite. Analysis using electrochemical methods demonstrated a higher redox capacity of$Co/PCN/MoS_2$ in comparison to $MoS_2$. Further, the composite $Co/PCN/MoS_2$ was used to activate the oxidant peroxymonosulfate (PMS), which is used to degrade organic pollutants. A 97.19 % RhB degradation rate was obtained by the $Co/PCN/MoS_2$/PMS system within 30 minutes, outperforming other systems such as Co-MOF/PMS, g-$C_3N_4$/PMSCo-MOF/$MoS_2$/PMS, $MoS_2$/PMS, and Co-MOF/g-$C_3N_4$/PMS. The catalyst showed outstanding resistance to interference from coexisting substances, displaying efficient performance of degradation over a broad pH range of 2–10. After 5 cycles, it maintained impressive stability and reusability, resulting in a RhB degradation rate of 91.10%. Quenching tests confirmed that the primary ROS accountable for RhB degradation is singlet oxygen ($^1O_2$).

Salarizadeh et al.[147] reported a nanocatalyst $MoS_2/Ni_3S_2$/rGO for methanol and ethanol oxidation reactions (MOR/EOR) in alkaline media. The catalyst was prepared via a one-step hydrothermal method, forming a porous starfish-like $MoS_2/Ni_3S_2$ structure uniformly dispersed over rGO nanosheets. Electrochemical studies demonstrated excellent catalytic efficiency, with stability retention of 106% for MOR and 104% for EOR after 200 CV cycles. The enhanced performance is attributed to the synergistic effect of $MoS_2$ and $Ni_3S_2$ which provides abundant active sites, and rGO, which increases surface area, electrical conductivity, and dispersion. The catalyst facilitates the adsorption of hydroxyl ions and alcohols, improving reaction kinetics. Fuel cell tests using $MoS_2/Ni_3S_2$/rGO as the anode catalyst achieved open circuit voltages (OCVs) of 0.59 V and 0.48 V for methanol and ethanol respectively. They also exhibited a high value of peak power densities of 4.42 mW $cm^{-2}$ for methanol and 3.5 mW $cm^{-2}$ for ethanol, offering promising potential for next-generation direct alcohol fuel cells.

Various $MoS_2$-based composites and their applications as catalysts are summarised in Table 2 and the diagrammatic representation of some $MoS_2$ composites is shown in Figure 23.

## 4. Conclusion and Future Prospects

$MoS_2$ has established itself as a highly efficient and versatile catalyst, demonstrating remarkable performance in electrocatalysis, photocatalysis, and cocatalysis. Its unique layered structure, abundant active sites, and tunable electronic properties make it an ideal candidate for various catalytic applications, including hydrogen evolution, hydrodesulfurization, pollutant degradation, and organic transformations. Additionally, $MoS_2$-based catalysts are stable, cost-effective, and environmentally friendly, reinforcing their potential in sustainable chemical processes. Looking ahead, the future of $MoS_2$-based catalysts lies in optimizing their structural and electronic properties to enhance


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**Table 2.** The preparation of $MoS_2$-based composites and their application as a catalyst.

| S. No. | Composites | Synthesis Method | Morphologies | Type of Catalyst | Catalytic Action | Results | References |
|---|---|---|---|---|---|---|---|
| 1 | $MoS_2$/Au Heterojunctions | Hydrothermal | Multilayered | Heterogeneous catalyst | Large surface area and more sulphur vacancies | Degradation of methylene blue | Wei et al.[94] |
| 2 | $MoS_2$-rGOA | Hydrothermal | Nanosheets | Heterogeneous catalyst | The volume of adsorption increased with the introduction of rGO | Reduction of various organic moieties | Kaushik et al.[95] |
| 3 | $MoS_2/Fe_3O_4$ MNPs | Biomimetic | Nanosheets | Heterogeneous catalyst | Aggregation of composite sustained and increased the catalytic activities | Degradation of organic dyes | Shen et al.[96] |
| 4 | $MoS_2$-MoC, $MoS_2$-$Mo_2N$ | Electrochemical | Nanocomposite | Heterogeneous catalyst | Increased presence of sulphur vacancies | Degradation of pollutants | Li et al.[97] |
| 5 | ZnS/$MoS_2$ | Hydrothermal | Nanocomposite | Heterogeneous catalyst | Band gap optimization and reduced rate of electron hole recombination | HER, | Jayan et al.[98] |
| 6 | $Ag_3O_4/MoS_2$ | Ultrasonic exfoliation and hydrotherml | Nanosheets | Heterogeneous catalyst | High electrochemical surface area | HER | Satheesh et al.[99] |
| 7 | $CoFe_2O_4/MoS_2$ | Hydrothermal | Nanosheets | Heterogeneous catalyst | Increased surface area and high recyclability leads to more economic production | esterification of oleic acid | Ibrahim et al.[100] |
| 8 | $MoS_2$ Flower | Hydrothermal | Nanosphere | Heterogeneous catalyst | Active Bronsted and Lewis acid sites | Acetalization and etherification of 5-HMF | Zheng et al.[101] |
| 9 | $MoS_2$ | Hydrothermal or chemical vapour deposition or solvothermal | layered | Heterogeneous catalyst | High selectivity and more tolerance for functional groups like $NH_2$, OH during reduction | Nitrobenzene reduction | Huang et al.[102] |
| 10 | $MoS_2$/ $\gamma$-$Al_2O_3$ catalyst. | Impregnation | Layered | Heterogeneous catalyst | Large surface area and more selectivity | Desulphurization and hydrogenation reaction | Salnikov et al.[103] |
| 11 | $MoS_2$-RGO | Hydrothermal | Nanocomposite | Heterogeneous catalyst | Large surface area and more active sites | Indole alkaloids synthesis in water | Bahuguna et al.[104] |
| 12 | Ni/$MoS_2$/SM | Hydrothermal | Nanosheets | Electrocatalyst | Weak Adsorption of $S_2$* on $MoS_2$ and strong adsorption S* and $S_2$* on Ni | Sulphides oxidation and $H_2$ production | Liu et al.[116] |
| 13 | $MoS_2$ nanoislands | Hydrothermal | Nanoislands | Electrocatalyst | Increase exposure of edges of $MoS_2$ for various catalytic properties | HER and OER | Chen et al.[81] |
| 14 | Co/Se-$MoS_2$-NF | Solvothermal | Nanofoam | Electrocatalyst | Selenium and cobalt stabilize the catalyst structure by the generation of both in-plane and out-plane active edge sites | HER | Zheng et al.[99] |


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**Table 2.** (Continued)

| S. No. | Composites | Synthesis Method | Morphologies | Type of Catalyst | Catalytic Action | Results | References |
|---|---|---|---|---|---|---|---|
| 15 | $MoS_2/CoS_2$ | Hydrothermal | Nanosheets | Electrocatalyst | Synergetic effect of $MoS_2$ and $CoS_2$ increase the surface are for fast electron transfer | HER | Gu et al.[118] |
| 16 | $MoS_2$- $Fe_2Mo_3O_8/MoO_2$ | Pyrolysis and Sulfidation | Nanocomposite | Electrocatalyst | Increase number of active sites promoting ion diffusion rate to increase current density | OER/HER | Chen et al.[119] |
| 17 | $Fe_3O_4$@$MoS_2$/RGO | Hydrothermal | Nanocomposite | Electrocatalyst | By increasing the active surface area of exposure by the collaborative impact of iron oxide and conductive RGO in both acidic and basic medium | Oxidation of Methanol | Askari et al.[120] |
| 18 | $MoS_2$/g-$C_3N_4$ | Hydrothermal | Nanosheets | Photocatalyst | Edge-enriched 1T phase with more active sites and heterojunction with g-$C_3N_4$ increase the charge transfer of and separation | Reduction of $H_2O$ | Hu et al.[131] |
| 19 | Au-$MoS_2$ | Hydrothermal | Nanosheets | Photocatalyst | Low excitation energy and gold nanoparticles enhance catalytic activity by enhancing electron-hole pair and ROS generation | Water purification | Chou et al.[132] |
| 20 | $MoS_2$ layered structure | Hydrothermal | Nanosheets | Photocatalyst | Efficient transfer of electrons between the large surface area of catalyst and substrate and no unwanted product formed | Degradation of Bromophenol blue | Kisala et al.[133] |
| 21 | $MoS_2$ graphene hybrid (MSG) | Hydrothermal | Nanosheets | Photocatalyst | Charge transfer and photocurrent adjusted, and defects increase on the introduction of graphene | Production of $H_2$ gas | Lee et al.[134] |
| 22 | $MoO_3/MoS_2$ | Hydrothermal | Nanorods | Photocatalyst | Increase surface area and more no of electron hole pair increase the photocatalytic activity | Antimicrobial and antibacterial, Formaldehyde removal | Zhou et al.[135] |


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**Table 2.** (Continued)

| S. No. | Composites | Synthesis Method | Morphologies | Type of Catalyst | Catalytic Action | Results | References |
|---|---|---|---|---|---|---|---|
| 23 | $MoS_2$-CdS | Solvothermal | Nanocomposite | Photocatalyst | photogenerated electrons from the CdS to $MoS_2$ decreased the recombination probabilities of photogenerated charge carriers, enhancing the photo degradation efficiency of the composite | Photocatalytic reduction of Cr(VI) | Kar et al.[136] |
| 24 | $NiMoO_4/MoS_2$ | Hydrothermal | Nanosheet | Co-catalyst | Large surface-active sites and good contact with electrolytes. Combined effects increase the electrochemical activities | OER | Solomon et al.[141] |
| 25 | $CoFe_2O_4/MoS_2$ | Hydrothermal | Nanosheets | Co-catalyst | Composite provides new active sites for the decomposition of $H_2O_2$and phenol | Degradation of organic pollutants | Yan et al.[145] |
| 26 | $MoS_2$/FeS | Hydrothermal | Nanocomposite | Co- catalyst | $MoS_2$ surface with an increase in active sites and sulfur vacancies due to the presence of FeS. | Fenton reaction (degradation of heavy metal ions | Yang et al.[143] |
| 27 | $MoS_2$/ $Zn_3In_2S_6$ | Hydrothermal | Nanocomposite | Co-catalyst | Increase the number of active sites with abundant sulfur vacancies | HER | Yan et al.[142] |
| 28 | $CaO_2/MoS_2$ | Hydrothermal | Nanosheets | Co-catalyst | Composite increased the decomposition of dyes and organic pollutants by effectively reducing $Fe^{3+}$ through the reactive position of $Mo^{4+}$ on $MoS_2$ | Degradation of various dyes like MO, MB, MTZ etc | Zhao et al.[24] |
| 29 | Co/PCN/$MoS_2$ | Hydrothermal | Nanocomposite | Co-catalyst | Increased exposure of mariginal active sites, increased electron hole transfer process | RhB Degradation | Jiang et al.[146] |
| 30 | $MoS_2/Ni_3S_2$/rGO | Solvothermal | Nanocomposite | Co-catalyst | Abundant active sites, increased surface area, electrical conductivity and dispersion | Alcohol fuel generation | Salarizadeh et al.[147] |


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

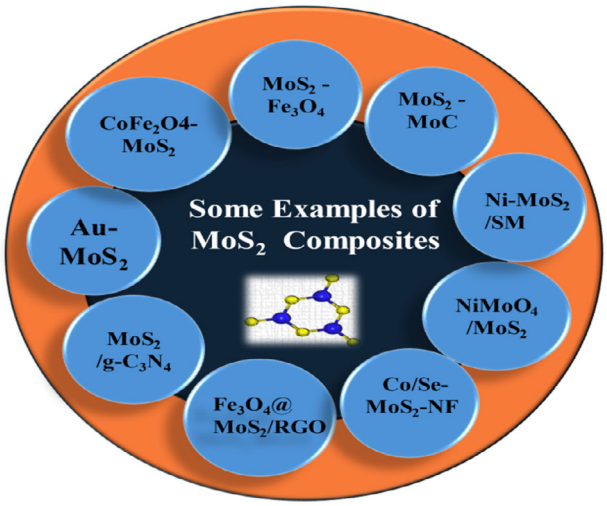



**Figure 23.** Diagrammatic representation of some composites of $MoS_2$.

catalytic activity and selectivity. Research efforts on advanced synthesis techniques, surface engineering, and heterostructure design to improve stability, recyclability, and efficiency shall continue in the future too. The integration of $MoS_2$ with other nanomaterials and hybrid systems will further expand its applicability in emerging fields such as $CO_2$ reduction, energy storage, and green hydrogen production. As materials science and catalysis continue to evolve, $MoS_2$ is expected to play a transformative role in next-generation catalytic systems. With ongoing innovations, $MoS_2$-based catalysts will not only drive advancements in clean energy and environmental remediation but also contribute significantly to the development of more efficient and sustainable chemical processes.

## Acknowledgements

Authors are highly obliged to K. R. Mangalam University for promoting research and providing the necessary infrastructure and facilities. Authors Ritika Khatri and Suman Srivastava would like to thank the seed research grant [KRMU/ADMIN/SEED/2024-25/4331(A)] and Ritika Khatri would also like to thank the seed research grant [KRMU/ADMIN/SEED/2023-24/3910(4)] received from K. R. Mangalam University.

## Conflict of Interests

The authors declare no conflict of interest.

## Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

**Keywords:** Cocatalyst · Electrocatalyst · Heterogeneous catalyst · $MoS_2$ · Photocatalyst

[1] R. Lv, J. Robinson, R. E. Schaak, D. Sun, Y. Sun, T. E. Mallouk, M. Terrones, *Acc Chem. Res.* **2015**, *48*, 56–64.
[2] H. Qi, L. Wang, J. Sun, Y. Long, P. Hu, F. Liu, X. He, *Crystals* **2018**, *8*, 35.
[3] T. Van Nguyen, M. Tekalgne, T. P. Nguyen, Q. Van Le, S. H. Ahn, S. Y. Kim, *Battery Energy* **2023**, *2*, 20220057.
[4] M. Velický, P. S. Toth, *Appl. Mater. Today* **2017**, *8*, 68–103.
[5] S. Joseph, J. Mohan, S. Lakshmy, S. Thomas, B. Chakraborty, S. Thomas, N. Kalarikkal, *Mater. Chem. Phys.* **2023**, *297*, 127332.
[6] X. Duan, H. Zhang, *Chem. Rev.* **2024**, *124*, 10619–10622.
[7] A. Hasani, M. Tekalgne, Q. V. Le, H. W. Jang, S. Y. Kim, *J. Mater. Chem. A* **2019**, *7*, 430–454.
[8] J. Xu, G. Shao, X. Tang, F. Lv, H. Xiang, C. Jing, S. Liu, S. Dai, Y. Li, J. Luo, Z. Zhou, *Nat. Commun.* **2022**, *13*, 2193.
[9] R. Khatri, N. K. Puri, *New J. Chem.* **2022**, *46*, 7477–7489.
[10] M. Acerce, D. Voiry, M. Chhowalla, *Nat. Nanotechnol.* **2015**, *10*, 313–318.
[11] X. Tian, S. Wang, H. Li, M. Li, T. Chen, X. Xiao, Y. Wang, *Sensors & Diagnostics* **2023**, *2*, 361–381.
[12] Y. Cao, *ACS Nano* **2021**, *15*, 11014–11039.
[13] A. Di Bartolomeo, A. Kumar, O. Durante, A. Sessa, E. Faella, L. Viscardi, K. Intonti, F. Giubileo, N. Martucciello, P. Romano, S. Sleziona, M. Schleberger, *Mater. Today Nano* **2023**, *24*, 100382.
[14] J. Ding, X. Hou, Y. Qiu, S. Zhang, Q. Liu, J. Luo, X. Liu, *Inorg. Chem. Commun.* **2023**, *151*, 110621.
[15] a) N. Thomas, S. Mathew, K. M. Nair, K. O'Dowd, P. Forouzandeh, A. Goswami, G. McGranaghan, S. C. Pillai, *Mater. Today Sustain.* **2021**, *13*, 100073; b) U. Sharma, S. Karazhanov, N. Alonso-Vante, S. Das, *Curr. Opin. Electrochem.* **2022**, *35*, 101067; c) M.-H. Fan, C.-H. Wang, X. Yu, J. Ding, A.-Q. Xiao, Y. Li, W.-Y. Huang, *Tungsten* **2024**, *6*, 473–487.
[16] K. B. M. Ismail, M. Arun Kumar, S. Mahalingam, J. Kim, R. Atchudan, *Materials* **2023**, *16*, 4471.
[17] S. Ogawa, S. Fukushima, M. Shimatani, *Materials* **2023**, *16*, 2005.
[18] L. Yadav, S. Pakhira, *J. Mater. Chem.* **2023**, *11*, 15215–15232.
[19] R. Nandee, M. A. Chowdhury, A. Shahid, N. Hossain, M. Rana, *Results Eng.* **2022**, *15*, 100474.
[20] J. Jampilek, K. Kralova, *Int. J. Mol Sci.* **2022**, *23*, 6253.
[21] A. N. Ghulam, O. A. L. dos Santos, L. Hazeem, B. Pizzorno Backx, M. Bououdina S. Bellucci, *J. Funct. Biomater.* **2022**, *13*, 77.
[22] S. Kumar, V. Maivizhikannan, J. Drews, V. Krishnan, *Vacuum* **2019**, *163*, 88–98.
[23] H. M. El Sharkawy, A. M. Shawky, R. Elshypany, H. Selim, *Sci. Rep.* **2023**, *13*, 8845.
[24] M. Zhao, Y. Xiang, X. Jiao, B. Cao, S. Tang, Z. Zheng, X. Zhang, T. Jiao, D. Yuan, *Sep. Purif. Technol.* **2021**, *276*, 119289.
[25] E. S. Kadantsev, P. Hawrylak, *Solid State Commun.* **2012**, *152*, 909–913.
[26] N. Thomas, S. Mathew, K. M. Nair, K. O'Dowd, P. Forouzandeh, A. Goswami, S. C Pillai, *Mater. Today Sustain.* **2021**, *13*, 100073.
[27] Q. Ding, B. Song, P. Xu, S. Jin, *Chem.* **2016**, *1*, 699–726.
[28] S. Ippolito, P. Samorì, *Small Sci.* **2022**, *2*, 2100122.
[29] Y. Guo, J. Li, *Mater. Sci. Eng. C* **2020**, *109*, 110511.
[30] A. V. Bandura, S. I. Lukyanov, D. D. Kuruch, A. R. Evarestov, *Physica. E Low Dimens. Syst. Nanostruct.* **2020**, *124*, 114183.
[31] D. Li, L. Zhao, Q. Xia, J. Wang, X. Liu, H. Xu, S. Chou, *Adv. Funct. Mater.* **2022**, *32*, 2108153.
[32] M. Ahmaruzzaman, V. Gadore, *J. Environ. Chem. Eng.* **2021**, *9*, 105836.
[33] J. Sun, X. Li, W. Guo, M. Zhao, X. Fan, Y. Dong, C. Xu, J. Deng, Y. Fu, *Crystals* **2017**, *7*, 198.
[34] R. Khatri, N. K. Puri, *Vacuum* **2020**, *175*, 109250.
[35] D. Pariari, D. D. Sarma, *APL Mater.* **2020**, *8*, 040909.
[36] X. Fan, P. Xu, D. Zhou, Y. Sun, Y. C. Li, M. Nguyen, M. Terrones, T. E. Mallouk, *Nano Lett.* **2015**, *15*, 5956–5960.
[37] W. Qiao, S. Yan, X. Song, X. Zhang, X. He, W. Zhong, Y. Du, *Appl. Surf. Sci.* **2015**, *359*, 130–136.
[38] M. Telkhozhayeva, E. Teblum, R. Konar, O. Girshevitz, I. Perelshtein, H. Aviv, Y. S. Tischler, G. D. Nessim, *Langmuir* **2021**, 374504–4514.
[39] X. Yang, Z. Chen, J. Fang, Q. Yang, W. Zhao, X. Qian, C. Liu, D. Zhou, S. Tao, X. Liu, *Mater. Lett.* **2019**, *255*, 126596.
[40] J. Zhang, A. Yang, X. Wu, J. Van De Groep, P. Tang, S. Li, B. Liu, F. Shi, J. Wan, Q. Li, Y. Sun, Z. Lu, X. Zheng, G. Zhou, C.-L. Wu, S.-C. Zhang, M. L. Brongersma, J. Li, Y. Cui, *Nat. Commun.* **2018**, *9*, 5289.
[41] S. Parida, A. Mishra, J. Chen, J. Wang, A. Dobley, C. B. Carter, A. M. Dongare, *J. Am. Ceram. Soc.* **2020**, *103*, 6603–6614.


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

[42] N. Liu, P. Kim, J. H. Kim, J. H. Ye, S. Kim, C. J. Lee, *ACS Nano* **2014**, *8*, 6902–6910.
[43] V. Forsberg, R. Zhang, J. Bäckström, C. Dahlström, B. Andres, M. Norgren, M. Andersson, M. Hummelgård, H. Olin, *PLoS One* **2016**, *11*, e0154522.
[44] H. Yu, H. Zhu, M. Dargusch, Y. Huang, *J. Colloid Interface Sci.* **2018**, *514*, 642–647.
[45] A. Gupta, V. Arunachalam, S. Vasudevan, *J. Phys. Chem. Lett.* **2016**, *7*, 4884–4890.
[46] H. Ma, S. Ben, Z. Shen, X. Zhang, C. Wu, S. Liao, S. F. An, *Appl. Surf. Sci.* **2020**, *512*, 145588.
[47] C. Wu, M. Xu, C. Zhang, L. Ye, K. Zhang, H. Cong, L. Zhuang, X. Ai, H. Yang, J. Qian, *Energy Stor. Mater.* **2023**, *55*, 154–165.
[48] K. Manna, H. N. Huang, W. T. Li, Y. H. Ho, Y. W. H. Chiang, *Chem. Mater.* **2016**, *28*, 7586–7593.
[49] D. Sahoo, B. Kumar, J. Sinha, S. Ghosh, S. S. Roy, B. Kaviraj, *Sci. Rep.* **2020**, *10*, 10759.
[50] X. Zheng, Y. Zhu, Y. Sun, Q. Jiao, *J. Power Sources* **2018**, *395*, 318–327.
[51] M. Mulu, D. Ramadevi, N. Belachew, K. Basavaiah, *RSC Adv.* **2021**, *11*, 24536–24542.
[52] S. A. Getaneh, A. G. Temam, A. C. Nwanya, P. M. Ejikeme, F. I. Ezema, *Technol.* **2024**, *40*, 185–212.
[53] A. Arora, K. Sharma, S. K. Tripathi, *J. Nanomater. Biostruct.* **2022**, *17*, 871–880.
[54] X. Zhang, G. Ma, J. Wang, *Tungsten* **2019**, *1*, 59–79.
[55] H. Wang, C. Li, P. Fang, Z. Zhang, J. Z. Zhang, *Chem. Soc. Rev.* **2018**, *47*, 6101–6127.
[56] J. Jeon, S. K. Jang, S. M. Jeon, G. Yoo, Y. H. Jang, J. H. Park, S. Lee, *Nanoscale* **2015**, *7*, 1688–1695.
[57] Y. H. Lee, X. Q. Zhang, X. W. Zhang, M. T. Chang, C. T. Lin, K. D. Chang, Y. C. Yu, J. T. W. Wang, C. S. Chang, L. J. Li, T. W. Lin, *arXiv preprint arXiv* **2012**, *1202*, 5458.
[58] X. Wang, H. Feng, Y. Wu, L. Jiao, *J. Am. Chem. Soc.* **2013**, *135*, 5304–5307.
[59] S. Balendhran, J. Z. Ou, M. Bhaskaran, S. Sriram, S. Ippolito, Z. Vasic, E. Kats, S. Bhargava, S. Zhuiykov, K. Kalantar-Zadeh, *Nanoscale* **2012**, *4*, 461–466.
[60] A. H. Motagamwala, J. A. Dumesic, *Chem. Rev.* **2020**, *121*, 1049–1076.
[61] S. Dery, B. Friedman, H. Shema, E. Gross, *Chem. Rev.* **2023**, *123*, 6003–6038.
[62] S. H. Albayati, N. G. Nezhad, A. G. Taki, R. N. Z. R. A. Rahman, *Int. J. Biol. Macromol.* **2024**, *276*, 133978.
[63] N. Parthiban, G. Dhanasekaran, G. Vijayakumar, R. Gopal, E. Thangavel, P. Sakthivel, S. Sangaraju, *Inorg. Chem. Commun.* **2024**, *167*, 112730.
[64] L. Lei, D. Huang, G. Zeng, M. Cheng, D. Jiang, C. Zhou, S. Chen, W. Wang, *Coord. Chem. Rev.* **2019**, *399*, 213020.
[65] A. Kumar, V. Navakoteswara Rao, A. Kumar, M. Venkatakrishnan Shankar, V. Krishnan, *ChemPhotoChem* **2020**, *4*, 427–444.
[66] Y. Shi, Y. Zhou, D.-R. Yang, W.-X. Xu, C. Wang, F.-B. Wang, J.-J. Xu, X.-H. Xia, H.-Y. Chen, *J. Am, Chem. Soc.* **2017**, *139*, 15479–15485.
[67] M. N. Lakhan, A. Hanan, Y. Wang, H. K. Lee, H. Arandiyan, *Chem. Sci.* **2024**, *15*, 15540–15564.
[68] M. Asadi, B. Kumar, A. Behranginia, B. A. Rosen, A. Baskin, N. Repnin, D. Pisasale, P. Phillips, W. Zhu, R. Haasch, R. F. Klie, P. Král, J. Abiade, A. Salehi-Khojin, *Nat. Commun.* **2014**, *5*, 4470.
[69] U. Aftab, M. Y. Solangi, A. Tahira, A. Hanan, M. I. Abro, A. Karsy, E. Dawi, M. A. Bhatti, R. H. Alshammari, A. Nafady, A. Gradone, R. Mazzaro, V. Morandi, A. Infantes-Molina, Z. H. Ibupoto, *RSC Adv.* **2023**, *13*, 32413–32423.
[70] D. Li, B. Ren, Q. Jin, H. Cui, C. Wang, *J. Mater. Chem. A* **2018**, *6*, 2176–2183.
[71] H. Wang, Y. Chen, X. Hou, C. Ma, T. Tan, *Green Chem.* **2016**, *18*, 3250–3256.
[72] Y. Zheng, Y. Jiao, Y. Zhu, Q. Cai, A. Vasileff, L. H. Li, Y. Han, Y. Chen, S. Z. Qiao, *J. Am. Chem. Soc.* **2017**, *139*, 3336–3339.
[73] Y. Zeng, X. Liu, C. Liu, L. Wang, Y. Xia, S. Zhang, S. Luo, Y. Pei, *Appl. Catal. B Environ.* **2018**, *224*, 1–9.
[74] W. Li, X. Wang, Q. Ma, F. Wang, X.-S. Chu, X.-C. Wang, C.-Y. Wang, *Appl. Catal. B Environ.* **2021**, *284*, 119688.
[75] Q. Yuan, M. Fan, Y. Zhao, J. Wu, J. Raj, Z. Wang, A. Wang, H. Sun, X. Xu, Y. Wu, K. Sun, J. Jiang, *Appl. Catal. B Environ.* **2023**, *324*, 122195.
[76] Z. W. Seh, K. D. Fredrickson, B. Anasori, J. Kibsgaard, A. L. Strickler, M. R. Lukatskaya, Y. Gogotsi, T. F. Jaramillo, A. Vojvodic, *ACS Energy Lett.* **2016**, *1*, 589–594.
[77] R. Wang, Z. Chang, Z. Fang, T. Xiao, Z. Zhu, B. Ye, C. Xu, J. Cheng, *Int. J. Hydrogen Energy* **2020**, *45*, 28190–28195.
[78] X. Yang, Q. Jia, F. Duan, B. Hu, M. Wang, L. He, Y. Song, Z. Zhang, *Appl. Surf. Sci.* **2019**, *464*, 78–87.
[79] X. Ren, J. Zhou, X. Qi, Y. Liu, Z. Huang, Z. Li, Y. Ge, S. C. Dhanabalan, J. S. Ponraj, S. Wang, J. Zhong, H. Zhang, *Adv. Energy Mater.* **2017**, *7*, 1700396.
[80] Z. Zhang, M. Khurram, Z. Sun, Q. Yan, *Inorg. Chem.* **2018**, *57*, 4098–4103.
[81] B. Chen, P. Hu, F. Yang, X. Hua, F. F. Yang, F. Zhu, R. Sun, K. Hao, K. Wang, Z. Yin, *Small* **2023**, *19*, 2207177.
[82] M. Yang, W. Bao, J. Zhang, T. Ai, J. Han, Y. Li, J. Liu, P. Zhang, L. Feng, *J. Colloid Interface Sci.* **2024**, *658*, 32–42.
[83] S. Chen, Y. Pan, *Int. J. Hydrog. Energy* **2021**, *46*, 21040–21049.
[84] R. Li, L. Yang, T. Xiong, Y. Wu, L. Cao, D. Yuan, W. Zhou, *J. Power Sources* **2017**, *356*, 133–139.
[85] Y. Jiang, S. Li, F. Zhang, W. Zheng, L. Zhao, Q. Feng, *Catal. Commun.* **2021**, *156*, 106325.
[86] C. Vidya, C. Manjunatha, A. Pranjal, I. Faraaz, K. Prashantha, *Mater. Res. Innov.* **2023**, *27*, 177–193.
[87] M. Mavvaji, S. Akkoc, *Coord. Chem. Rev.* **2024**, *505*, 215714.
[88] A. Bahuguna, A. Kumar, S. Kumar, T. Chhabra, V. Krishnan, *ChemCatChem* **2018**, *10*, 3121–3132.
[89] R. Gao, L. Pan, H. Wang, X. Zhang, L. Wang, J. J. Zou, *ACS Catal.* **2018**, *8*, 8420–8429.
[90] M. El Kadiri, T. El Assimi, P. Thébault, A. El Meziane, S. Royer, A. El Kadib, G. Gouhier, M. Lahcini, *ACS Appl. Nano Mater.* **2023**, *6*, 4017–4027.
[91] B. Nanda, M. Sailaja, P. Mohapatra, R. K. Pradhan, B. B. Nanda, *Mater. Today* **2021**, *47*, 1234–1240.
[92] X. Chen, X. Q. Shen, Z. Li, W. Wan, J. Chen, J. Zhang, *ACS Appl. Mater. Interfaces* **2019**, *12*, 654–666.
[93] K. E. Dungey, M. D. Curtis, *J. Am. Chem. Soc.* **1997**, *119*, 842–843.
[94] Q. Wei, B. Lu, Q. Yang, C. Shi, Y. Wei, M. Xu, C. Zhang, Y. Yuan, *Materials* **2023**, *16*, 1169.
[95] J. Kaushik, C. Sharma, N. K. Lamba, P. Sharma, G. S. Das, K. M. Tripathi, R. K. Joshi, S. K. Sonkar, *Langmuir* **2023**, *39*, 12865–12877.
[96] K. Shen, Y. Cui, D. Zhang, M. Liu, H. Huang, X. Sha, F. Deng, N. Zhou, X. Zhang, Y. Wei, *J. Environ. Chem. Eng.* **2020**, *8*, 104125.
[97] S. Li, W. Yu, X. Zhang, L. Liu, H. Wang, Y. Peng, Z. Bian, *ACS Appl. Mater. Interfaces* **2022**, *15*, 1326–1338.
[98] G. Jayan, L. Elias, A. Anil, T. C. Bhagya, S. M. A. Shibli, *Int. J. Hydrog. Energy* **2024**, *51*, 1375–1386.
[99] D. Satheesh, L. Baskar, Y. Jayavelu, A. Dekshinamoorthy, V. R. Sakthinathan, P. J. Daniel, S. Vijayaraghavan, K. Krishnan, R. Rajendran, R. Pachaiappan, K. Manavalan, *Chemosphere* **2024**, *351*, 141220.
[100] A. S. I. Ibrahim, B. Gözmen, Ö. Sönmez, *Fuel* **2024**, *371*, 131988.
[101] J. Zheng, M. Wang, X. Tong, Y. Yuan, *Catal. Today* **2024**, *443*, 114984.
[102] L. Huang, P. Luo, M. Xiong, R. Chen, Y. Wang, W. Xing, J. Huang, *Chinese J. Chem.* **2013**, *31*, 987–991.
[103] O. G. Salnikov, D. B. Burueva, D. A. Barskiy, G. A. Bukhtiyarova, K. V. Kovtunov, I. V. Koptyug, *ChemCatChem* **2015**, *7*, 3508–3512.
[104] A. Bahuguna, S. Kumar, V. Sharma, K. L. Reddy, K. Bhattacharyya, P. C. Ravikumar, V. Krishnan, *ACS Sustainable Chem. Eng.* **2017**, *5*, 8551–8567.
[105] P. Banoth, C. Kandula, P. K. Lavudya, S. Akaram, L. De Los Santos Valladares, R. Ammanabrolu, G. K. Mamidipudi, P. Kollu, *ACS Omega* **2023**, *8*, 18653–18662.
[106] Y. Jia, Y. Zhang, H. Xu, J. Li, M. Gao, X. Yang, *ACS Catal.* **2024**, *14*, 4601–4637.
[107] X. Zhang, S. Hua, L. Lai, Z. Wang, T. Liao, L. He, H. Tang, X. Wan, *RSC Adv.* **2022**, *12*, 17959–17983.
[108] R. Shwetharani, R. Kumar, *Inorg. Chem. Commun.* **2024**, *170*, 113255.
[109] S. Kumar, N. L. Reddy, S. H. Kushwaha, A. Kumar, M. V. Shankar, K. Bhattacharyya, A. Halder, V. Krishnan, *ChemSusChem* **2017**, *10*, 3588–3603.
[110] L. Liu, N. Liu, B. Chen, C. Dai, N. Wang, *Catalysts* **2024**, *14*, 126.
[111] M. A. Qadeer, X. Zhang, M. A. Farid, M. Tanveer, Y. Yan, S. Du, Z.-F. Huang, M. Tahir, J.-J. Zou, *J. Power Sources* **2024**, *613*, 234856.
[112] X. Chia, A. Y. S. Eng, A. Ambrosi, S. M. Tan, M. Pumera, *Chem. Rev.* **2015**, *115*, 11941–11966.


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

[113] B. W. H. Baugher, H. O. H. Churchill, Y. Yang, P. Jarillo-Herrero, *Nano Lett.* **2013**, *13*, 4212–4216.
[114] S. Dong Z. Wang, *Electrocatalysis* **2018**, *9*, 744–751.
[115] W. Raza, K. Ahmad, F. G. Alvarado, T. H. Oh, *Inorganics* **2025**, *13*, 47.
[116] F. Liu, X. Cai, Y. Tang, W. Liu, Q. Chen, P. Dong, S. Bao, *Energy Environ. Mater.* **2023**, *7*, e12644.
[117] Z. Zheng, L. Yu, M. Gao, X. Chen, W. Zhou, C. Ma, L. Wu, J. Zhu, X. Meng, J. Hu, Y. Tu, S. Wu, J. Mao, Z. Tian, D. Deng, *Nat. Commun.* **2020**, *11*, 3315.
[118] Y. H. Gu, M. F. Shao, J. Zhang, R. Li, N. Huang, Q. Liu, J. G. Jhao, W. Y. Zhang, X. H. Zhang, F. Peng, W. Q. Li, J. Li, *Langmuir* **2024**, *40*, 10405–10834.
[119] H. Chen, Y. Wang, S. Zhou, N. Li, X. Liu, A. Abdukayum, L. Zhang, G. Hu, *Ceram. Int.* **2024**, *50*, 16714–16721.
[120] M. B. Askari, A. Beheshti-Marnani, M. Seifi, S. M. Rozati, P. Salarizadeh, *J. Colloid Interface Sci.* **2019**, *537*, 186–196.
[121] I. Ahmad, Y. Zou, J. Yan, Y. Liu, S. Shukrullah, M. Y. Naz, H. Hussain, W. Q. Khan, N. R. Khalid, *Adv. Colloid Interface Sci.* **2023**, *311*, 102830.
[122] J. Wang, K. Fu, X. Zhang, Q. Yin, G. Wei, Z. Su, *J. Mater Chem.* **2021**, *9*, 8466–8482.
[123] A. Rahman, J. R. Jennings, A. L. Tan, M. M. Khan, *ACS Omega* **2022**, *7*, 22089–22110.
[124] Z. Saleem, E. Pervaiz, M. U. Yousaf, M. B. K. Niazi, *Catalysts* **2020**, *10*, 464.
[125] G. Mahalakshmi, M. Rajeswari, P. Ponnarasi, *Inorg. Chem. Commun.* **2020**, *120*, 108146.
[126] H. Ullah, Z. Haneef, A. Ahmad, I. S. Butler, R. Nasir Dara, Z. Rehman, *Inorg. Chem. Commun.* **2023**, *153*, 110775.
[127] V. Dhiman, N. Kondal, *Physica. B Condens. Mater.* **2022**, *628*, 413569.
[128] M. Wang, P. Ju, J. Li, Y. Zhao, X. Han, Z Hao, *ACS Sustainable Chem. Eng.* **2017**, *5*, 7878–7886.
[129] P. P. Singh, S. Sinha, G. Pandey, V. Srivastava, *RSC Adv.* **2022**, *12*, 29826–29839.
[130] V. Maivizhikannan, V. N. Rao, S. Kumar, A. Kumar, A. Kumar, M. V. Shankar, V. Krishnan, *Ceram. Int.* **2021**, *47*, 10301–10313.
[131] Y. Hu, X. Yu, Q. Liu, L. Wang, H. Tang, *Carbon* **2022**, *188*, 70–80.
[132] T. M. Chou, S. W. Chan, Y. J. Lin, P. K. Yang, C. C. Liu, Y. J. Lin, J. M. Wu, J. T. Lee, Z. H. Lin, *Nano Energy* **2019**, *57*, 14–21.
[133] J. Kisała, A. M. Ferraria, N. Mitina, B. Cieniek, P. Krzemiński, D. Pogocki, R. Nebesnyi, O. Zaichenko, Y. Bobitski, *RSC Adv.* **2022**, *12*, 22465–22475.
[134] G. J. Lee, Y. H. Hou, C. Y. Chen, C. Y. Tsay, Y. C. Chang, J. H. Chen, T. L. Horng, S. Anandan, J. J. Wu, *Int. J. Hydrog. Energy* **2021**, *46*, 5938–5948.
[135] B. Zhou, C. Wang, J. Lu, X. Zhang, X. Liu, D. P. Yang, S. Wu, *ACS Appl. Nano Mater.* **2024**, *7*, 7883–7895.
[136] S. Kar, S. Ghosh, T. Pal, *Sci. Rep.* **2024**, *14*, 18674.
[137] M. H. Fan, C. H. Wang, X. Yu, J. Ding, A. Q. Xiao, Y. Li, W. Y. Huang, *Tungsten* **2023**, *6*, 473–487.
[138] C. Feng, M. B. Faheem, J. Fu, Y. Xiao, C. Li, Y. Li, *ACS Catal.* **2020**, *10*, 4019–4047.
[139] Y. Qi, Y. Zhang, L. Yang, Y. Zhao, Y. Zhu, H. Jiang, C. Li, *Nat. Commun.* **2022**, *13*, 4602.
[140] Y. Wang, H. Su, Y. He, L. Li, S. Zhu, H. Shen, P. Xie, X. Fu, G. Zhou, C. Feng, D. Zhao, F. Xiao, X. Zhu, Y. Zeng, M. Shao, S. Chen, G. Wu, J. Zeng, C. Wang, *Chem. Rev.* **2020**, *120*, 12217–12314.
[141] G. Solomon, M. G. Kohan, R. Mazzaro, M. Jugovac, P. Moras, V. Morandi, I. Concina, A. Vomiero, *Adv. Sustain. Syst.* **2023**, *7*, 2200410.
[142] H. Zhang, X. Zhou, Y. Su, C. Du, *Fuel* **2024**, *371*, 131939.
[143] Y. Yang, Q. Wang, R. Aleisa, T. Zhao, S. Ma, G. Zhang, Y. Yin, **2021**, *13*, 51829–51838.
[144] M. H. Zhang, H. Dong, L. Zhao, D. X. Wang, D. Meng, *Sci. Total Environ.* **2019**, *670*, 110–121.
[145] Q. Yan, C. Lian, K. Huang, L. Liang, H. Yu, P. Yin, J. Zhang, M. Xing, *Angew. Chem., Int. Ed.* **2021**, *60*, 17155–17163.
[146] R. Jiang, D. Zhong, Y. Xu, H. Chang, P. Liao, Y. He, J. Zhang, *Colloids Surf. A Physicochem. Eng. Asp.* **2024**, *685*, 133184.
[147] P. Salarizadeh, M. B. Askari, A. Di Bartolomeo, *ACS Appl. Nano Mater.* **2022**, *5*, 3361–3373.

Manuscript received: November 5, 2024


© 2025 Wiley-VCH GmbH

23656549, 2025, 19, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/slct.202405247 by Delhi Technological University, Wiley Online Library on [25/05/2025]. 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

# Risk Analysis of Indian Banking Stocks Using Machine Learning Techniques

Vidit Tayal,
Delhi Technological University,
Delhi, INDIA
vidittayal21@gmail.com

Snehal Kumar
Delhi Technological University,
Delhi, INDIA
snehalkumar8429@gmail.com

Shweta Meena
Delhi Technological University,
Delhi, INDIA
shwetameena@dtu.ac.in

*Abstract*— **Indian banking stocks are highly volatile in nature. Although these stock values are dependent on their fundamentals. However, the banking stocks are susceptible to economic ups and downs, market sentiments, manufacturing factors and employment conditions, etc. Existing studies largely focus on implementing machine learning models for stock price prediction. However, minimal studies exist for stock risk analysis. In order to safeguard the investment of retail customers, there is a need to analyze the risk associated with them by simple tools in advance. In this paper, machine learning techniques viz. Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost) are investigated to predict the banking stocks risk. The statistical risk measures used are beta and value at risk (VaR). The simulation results indicate that the least Mean Absolute Error (MAE) for beta and Value at Risk are obtained by XGBoost and LSTM techniques respectively.**

***Keywords: Beta, Value at Risk (VaR), Long Short-Term Memory (LSTM), Extreme Gradient Boosting (XGBoost), Risk, Banking Stocks, Mean Absolute Error (MAE)***

## I. INTRODUCTION

Machine learning has gained substantial attention in financial risk management. This is due to its capability to identify complex patterns in large datasets. Credit Risk Assessment Model by Hu et al. [1] highlights the benefits of using various machine learning models over traditional methods to reveal non-linear relationships and refining the accuracy of risk predictions. Various ML algorithms, such as gradient boosting, support vector machines (SVM) and neural networks, have been successfully utilized to forecast market risks, including volatility and extreme tail events. To evaluate the financial time series data, the effective and proven techniques are LSTM networks, GANs and RNNs. The stock volatility may be analysed efficiently and effectively by RNNs and CNNs [2]. The high frequency data source acts as a source to fetch the patterns. To make precise risk predictions, the temporal dependencies can be fetched from the given data by application of LSTM models [3-7].

There is lack of satisfactory work in utilizing machine learning models in the domain of stock market risk analysis. Additionally, traditional machine learning models are rarely evaluated in live trading environments. Although existing machine learning models are undoing intense upgradations and advancements, the amount of research on stock market risk analysis using advanced machine learning models is very minimal. Therefore, there is a need to use advance machine learning algorithms for the analysis of the risk metrics associated with an investment portfolio. This work aims to address this research gap by comparing LSTM and XGBoost models for risk analysis. This study presents the application of computationally complex algorithms such as LSTM and XGBoost etc for the risks analysis of Indian banking stocks.

This paper is organised as follows-section II presents the literature review, section III discusses about risk relative to the market (beta), VaR is depicted in section IV, section V presents LSTM, section VI discusses about XGBoost, MAE is described in section VII, Simulation Results are presented in section VIII and finally in section IX, Conclusion is presented.

## II. LITERATURE REVIEW

The growth and financial stability of any country or society is mainly contributed by Banks [8]. They provide ease of doing business, various investment opportunities and lending facilities etc. For risk prediction in banking system, deep learning models are far superior in comparison with traditional models [9-14]. The fog computing algorithm is explored for market risk analysis. This utilized decision trees, ensemble methods and logistic infrastructure [15]. To judge the model accuracy, confusion matrix is used. Further, for comprehensive risk analysis, machine learning techniques including ANN. KNN, XGBoost and SVM are used. Various deep learning and mathematically intense algorithms and models should be tried against the stock prices historical dataset. Sujay Lokesh et al. [16] combined machine learning models with NLP techniques for stock market analysis. NLP is used to perform Sentiment Analysis on the input news data. These sentiments are then utilized to evaluate any changes in the stock market. Sentiment data can be collected using social media applications, web pages and financial news. This study provided a comprehensive stock risk analysis based on financial news. Their approach suggested that the integration of sentiment analysis techniques using NLP with stock historical data can improve the predictions for stock risk. The machine learning techniques such as Support Vector Machine (SVM), Decision Tree (DT) and Bayesian Networks (BN) are used to classify the payers in corporate financing by banks [17]. The best results were obtained by using SVM technique. The overview of various machine learning techniques for risks associated in banking sector is presented by Valentin et al [18]. Further, future directions were also presented by the study.

## III. MARKET RELATIVE RISK (BETA)

Greek letter beta (β) denotes the quantitative risk of an investment relative to NIFTY 50 or other market. The market is assumed to have a beta value of unity. Beta value more than

2025 3rd International Conference on Disruptive Technologies (ICDT) | 979-8-3315-1958-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICDT63985.2025.10986312


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 07:09:17 UTC from IEEE Xplore. Restrictions apply.

one indicates higher volatility of the security in comparison with market. The computation of beta provides an idea whether a particular stock is moving in synch or opposite to the market. The stocks with little risk have less potential of greater return and vice versa. Thus, in order to purchase a stock, investor should look at its beta value and fundamentals.

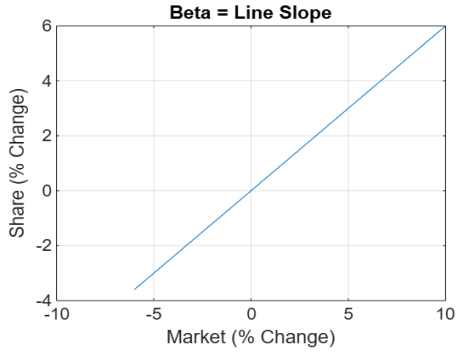



Fig. 1. Measure of Market Relative Risk (Beta)

The % change in a particular stock return vs %change in Market is depicted in Fig. 1. The slope of the straight line indicates the beta value. In a given time, the beta is mathematically represented by equation (1):

$$Beta = \frac{\text{Covariance (stock return, index return)}}{\text{Variance (index return)}} \quad (1)$$

Covariance is given by equation (2)-

$$Covariance = \frac{\sum (Stock\ Return - Stock\ Average\ Return) \times (Market\ Return - Market\ Average\ Return)}{Sample\ Size - 1} \quad (2)$$

## IV. Value At Risk (VaR)

Value at risk (VaR) is utilized to determine the maximum probable loss with certain degree of confidence for an investment. Value at Risk is utilized for risk management of a stock, trading position, or financial institution by quantitatively measuring the maximum potential loss that could incur over a given time. VaR is a valuable measure for quantifying market risk, particularly in the context of investment portfolios and trading activities. The VaR is mathematically represented by equation (3)-

$$VaR = \mu - \sigma \times Z \quad (3)$$

z indicates the Z score for a given confidence level, σ denotes the standard deviation and μ is the mean return of historical returns

### A. Working of VaR

Time Horizon: VaR calculations are performed for a given time horizon e.g. one month, one week, or one day. The choice of time depends on the risk management goals and the assets or positions being analyzed. Confidence Level: VaR is represented by a confidence level, often denoted as a percentage (e.g., 95% or 99%). This percentage represents the level of confidence that the calculated VaR will not be exceeded over the chosen time horizon. e.g. a 90% VaR shows that there is a 10% chance that losses will exceed the estimated VaR.

Probability Distribution: VaR is based on probability distributions of potential portfolio returns or losses. The frequently used distribution is the normal distribution (bell-shaped curve), but other distributions, such as the t-distribution or empirical distributions, can also be employed when assumptions about the normal distribution do not hold.

Portfolio Valuation: The portfolio's value is determined at the beginning and end of the time horizon. Changes in the VaR of the portfolio are used to calculate potential losses.

Volatility and Risk Factors: VaR models consider the historical volatility of the portfolio's assets or risk factors. Volatility measures how much the portfolio's value can fluctuate over time.

VaR assumes that market returns follow a specific probability distribution and may not capture rare but severe events (known as "tail risk"). Risk managers often use other risk measures alongside VaR to provide a more comprehensive assessment of risk, especially in situations where extreme events are of concern.

## V. Long Short Term Memory (LSTM)

LSTM models are a particular type of RNN that are very helpful in learning and modeling sequential data.

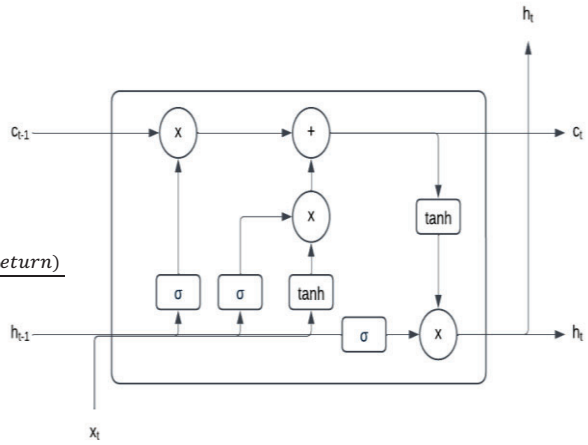



Fig. 2. LSTM Model

Standard RNNs face various problems with long-term dependencies because of vanishing issues and exploding gradients. LSTMs are designed to mitigate these challenges. They are mainly utilized for time-series data, where the sequence of previous inputs has a major impact on future predictions. This makes the LSTM model (Fig. 2) ideal for financial data and risk analysis in stock markets.

### A. Structure of LSTM

The ability to retain long-term memory through a sophisticated gating mechanism is held as the core advantage of LSTM networks. LSTMs comprise of memory cells that preserve data for prolonged periods, which enables the network to "remember" important patterns from past data. As depicted in Fig, 3, the memory cells are con-trolled by three primary gates- Forget Gate: The data from the previous time step to be discarded from the memory is decided by this gate, Input Gate: Decides which new data should be add into the memory cell, Output Gate: Controls which data from the memory should be used to compute the output for the current time step. This gating mechanism ensures that LSTMs can capture long term as well as short term dependencies within time series data and needed for accurately predicting stock prices and risk metrics.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 07:09:17 UTC from IEEE Xplore. Restrictions apply.

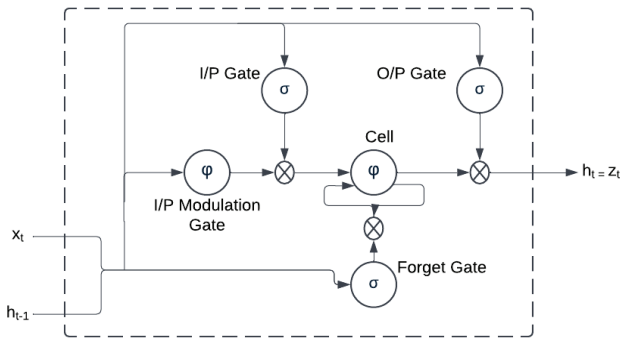



Fig. 3. Structure of LSTM cell

### B. Role of LSTM in Financial Risk Analysis

In financial risk analysis, market indicators like stock prices exhibit complex temporal patterns. LSTMs are particularly effective in capturing these temporal relationships by considering the influence of past stock prices on future values. For example, in risk metrics like volatility or value at risk (VaR), the model considers how the stock has moved over time and the potential for future fluctuations. The LSTM's ability to "remember" critical historical data points makes it an ideal tool for evaluating stock market risks. By training an LSTM model on historical stock price data, it might identify patterns that may indicate periods of high volatility, helping investors identify potential risks. LSTM networks have several advantages when applied to stock risk analysis: Handling Long-Term Dependencies: Financial data often contains trends and cycles that can span weeks, months, or even years. LSTMs can effectively model these long term dependencies to work better for forecasting future stock behavior and associated risks. Non-Linearity Handling: Stock market data is non-linear in nature, meaning that past prices do not always follow a simple linear pattern. These complex, non-linear relationships are captured by LSTMs, allowing them to make better predictions as compared to the predictions by linear models on similar data. Adaptability: LSTMs can adapt to changes in the market, learning to adjust their predictions as new data becomes available. This is mainly important in financial markets, where conditions can change rapidly. Flexibility in Input Data: LSTMs can be applied not only to stock prices but also to other types of time-series data like trading volume, sentiment indicators as well as macroeconomic factors, further enhancing their applicability to comprehensive risk.

### C. Training LSTM Models for Risk Metrics

To apply LSTM for stock risk analysis, the first task is to preprocess the historical stock price data, with the adjusted closing price feature. After preprocessing the data, the LSTM model is trained on this time-series data. While training, the model studies to associate past prices and features with future stock movements.

The following steps are typically involved in training an LSTM model for stock risk prediction:

1. Data Preprocessing: This includes scaling the stock prices and preparing them in a format that is suitable for time-series forecasting.
2. Model Architecture: A typical LSTM architecture consists of single or multiple LSTM layers and a dense layer for the final prediction. Dropout layers may also be included to prevent overfitting.
3. Training: Historical data is utilized for model training with suitable loss functions and optimization algorithms.
4. Evaluation: Once training is completed, the evaluation of model performance is carried out using test data that was not seen during training. Metrics like RMSE or MAE can be utilized to assess the performance.

## VI. Extreme Gradient Boosting (XGBoost)

XGBoost is an efficient, scalable, and robust machine learning library that has gained widespread popularity for its ability in regression and classification problems. A family of boosting algorithms, to which it belongs, combines number of weak learners i.e. decision trees to make a strong learner. XGBoost, as shown in Fig, 4, is an implementation of gradient boosting that includes several advanced features for performance enhancement, making it one of the most powerful algorithms for structured/tabular data. The errors generated by the former trees in the ensemble are corrected by each new tree trained in the ensemble built by XGBoost.

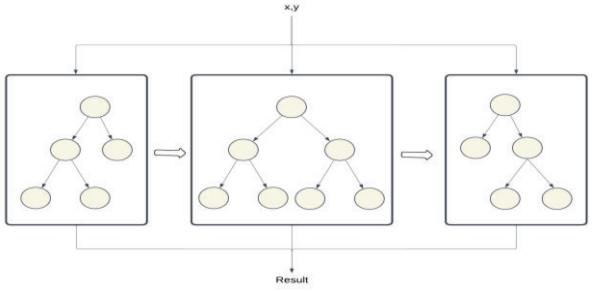



Fig. 4. XGBoost Model

Unlike traditional decision trees, which are built independently, the trees in boosting algorithms are built sequentially. This sequential nature allows XGBoost to focus on the most difficult cases where previous trees performed poorly. The loss function is minimized using gradient descent during the training process, with each tree being trained based on residual errors of the former tree. This boosts the performance of the model iteratively, producing an ensemble of weak learners that work together to make accurate predictions.

### A. Key Features of XGBoost

XGBoost stands out for several reasons as follows-

*1) Regularization:* A regularization term is included in the objective function of XGBoost, which helps prevent overfitting. The regularization can be adjusted using parameters like alpha (L1 regularization) and lambda (L2 regularization).

*2) Parallelization:* Unlike other gradient boosting implementations, XGBoost supports parallelization, meaning multiple trees can be built simultaneous-ly during training. This greatly speeds up the model's training process, particularly in complex datasets.

*3) Handling Missing Values:* XGBoost can manage missing data by learning the optimal direction to split the data even when some features are miss-ing, making it highly robust in real-world datasets.

*4) Tree Pruning:* XGBoost utilizes a method called max_depth to bound the depth of trees, which helps improve


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 07:09:17 UTC from IEEE Xplore. Restrictions apply.

generalization and prevents overfitting. Additionally, it uses depth-first pruning, which allows the algorithm to choose the most important splits in the data and prune less important branches.

*5) Weighted Quantile Sketch*: For handling weighted datasets, XGBoost uses a special algorithm called the weighted quantile sketch, which improves the algorithm's efficiency and accuracy in these cases.

### B. XGBoost in Stock Risk Analysis

XGBoost is efficient for analyzing and forecasting stock risks due to its capacity to manage structured data with tedious relationships. For stock risk analysis, features including market sentiment, trading volumes and historical prices data can be used as input variables for the model. By training on historical stock data, XGBoost can be used to predict various risk metrics, such as-

*1) Volatility: XGBoost can predict the expected range of price fluctuations, helping users track stock price movement within a specific time frame.*

*2) Beta:* The beta coefficient measures the stock's risk relative to the overall market.

*3) Value at Risk (VaR*): XGBoost can be applied to estimate the major loss in the stock value for a given confidence interval, helping investors assess potential risk exposure.

The capacity of XGBoost to manage huge datasets and model complex, non-linear relationships between features makes it particularly suitable for financial markets, where relationships between variables are often non-linear and influenced by multiple factors.

### C. Training an XGBoost Model for Risk Prediction

To apply XGBoost for stock risk analysis, the first step is to gather relevant data. This data includes - technical indicators, sentimental data, historical stock prices and fundamental data (e.g., Profit-Expense ratio). After gathering the data, the following steps are generally followed:

*1) Data Preprocessing:* This step involves data cleaning, handling missing values, and creating new features, including momentum indicators and moving averages. Feature scaling (e.g., normalization) may also be necessary to improve model performance.

*2) Model Training:* The dataset is distributed into testing set/training set. Using training set, the training of model is done, by means of the boost-ing technique in XGBoost. The performance is optimized by hyperparameter tuning.

*3) Evaluation:* Evaluation of the trained model is done by using MAE, MSE or RSE performance measures for regression problems. Moreover, cross-validation techniques may be used to confirm that the model fits properly with any unobserved data.

*4) Model Interpretation:* A feature importance score is provided by XGBoost, which helps interpret which features are most significant in prediction of the model. XGBoost is highly utilized in various financial applications to understand the key risk factors.

XGBoost is known for its high performance and accuracy in predictive tasks, making it a top choice for stock risk analysis. Due to its parallelization capabilities, XGBoost can process large datasets more efficiently than many other machine learning models, reducing training time significantly. XGBoost is robust to outliers and can handle noisy data better than many other algorithms. This is important in stock market data, where volatility and anomalies are common. XGBoost can be applied to different kinds of financial data, including time-series data, structured tabular data, and even sentiment data, making it versatile for different stock risk prediction tasks.

## VII. Mean Absolute Error (MAE)

Mean absolute error (MAE) is termed as quantitative measure of errors between predicted and observed values. In other words, it is defined as the sum of absolute errors divided by the sample size (n). This is used in machine learning regression models as a performance measure. The smaller the MAE, the better the model's predictions align with the actual data.

The MAE is expressed mathematically by equation (4):

$$MAE = \frac{1}{n}\sum_{=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

The number of observations is indicated by n, the true value is denoted by $y_i$ and the predicted value is shown by $\hat{y}_i$.

## VIII. Simulation Results

In this work, NIFTY 50 index and Indian Banking Sector Stocks are considered from January 2020 to May 2024. The data set is taken from the python library provided by Yahoo Finance. This has been observed that the MAE values (Table I & Fig. 5) for the calculation of beta, using LSTM, for Axis, Bandhan, Bank of Baroda, Canara, HDFC, ICICI, IDFC First, Indian, IndusInd, Karur Vysya, Kotak Mahindra, Punjab & Sind, Punjab National, State Bank of India, Union Bank of India, and Yes Banks are 0.011, 0.015, 0.014, 0.029, 0.009, 0.008, 0.013, 0.018, 0.013, 0.018, 0.009, 0.077, 0.024, 0.012, 0.025 and 0.023 respectively.

TABLE I. Mean Absolute Error (MAE) for Beta Prediction

| S.N. | Bank | Mean Absolute Error (MAE) | |
|---|---|---|---|
| | | LSTM | XGBoost |
| 1 | Axis Bank | 0.011 | 0.01 |
| 2 | Bandhan Bank | 0.015 | 0.014 |
| 3 | Bank of Baroda | 0.014 | 0.015 |
| 4 | Canara Bank | 0.029 | 0.014 |
| 5 | HDFC Bank | 0.009 | 0.008 |
| 6 | ICICI Bank | 0.008 | 0.008 |
| 7 | IDFC First Bank | 0.013 | 0.012 |
| 8 | Indian Bank | 0.018 | 0.018 |
| 9 | IndusInd Bank | 0.013 | 0.011 |
| 10 | Karur Vysya Bank | 0.018 | 0.014 |
| 11 | Kotak Mahindra Bank | 0.009 | 0.008 |
| 12 | Punjab & Sind Bank | 0.077 | 0.024 |
| 13 | Punjab National Bank | 0.024 | 0.017 |
| 14 | State Bank of India | 0.012 | 0.01 |
| 15 | Union Bank of India | 0.025 | 0.018 |
| 16 | Yes Bank | 0.023 | 0.019 |
| **Average MAE** | | **0.019** | **0.013** |

The MAE values using XGBoost for Axis, Bandhan, Bank of Baroda, Canara, HDFC, ICICI, IDFC First, Indian, IndusInd, Karur Vysya, Kotak Mahindra, Punjab & Sind, Punjab National, State Bank of India, Union Bank of India, and Yes Banks are 0.01, 0.014, 0.015, 0.014, 0.008, 0.008, 0.012, 0.018, 0.011, 0.014, 0.008, 0.024, 0.017, 0.01, 0.018 and 0.019 respectively. The average MAE values for LSTM


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 07:09:17 UTC from IEEE Xplore. Restrictions apply.

and XGBoost are 0.019 and 0.013 respectively. Thus, for the prediction of beta, the average MAE value for XGBoost is far smaller than the average MAE value for LSTM.

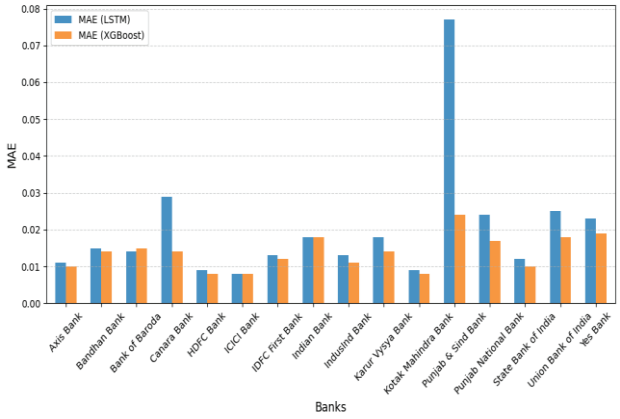



Fig. 5. Mean Absolute Error (MAE) Comparison of LSTM and XGBoost for Beta Prediction

Table II depict that the MAE values for the calculation of VaR, using LSTM, for Axis, Bandhan, Bank of Baroda, Canara, HDFC, ICICI, IDFC First, Indian, IndusInd, Karur Vysya, Kotak Mahindra, Punjab & Sind, Punjab National, State Bank of India, Union Bank of India, and Yes Banks are 0.002, 0.006, 0.002, 0.003, 0.003, 0.003, 0.005, 0.004, 0.003, 0.005, 0.002, 0.004, 0.003, 0.002, 0.004 and 0.004 respectively. These values are further pictorially presented for better understanding by means of graphs as shown in Fig 6.

TABLE II. MEAN ABSOLUTE ERROR (MAE) FOR VAR PREDICTION

| S.N. | Bank | Mean Absolute Error (MAE) | |
|---|---|---|---|
| | | LSTM | XGBoost |
| 1 | Axis Bank | 0.002 | 0.007 |
| 2 | Bandhan Bank | 0.006 | 0.008 |
| 3 | Bank of Baroda | 0.002 | 0.011 |
| 4 | Canara Bank | 0.003 | 0.016 |
| 5 | HDFC Bank | 0.003 | 0.009 |
| 6 | ICICI Bank | 0.003 | 0.005 |
| 7 | IDFC First Bank | 0.005 | 0.006 |
| 8 | Indian Bank | 0.004 | 0.01 |
| 9 | IndusInd Bank | 0.003 | 0.012 |
| 10 | Karur Vysya Bank | 0.005 | 0.017 |
| 11 | Kotak Mahindra Bank | 0.002 | 0.006 |
| 12 | Punjab & Sind Bank | 0.004 | 0.015 |
| 13 | Punjab National Bank | 0.003 | 0.013 |
| 14 | State Bank of India | 0.002 | 0.011 |
| 15 | Union Bank of India | 0.004 | 0.025 |
| 16 | Yes Bank | 0.004 | 0.010 |
| **Average MAE** | | **0.003** | **0.011** |

The MAE values using XGBoost for Axis, Bandhan, Bank of Baroda, Canara, HDFC, ICICI, IDFC First, Indian, IndusInd, Karur Vysya, Kotak Mahindra, Punjab & Sind, Punjab National, State Bank of India, Union Bank of India, and Yes Banks are 0.007, 0.008, 0.011, 0.016, 0.009, 0.005, 0.006, 0.012, 0.017, 0.006, 0.015, 0.013, 0.011, 0.025 and 0.010 respectively.

The average Mean Absolute Error (MAE) values for LSTM and XGBoost are 0.003 and 0.011 respectively. Thus, for the prediction of VaR, the average MAE value obtained from LSTM is far smaller than the average MAE value obtained from XGBoost. This indicates the superiority of LSTM technique over the XGBoost technique in risk analysis of Indian Banking stocks.

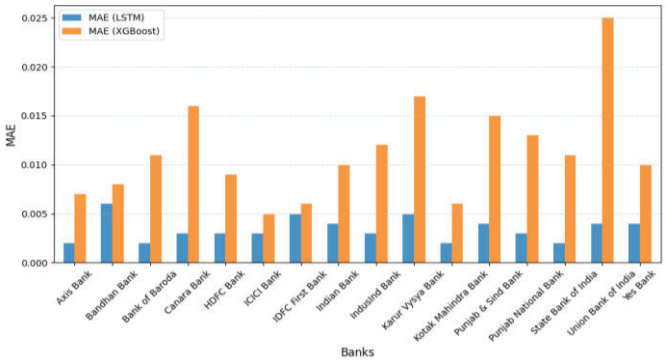



Fig. 6. Mean Absolute Error (MAE) Comparison of LSTM and XGBoost for VaR Prediction

## IX. CONCLUSION

Due to frequent volatile conditions in stock markets, there is a financial loss to retail investors daily. This generates a loss of confidence among them to take correct investment decisions. Thus, there is a need for suitable machine learning techniques to analyze the risk associated with a particular stock in comparison with benchmark market index such as NIFTY50. In this work two machine learning techniques namely LSTM and XGBoost are explored for risk analysis of sixteen Indian banking stocks. The performance measures used are beta and Var. For beta prediction, smaller average MAE value from XGBoost technique ensures its better efficacy in comparison with LSTM technique. Whereas for the prediction of VaR, LSTM is more precise than XGBoost due to its smaller average MAE value. Thus, XGBoost is preferred for beta prediction, while LSTM is the better choice for VaR prediction. This risk analysis may help retail as well as financial institutional investors to take correct decisions for investment in banking stocks.

In general, a retail investor only analyzes the historical prices for a stock to make an investment. Thus, in the study only historical stock prices are considered as the only input for analysis. However, there are many other factors which impact the price of a stock, such as news sentiments, stock market events, government taxation rules modifications, etc. Thus, this study may be further extended by including the data set based on various rare market events or government/corporate structural breaks.

## REFERENCES

[1] X. Hu, J. Hu, L. Chen and Y. Li, "Credit Risk Assessment Model for Small, Medium and Micro Enterprises Based on RS-PSO-SVM Integration", IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, pp. 342-345, 2021.

[2] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model", International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, pp. 1643-1647, 2017.

[3] M. A. Istiake Sunny, M. M. S. Maswood and A. G. Alharbi, "Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model", 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, pp. 87-92, 2020.

[4] Y. Wen, P. Lin, and X. Nie, "Research of stock price prediction based on PCA-LSTM model", OP Conf. Ser., Mater. Sci. Eng., 790, 2020.

[5] Y. Song and J. Lee, "Importance of event binary features in stock price prediction", Appl. Sci., vol. 10, issue 5, pp. 1597, 2020.

[6] A. Kelotra and P. Pandey: Stock market prediction using optimized Deep-ConvLSTM model. Big Data, vol. 8, issue 1, pp. 5-24, 2020.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 07:09:17 UTC from IEEE Xplore. Restrictions apply.

[7] S. Borovkova and I. Tsiamas: An ensemble of LSTM neural networks for high-frequency stock market classification. J. Forecasting, vol. 38, issue 6, pp. 600–619, 2019.

[8] Chockalingam, A., Dabadghao, S. and Soetekouw, R., "Strategic risk, banks, and Basel III: estimating economic capital requirements", Journal of Risk Finance, Vol. 19 No. 3, pp. 225-246, 2018.

[9] P. Patil, C. Wu, K. Potika, and M. Orang: Stock Market Prediction using Ensemble of Graph Theory, Machine Learning and Deep Learning Models. Proc. 3rd Int. Conf. Softw. Eng. Inf. Manage., pp. 85–92, 2020

[10] A. Fazeli and S. Houghten, "Deep learning for the prediction of stock market trends", Proc. IEEE Int. Conf. Big Data (Big Data), pp. 5513–5521, 2019.

[11] G. S. Atsalakis and K. P. Valavanis, "Surveying stock market forecasting techniques", Part II: Soft computing methods. Expert Syst. Appl, pp. 5932–5941, 2009.

[12] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions", Artif. Intell. Rev., vol. 53, pp. 3007–3057, 2019.

[13] Y. Xu and V. Keselj, "Stock prediction using deep learning and sentiment analysis", Proc. IEEE Int. Conf. Big Data (Big Data), pp. 5573–5580, 2019.

[14] M. Agrawal, A. U. Khan, and P. K. Shukla, "Stock price prediction using technical indicators: A predictive model using optimal deep learning", Learning, vol. 6, issue 2, pp. 7, 2019.

[15] G. Wali, C. Bulla, A. Kori, "Market Risk Assessment Using Deep Learning Model and Fog Computing Infrastructure", International Journal of Advances in Engineering Architecture Science and Technology, pp. 31-42, 2024.

[16] S. Lokesh, S. Mitta, S. Sethia, S. Kalli and M. Sudhir, "Risk Analysis and Prediction of the Stock Market using Machine Learning and NLP", International Journal of Applied Engineering Research, pp. 16036-16041, 2018.

[17] Fenerich, M. Steiner, P. Steiner Neto, E. Tochetto, D. Tsutsumi, F. Assef and B. Samways dos Santos, "Use of machine learning techniques in bank credit risk analysis", Rev. int. métodos numér. cálc. diseño ing. Vol. 36, (3), no. 40, pp. 1-15, 2020.

[18] Valentin Lennart HeB and Bruno Damasio, "Machine learning in banking risk management: Mapping a decade of evolution", International Journal of Information Management Data Insights, vol. 5, pp. 1-17, Issue 1, 2025.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 20,2025 at 07:09:17 UTC from IEEE Xplore. Restrictions apply.



IETE Journal of Research

ISSN: 0377-2063 (Print) 0974-780X (Online) Journal homepage: www.tandfonline.com/journals/tijr20

# Routing Optimization in Hybrid Software-Defined Networks: A Heuristic Approach

Samiullah Mehraban & Rajesh K Yadav

**To cite this article:** Samiullah Mehraban & Rajesh K Yadav (22 May 2025): Routing Optimization in Hybrid Software-Defined Networks: A Heuristic Approach, IETE Journal of Research, DOI: 10.1080/03772063.2025.2506011

**To link to this article:** https://doi.org/10.1080/03772063.2025.2506011

Published online: 22 May 2025.

Submit your article to this journal

Article views: 3

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tijr20

IETE JOURNAL OF RESEARCH
https://doi.org/10.1080/03772063.2025.2506011

Check for updates

# Routing Optimization in Hybrid Software-Defined Networks: A Heuristic Approach

Samiullah Mehraban and Rajesh K Yadav

IoT lab, Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

**ABSTRACT**

In modern networking infrastructure, integrating Software-Defined Networking (SDN) ideas with traditional networking methodologies has given rise to Hybrid Software Defined Networks (HSDN); this type of network leverages the benefits of both conventional and SDN networks, the flexibility and programmability of SDN with the familiarity and stability of conventional network architectures. As both network devices integrate in a hybrid network, Traffic Engineering (TE) necessitates creative solutions to improve network efficiency, performance, and resource utilization. In this article, we studied the challenges associated with TE in a hybrid SDN environment, where traditional network devices coexist alongside SDN nodes; we investigated the optimal migration sequence from a conventional network to a hybrid SDN to address the optimization impact on controllable traffic and then for the routing optimization in the migrated hybrid network, we introduced a heuristic algorithm (H-STE) that optimizes both the OSPF weight setting and the splitting ratio of SDN nodes in the hybrid environment. Extensive tests on real network topologies are used to validate our methods, and the results demonstrate that with a 30% SDN deployment ratio, we can control more than 70% of network traffic; we observe a reduction in the Maximum Link Utilization (MLU) as it stabilizes, leading to a near-optimal result.

**KEYWORDS**

SDN; HSDN; TE; MLU; QoS; BGP

## ABBREVIATIONS

| | |
|---|---|
| SDN | Software Defined networking |
| hSDN | Hybrid software-defined networking |
| TE | Traffic Engineering |
| MLU | Maximum Link Utilization |
| IoT | Internet of Things |
| IoE | Internet of Everythings |
| OF | OpenFlow |
| OSPF | Open Shortest Path First |
| ECMP | Equal Cost Multiple Path |
| ONOS | Open Network Operating System |
| BGP | Border Gateway Protocol |
| AS | Autonomous System |
| CPLEX | Constraint Programming Linear Programming with Extensions |
| FCM | Flow Control Method |
| FRR | Free Range Routing |
| VTY | Virtual Teletype |

## 1. INTRODUCTION

The primary challenge with the existing network design is network Management, which both experts and industry require assistance with. To solve these challenges, a number of ideas and architectures have been proposed worldwide for future network design. Software-defined networking (SDN), whose core idea is the separation of data and control layers, was developed in response to these challenges. SDN increases network throughput and link utilization by enabling flexible and dependable network management through its global view and centralized management. Additionally, it offers a special flow scheduling method to improve traffic engineering (TE). SDN networks have gained a lot of attention in the telecommunications industry in recent years. With the development of technologies like mobility, virtualization, and the Internet of Things (IoT), this crucial and innovative network design idea came into being.

The fast growth of Internet applications, such as games, chat tools, and movies, has resulted in a noticeable increase in traffic on Internet Service Providers' (ISPs') networks in recent decades. In order to provide a network with maximum bandwidth and minimal latency, ISPs have been investing more in building their network infrastructure; yet, this increase in traffic has led to frequent congestion, which has resulted in a lack of assured customer experience and satisfaction. Emerging internet applications have caused network traffic to skyrocket as a result of the Internet of Things (IoT) and the

© 2025 IETE

Internet of Everything's (IoE) exponential progress [1]. Consequently, there is a pressing need to develop network management technologies while concurrently focusing on developing ways to prevent network congestion and achieve traffic load balance.

By optimizing traffic routing techniques, traffic engineering is a very powerful network administration tool that balances network load and improves network performance [2]. Over the last few decades, TE has attracted a lot of attention from both industry and academics because of its effectiveness in balancing traffic flow, preventing congestion, and improving network performance. It also gives users control over how data is routed over the network. However, traffic engineering optimization is still an unexplored field of study because of the network's complex structure.

SDN is a well-known network design that separates the data and control planes of the network. From a TE standpoint, it provides a versatile and effective method for optimising traffic routing. SDN switches can distribute traffic to multiple paths in a flexible and unrestricted way, and policies or routing algorithms can be implemented in the control plane of an SDN network to provide flexible control over the distribution of traffic to any outgoing links in SDN switches [3,4]. By customizing the forwarding rules for each flow, SDN networks may attain more intelligent, adaptable, and user-friendly network control [5].

An SDN network with a centralized controller can manage all networks with programmable devices. In terms of traffic engineering, link utilization, and network throughput, centralized management may greatly improve overall network performance, outperforming traditional networks. In recent years, SDN has drawn a lot of attention from around the world, and network operators have been adopting it at a much higher rate. This is because network operators can now dynamically control the routing flows within the network and quickly change the routing choice when necessary. For example, Microsoft [3] and Google [4] have already built a full SDN-enabled networks, which can achieve almost 100% link utilisation. By enabling dynamic and timely network route alterations, the SDN network offers substantial benefits in traffic engineering, successfully resolving unforeseen issues such as traffic fluctuations, connection failures, etc. Because SDN flexibly optimizes TE, the SDN network is a distinctive and alluring technology.

Figure 1 illustrates the fundamental architecture of three distinct network designs: (a) represents the conventional network architecture, (b) depicts the fully SDN architecture, and (c) presents the hybrid SDN design.

Despite the strong motivation and numerous benefits, the direct transition to a fully SDN network is a difficult task; upgrading the entire conventional network to SDNs appears risky and expensive, particularly for large networks with thousands of conventional network devices. These challenges include technical, economic, security, service outages, and more. Thus, the ideal approach to deal with these issues would be to use a hybrid SDN and deploy SDN nodes incrementally. Because SDN and conventional IP network devices coexist, a hybrid SDN blends the programmability and flexibility of an SDN network with the stability and familiarity of traditional network architecture. Organization and service providers may incrementally transition to a hybrid SDN network before moving on to a completely SDN network, negating the need to update all of the conventional network equipment to the SDN network.

Hybrid SDN has become a more prevalent network design in recent years. Furthermore, business and academics have shown a great deal of interest in the study of traffic engineering in the hybrid SDN [6]. One of the most often used internal gateway protocols in traditional IP networks is OSPF (Open Shortest Path First) [7]. In a conventional IP network, each connection is given a weight (or cost), and routing protocols that give priority to shortest path routing, such as OSPF, are used to route traffic from the source to the destination nodes via the shortest pathways (least cost). When there are several shortest paths, traffic is evenly split across Equal Cost Multiple Paths (ECMP). Consequently, the weights given to the network have a direct impact on how successful TE is in a traditional IP network. Conventional legacy devices can only support the OSPF protocol in a hybrid SDN; on the other hand, SDN network nodes operate in hybrid mode by supporting both OSPF and OpenFlow protocols.

The main topic of this study is a hybrid SDN architecture where the SDN nodes are incrementally deployed; we focus on the effective deployment of the hybrid SDN. We study and address both the SDN migration and minimizing MLU problems, since they are closely related and intricately entwined. In order to maximize controlled traffic and minimize the network's MLU in a hybrid SDN environment, we first tackle the SDN migration challenge, which involves migrating from a conventional network to an SDN network or successfully implementing a hybrid SDN. Second, how can traffic routing be optimized in a hybrid SDN environment, or how can

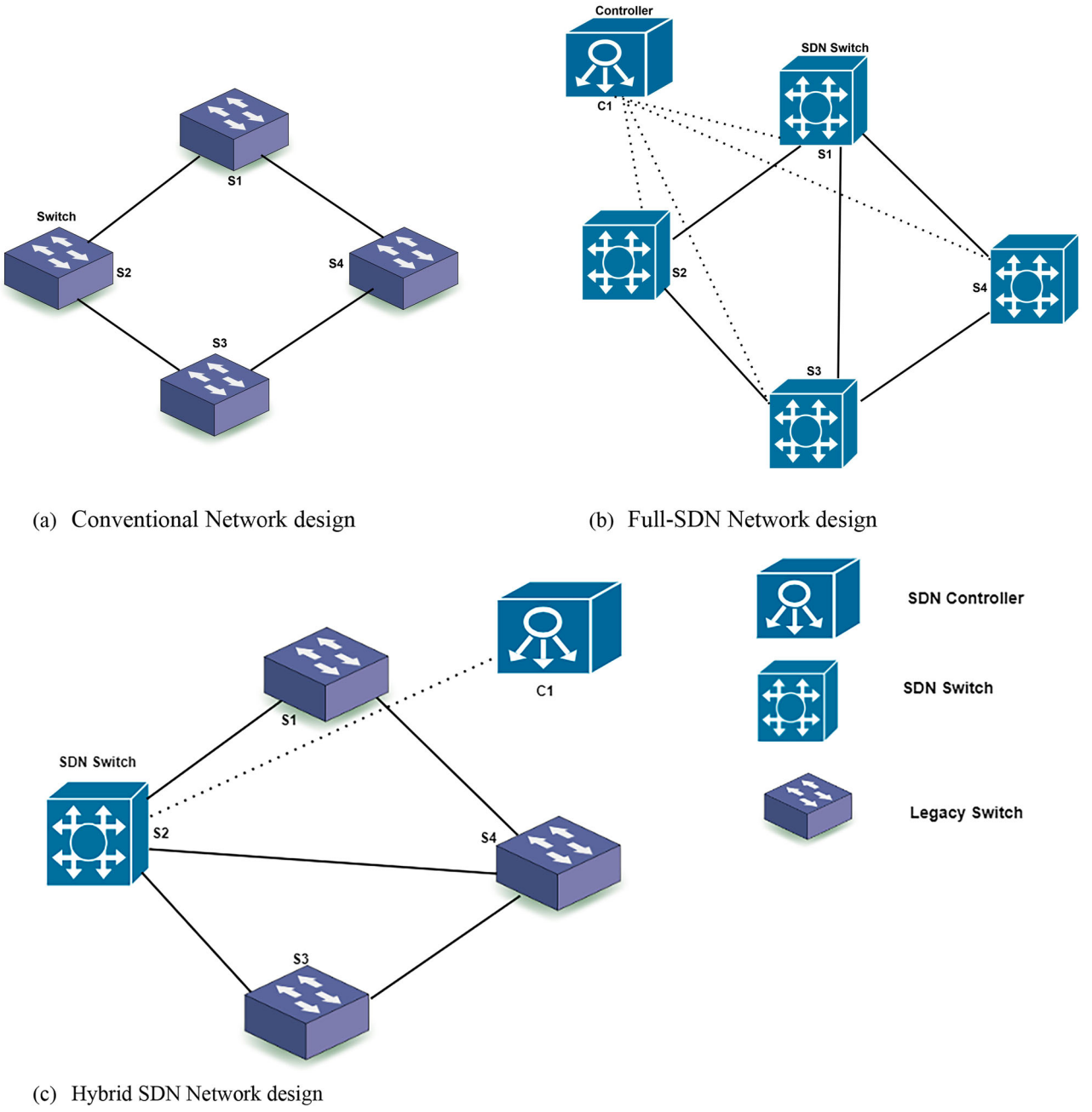



**Figure 1:** Different Network Designs: (A) Conventional Network design (B) Full-SDN Network design (C) Hybrid SDN Network design

MLU be decreased in the hybrid SDN once SDN nodes are deployed? We presented H-STE (Hybrid SDN Traffic Engineering), a heuristic algorithm for routing optimisation that optimizes both SDN nodes splitting ratios with the OSPF weight setting. This improves routing flexibility, reduces MLU, and boosts network performance. The H-STE method helps bridge the gap between both networks by optimizing the whole network performance while adhering to the limitations imposed by traditional routing protocols. This method adjusts the splitting ratio of SDN nodes to reduce network congestion while minimizing maximum link utilization in a hybrid SDN netwrok. Compared to the traditional routing method, our method can minimize MLU more efficiently in a hybrid SDN.

Furthermore, the experimental analysis indicates that upgrading only 30% of SDN devices can yield the most benefits and achieve nearly optimal performance. This paper first addresses the SDN migration sequence problem for maximum controllable traffic in a hybrid SDN. Then, we proposed a heuristic algorithm to minimize MLU in a hybrid SDN by optimizing the traffic splitting ratio of SDN nodes and the OSPF weight setting

of conventional network devices. Lastly, we evaluate and examine the effects of our suggested method on various topologies and real network traffic datasets. The key contributions of our study are summed up as follows:

- First, investigate the SDN migration sequence to maximize controllable traffic and affect the traffic control approach.
- To minimize MLU in a hybrid SDN, we introduced a heuristic algorithm that optimizes the traffic splitting ratio of SDN nodes and the OSPF weight setting of conventional network devices.
- Integrating conventional network devices with SDN nodes and ONOS controllers.
- After evaluating and comparing our proposed algorithm on various topologies and network traffic datasets, our experimental analysis demonstrates that with 30% of SDN device deployment, we can get the most significant benefits, resulting in nearly equivalent performance.

The rest of the article is structured as follows: Part 2 presents the related work. Part 3 presents the system description. Part 4 introduces the algorithms. Part 5 discusses the experiments and analysis of our method on various network topologies. Finally, Part 6 presents the conclusion.

## 2. RELATED WORK

In this part, we discuss the related works on migration sequence and routing optimization challenges in a hybrid SDN environment. In recent years, researchers and industries have been drawn to the hybrid SDN, with the many aspects that can be optimized while migrating to a hybrid SDN, different algorithms have recently been deployed to optimize different network characteristics. Guo et al. [8] investigate the incremental deployment method for the hybrid SDN by presenting it as a traffic engineering optimization problem, identifying the location and the number of migrated devices to minimize MLU in the network. They proposed greedy and genetic searching algorithms as a solution to this challenge. Their findings demonstrated that the genetic algorithm outperformed other techniques in identifying optimal migration sequences, as with the deployment of 40% of the nodes, we can get optimal MLU. Kar et al. [9] studied the transition of traditional network devices to the hybrid SDN by framing improving hop/path coverage while reducing cost; they proved that the optimization issue in this transition is an NP-hard problem. They proposed two different heuristic algorithms to model the migration sequence problem concerning path/hop coverage. Their result showed that their algorithms achieve more hop and path coverage with the same investment as previous algorithms. Jia et al. [10] proposed a heuristic algorithm for the incremental deployment of a hybrid SDN to maximize network control ability while minimizing upgrade costs. The simulation result showed that with a 10% upgrading cost, it is possible to control 95% of the network flow through SDN devices. Yang et al. [11] concentrate on the problem of selecting suitable SDN candidates, which involves identifying the least number of legacy routers to upgrade to SDN controllers to defend against any single point of failure within a specified network. The technique they created intends to incrementally convert traditional network devices to a hybrid SDN by limiting the repair path of damaged links while reducing delay and improving network bandwidth utilization. Xu et al. [12] proposed a heuristic algorithm for the incremental deployment of hybrid SDN under budget constraints, and they used randomized rounding and depth-first search techniques to solve multi-commodity issues in a hybrid SDN to maximize performance.

The primary focus of routing optimization algorithms in conventional distributed networks is often on improving the OSPF link weight setting. Fortz et al. [13] presented an IP-based approach for optimizing intra-domain routing. They used a sophisticated tabu search heuristic technique for the best OSPF weight configuration. Similarly, Ericsson et al. [14] present a genetic method to optimize OSPF weight setting, as traffic in conventional networks is restricted to the shortest path, resulting in a lack of flexibility in routing possibilities. Srivastava et al. [15] provide a Lagrange relaxation-based strategy to enhance traditional network routing. Prior research efforts have primarily focused on optimizing routing in pure conventional or fully SDN networks, and these techniques cannot be applied directly to a hybrid SDN environment. This is the first study that addresses and elucidates TE concerns in hybrid SDN [16]. Agarwal et al. define TE as a linear programming problem in this study. They presented the Fully Polynomial Time Approximation approach to balance the flow and enhance routing in a hybrid SDN. A module for the incremental implementation of hybrid SDN is presented by Hong et al. [17], who also offer a variety of heuristic approaches to lower MLU from a traffic engineering standpoint. To enable fine-grained traffic control across subdomains, Caria et al. separated the OSPF domain into subdomains and implemented SDN at the border routers in paper number [18]. He et al. [19] study and examine TE in two distinct hybrid network modes, namely barrier mode and hybrid mode. The two-hybrid technique provides a polynomial-time

approximation strategy for traffic engineering problems with an approximation of (1+w). Guo et al. suggest SOTE in [20] to optimize the OSPF weight setting and SDN node splitting ratio simultaneously in a hybrid SDN; in their case, they jointly optimized the OSPF weight setting and SDN node traffic splitting ratio. Xu et al. [21] examined the classical integration of SDN switching with all devices in a fully SDN environment.

Much research has been conducted recently to solve TE challenges in the network using ML and DL. Yinga et al. [22] proposed an RL-based (Reinforcement Learning) approach to handle the traffic that is constantly changing to achieve link load balancing in the hybrid SDN network. Xu et al. [23] developed a learning-based traffic engineering approach, TEAL, to optimize WAN traffic engineering by utilizing GPUs' parallel processing capabilities. Liu et al. [24] present DRL-OR, a multi-agent deep reinforcement learning-based online routing algorithm, to enhance the Quality of Service (QoS) by optimizing traffic flow routing with diverse QoS requirements. Lin et al. [25] proposed a model to integrate the transformer model with deep reinforcement learning to enhance traffic engineering performance in a hybrid SDN with fluctuating traffic demands. Guo et al [26] proposed a multi-agent RL-based traffic engineering method that promptly chooses the network flow path in hybrid SDNs.

For routing optimization in a hybrid SDN environment, both the OSPF weight setting and the splitting ratio of SDN nodes must be optimized, as both network devices coexist and communicate. Guo et al. [27] formulated the traffic engineering problem in hybrid SDN over multiple traffic matrices by mixing online splitting ratio optimization with offline weight setting optimization. Yingya et al. [28] formulated the routing optimization problem in a hybrid SDN with path cardinality constraints and proposed the H-permissible Path Routing Scheme, which effectively routes traffic flows under path cardinality constraints. Tajiki et al. [29] present a novel traffic engineering idea for an SDN/MPLS network, combining SDN with MPLS technologies, while our emphasis is on a hybrid SDN architecture. These hybrid network scenarios addressed in the studies differ from the hybrid SDN scenario investigated in our context. This paper addresses two main issues in the hybrid SDN scenario; firstly, we address the SDN migration sequence problem to maximize controllable traffic in a hybrid SDN environment. We have introduced a simple greedy approach for upgrading traditional legacy devices to the SDN nodes, and then, for routing optimization, we introduce a heuristic algorithm to minimize MLU in a hybrid SDN by optimizing the traffic splitting ratio of SDN nodes and the

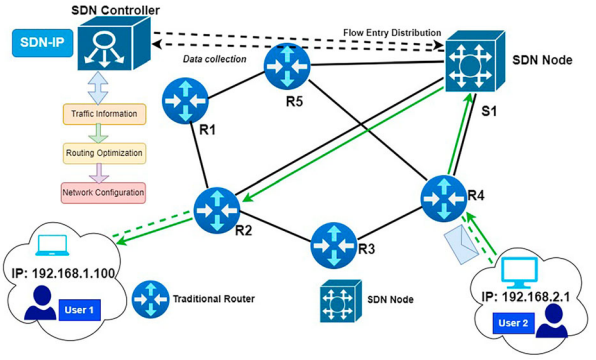



**Figure 2:** Hybrid SDN design

OSPF weight setting of conventional network devices. We have built the conventional network setup using Mininet and Mini-Edit tools, then incrementally deployed SDN nodes into our topology. Selecting the controller in the hybrid SDN is another issue that needs to be considered; in this simulation, we have studied and tested multiple controllers. Finally, the ONOS controller is chosen to be the SDN controller, and the integration between the legacy network and the SDN network is carried out through the SDN-IP. This program operates on top of the ONOS controller and allows SDN nodes to communicate with legacy routers using BGP protocols.

## 3. SYSTEM DESCRIPTION

In this part, we initially present the hybrid SDN architecture. Subsequently, we declare the variables needed in this optimization problem. Ultimately, we analyze the problem's complexity and prove it to be an NP-complete problem.

Figure 2 shows the Hybrid SDN network scenario in which conventional network devices coexist with SDN network devices. R1-R5 represents the traditional router, and S1 represents the SDN switch.

### 3.1 Hybrid SDN Network

This part describes the hybrid software-defined network that is considered in our paper. A hybrid SDN integrates traditional network equipment with SDN to enhance user experience and performance by regulating, modifying, configuring, and managing network behaviors. It combines SDN's centralized management and programmability with conventional networks' reliability and stability. While traditional devices ensure reliable routing, SDN nodes focus on network optimization. Hybrid SDN comprises three main components: conventional network devices, SDN nodes, and controllers. In this

architecture, traditional devices manage certain network parts, such as access layers or legacy systems; in contrast, centralized controllers will control the other parts of the network, such as data centers or virtualized environments. Hybrid SDN combines the advantages of centralized management and programmability of SDN with the familiarity, reliability, and stability of conventional networks. Conventional network devices provide reliable routing, whereas SDN devices focus on network optimization. Both network devices coexist in this type of networking, as shown in Figure 2; the network design includes R1-R5, which are conventional legacy routers responsible for routing traffic based on the shortest path between source and destinations; additionally, S1 represents an SDN node; with a controller in place to control the SDN node.

There is no need to upgrade all the conventional network devices to SDN; we can deploy SDN nodes incrementally to migrate to a hybrid SDN first, then eventually to a fully integrated SDN network. Traditional network devices support distributed routing protocols like IGP and OSPF, and forward the packets to the destinations based on the shortest path defined by the OSPF protocol. On the other hand, SDN devices work in hybrid mode, which supports both OSPF and SDN protocols, OpenFlow [30]. The SDN controller manages the forwarding operations of SDN nodesand routes network traffic from a worldwide perspective. It provides the flow entries to the flow table of SDN nodes, and the devices subsequently forward the packets based on the guidelines defined in the deployed rules. In this case, in a hybrid SDN, some traffic will be managed through conventional network devices, and the rest of the traffic will be managed flexibly by the SDN centralized controller.

The network's most important component, the SDN controller, is in charge of overseeing all data plane functions. Achieving optimal network performance depends heavily on the controller's efficacy and efficiency. With so many alternatives available, choosing the right SDN controller, depending on the network architecture, is essential to achieving optimal network performance. Since there are several controllers available for the hybrid SDN environment, such as NoX, POX, Floodlight, Ryu, ODL, and ONOS controllers. After careful consideration, we have chosen the ONOS controller as the SDN controller for our hybrid SDN topology. This decision was made as the ONOS controller is compatible with the hybrid SDN network and can support SDN-IP applications. SDN-IP is an application program that integrates traditional and SDN devices; it operates on the ONOS controller and enables SDN nodes to connect and exchange information with traditional routers using Border Gateway Protocol (BGP) and Autonomous Systems (AS).

**Table 1: Key notations**

| Variable | definitions |
|---|---|
| $G = (V, E)$ | Undirected graph |
| V | Set of all forwarding devices |
| E | Set of all links in the network |
| Vl | legacy node set |
| Vs | SDN node set |
| $V_c$ | Subset of migrated nodes |
| $C(e)$ | Capacity of link $e$ |
| U | MLU value |
| $\omega(e)$ | Link weight |
| $D$ | Set of Traffic Matrices |
| $\omega$ | OSPF weight matrix |
| $P_{sd}$ | In the OSPF protocol, the shortest path from s to d |
| $L^t_{sd}$ | In a period of t, the traffic weight of flow from s to d |
| $r_i$ | Weight coefficient of matrix $D_i$ |
| $f$ | Splitting ratio of SDN devices. |
| Ucn | Link utilization of SDN devices cn |

A number of factors, including load balancing, traffic engineering, network performance, security, and link failure recovery, must be taken into account while building an SDN network. Deploying SDN nodes and controllers incrementally into traditional networking is a workable way to benefit from both ideas at the same time. This method minimizes network service interruptions while improving overall network performance.

### 3.2 Variable Declaration

We made the network a graph, where *Vrepresents* the set of all forwarding nodes, and E represents the set of links in the network. Each forwarding device can either be an SDN device $V_s$ or a legacy device $V_l$, where $V_l \cap V_S = \emptyset$. For $\forall\, \upsilon 1, \upsilon 2 \in V, \upsilon 1 \neq \upsilon 2$, for $\upsilon 1$ to $\upsilon 2$, at least one path should exist to reach. Vc represents a subset of migrated SDN devices. Matrix C denotes the link's capacity $e \in E$ and $\omega(e)$ represents the link weight configuration that implements the OSPF routing protocol $e \in E$. D represents the Traffic Matrices (TMs) set, where D = D1, D2, … Dn and $r_i$ represent the coefficient of traffic matrices $D_i$. The variable $e$ represents the traffic splitting ratio of the SDN devices. We use N = /V/ to denote the number of devices in the network. All the sets of notations used in this paper are summarized in Table 1 for a better understanding.

The purpose of TE in a hybrid SDN for the given set of traffic matrices is to improve the overall network performance by minimizing the MLU of the network. Our main objective is to minimize the MLU with the minimum number of SDN nodes while maintaining the network's load balance to improve network performance. To discover the best migration sequence $S = (v1, v2, v3 \ldots vs)$

to control the traffic and minimize the MLU of the network after migrating legacy devices. The network topology and traffic matrix are considered to be fixed for this experiment. In each step, there are two routes for traffic: traditional devices will direct traffic along the shortest path, while SDN devices will route the traffic based on the optimal splitting ratio. The primary goal is to minimize the total MLU after the entire migration process (1-8); as we aim to minimize the MLU of the network, we formulate the objective function $\phi$ as follows:

$$\phi = \text{minimize} \sum_{s=1}^{N} U_{\{v_1, v_2, \ldots\ldots v_s\}} \tag{1}$$

$$\sum_{e \in \text{In}(c)} e(\omega) + TM_{ct} = \sum_{e \in Out(c)} x_e(\omega) c \in CN, t \in V \tag{2}$$

$$\sum_{q \in V} f^q_{e,D_i}(\omega) \leq U_{\{v_1, v_2, \ldots\ldots v_s\}} C(e), \forall e \in E, \forall D_i \in D \tag{3}$$

$$\sum_{j=1}^{n} r_j = 1, 0 \leq r_j \leq 1 \tag{4}$$

$$\omega(e) \in \text{N}, \forall e \in E \tag{5}$$

$$x_e(\omega) \geq 0 e \in A \tag{6}$$

$$0 \leq U^{\leq}_{\{v_1, v_2, \ldots\ldots v_s\}} 1, \forall D_i \in D \tag{7}$$

$$f^q_{e,D_i}(\omega) \geq 0, \forall e \in E, \forall t \in V, \forall D_i \in D \tag{8}$$

Looking for an optimal migration sequence from a pure traditional network and considering TE in a hybrid SDN, we determine the best migration sequence with the migration of each SDN node. Equation (1) represents optimization's primary goal. It aims to minimize the sum of MLU across the network, which for each link is computed as the maximum ratio of the link load divided by that link's capacity.

Equation (2) embodies the principle of flow conservation, ensuring that the device's flow and generated flows of the devices are equal to the flow out of the device. Simultaneously, the equation suggests that the traffic demand should be met equally. The capacity restriction is represented by equation (3), which establishes a constraint on link capacity, which means the total of flows on an edge device must not exceed its capacity. Equation (4) stipulates that the sum of the weight coefficients, ri, should equal 1. Equations (5-6) represent that the weight of the link is an integer, and the flow on the link should be positive. Equations (7-8) accordingly offer the permissible value range of the link weight, MLU, and traffic split ratio.

### 3.3 Problem formulation

Let's consider migration to a hybrid SDN, the coexistence of the traditional network with the SDN network devices, ECMP, OSPF rules, and the constraint splitting ratio of SDN traffic; traffic engineering in the hybrid SDN can be expressed as a MINLP (Mixed Integer Non-Linear Programming) problem.

Equation (3) represents the constraint on link capacity, which means the total flow on a given link should not exceed its capacity. We aim to migrate to a hybrid SDN and minimize the network's MLU, so our aim function can easily be modified to maximize network utilization, reduce the total cost, or maximize the network's throughput.

### 3.4 Problem Complexity

Looking for an optimal migration sequence and then minimizing MLU in the migrated hybrid SDN is an NP-complete problem. The MLU in a migrated hybrid SDN is dependent on the deployment ratios and the nodes that were migrated before, which is an NP-complete problem; if the splitting ratio and the link weight of the SDN nodes are given, then the network MLU can be computed in polynomial time and can be compared to the threshold U, whether it is less or more than threshold U. Still when the weight settings are in integer or unknown values, the problem becomes much more complex, so looking for migration sequence and minimizing the MLU is shown to be more difficult than an NP-complete problem. As a result, the complexity of our problem is an NP-complete problem.

### 3.5 SDN migration sequence

Finding the optimal migration sequence is crucial when migrating from a conventional network to an SDN. The location of SDN nodes can impact the available routes and controllable data flows. The migration sequence's primary goal is to optimize controllable traffic, which the SDN controller controls. The traffic volume in a real network fluctuates with time, potentially leading to varying update priorities at various times. In actual deployment situations, once a device is selected to be updated, it doesn't revert to its original configuration. In general, whereas short-term observations may show surprising and unexpected increases in traffic, the traffic flows tend to be relatively stable in long-term observations [31]. Based on this realization, we aim to investigate a traffic pattern that may reflect the features of stable traffic; based

on this traffic pattern, we may solve the SDN migration sequence problem. The traffic control strategy can be used to control unexpected traffic spikes. As a result, the deployment approach can deal with most flow distributions, while the traffic control approach effectively manages unanticipated traffic changes.

First, we will presume that the degree of precision in traffic measurement is constant and the total number of periods donated by T. We calculate the traffic proportion for each flow during each time interval according to the traffic measurements. During each time interval, the specified flow traffic proportion can be calculated by calculating the ratio of the flow traffic demand to the total network traffic for a specific period. Computed results assist in obtaining the traffic distribution of the flows. A polynomial function $F_{sd}(x)$ representing flow (s,d) in the period x can be calculated using the polynomial curve fitting technique, and the definition is specified as follows:

$$F_{sd}(x) = w_0 + w_1 x + \cdots + w_n x^n = \sum_{k=0}^{n} w_k x^k \tag{9}$$

$$I = \sum_{i=0}^{m} (F_{sd}(x_i) - y_i)^2 = \sum_{i=0}^{m} \left( \sum_{k=0}^{n} w_k x_i^k - y_i \right)^2 = min \tag{10}$$

The variable $n$ in the equation denotes the degree of the polynomial, $w$ denotes the polynomial coefficient, and $y_i$ signifies the relative amount of traffic flow in the period $x$. The following equations can be used to get the answer to equation (10):

$$\frac{\partial I}{\partial w_j} = 2 \sum_{i=0}^{m} \left( \sum_{k=0}^{n} \left( \sum_{k=0}^{n} w_k x_i^k - y_i \right) \right) x_i^j = 0,$$

$$j = 0, 1, \ldots, n \tag{11}$$

$$\sum_{k=0}^{n} \left( \sum_{i=0}^{m} x_i^{j+k} \right) w_k = \sum_{i=0}^{m} y_i, j = 0, 1, \ldots, n \tag{12}$$

$$\begin{bmatrix} m+1 & \sum_{i=0}^{m} x_i & \cdots & \sum_{i=0}^{m} x_i^n \\ \sum_{i=0}^{m} x_i & \sum_{i=0}^{m} x_i^2 & \cdots & \sum_{i=0}^{m} x_i^{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=0}^{m} x_i^n & \sum_{i=0}^{n+1} & \cdots & \sum_{i=0}^{m} x_i^{2n} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \cdots \\ w_n \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{m} y_i \\ \sum_{i=0}^{m} x_i y_i \\ \cdots \\ \sum_{i=0}^{m} x_i^n y_i \end{bmatrix} \tag{13}$$

Using the above equations, we can determine the coefficients $w$ for the function Fsd(x). As a result, the traffic weight $L^t{}_{sd} = F_{sd}(x)$ illustrates the flow traffic distribution during period $t$.

Therefore, the migration problem's primary goal is to acquire the migrated SDN updated list to optimize the efficiency of controlled traffic weights. The formulation of the problem is articulated as follows:

$$\text{Max} \sum_{t \in T} \emptyset^t \exists \begin{cases} \sum_{i=1}^{n} H(v_i) = k + |V_s| \\ 0 \langle k \leq | V_l |, v_i \in V, H(v_i) = \{0, 1\} \\ F(P_{sd}) = \begin{cases} 0, & \text{if } \sum_{v \in P_{sd}} H(v) = 0 \\ 1, & \text{if } \sum_{v \in P_{sd}} H(v) > 0 \end{cases} \\ \emptyset^t = \sum_{s,d \in V} \sum_{v \in P_{sd}} F(P_{sd}) \times L_{sd}^t \\ s \neq d \end{cases} \tag{14}$$

Equation 14 represents the problem of maximizing the number of controllable traffic weights such that the initial inequality signifies the K number of migrated traditional network devices, $H(v_i)$ indicating whether, after migration $v_i$, it becomes an SDN node or not. If $v_i$ is an SDN node, then $H(v_i) = 1$, otherwise $H(v_i) = 0$. For all $H(v_i) \in v_s H(v_i)$ is always equal to 1. The second equation assesses whether, after the migration process, the path $P_{sd}$ includes SDN nodes or not, and the last equation counts the total number of controlled traffic weights during a specific period $t$ after the migration of several nodes into SDN.

### 3.6 Routing Optimization

After migrating conventional network devices to an SDN and creating a hybrid SDN environment, routing flows in the hybrid SDN must be optimized. We introduce a heuristic approach for optimizing flow routing in a hybrid SDN. As in the hybrid SDN, conventional network devices coexist with the SDN network devices, and the traditional network devices use OSPF, while the SDN network devices use flow routing; in this case, there is a need for the optimization of both OSPF weight settings and the splitting ratio of the SDN network devices. For conventional network devices, the flows are restricted using the ECMP (Equal cost multi-path) rule, which routes the packet based on the shortest paths. To achieve flow balancing, we optimize the OSPF weight setting. As the traffic is splitting along their outgoing links for the SDN nodes, we optimize the splitting ratio of the SDN nodes to achieve load balancing. So, for optimizing both in offline mode, we have introduced a heuristic approach, H-STE, using different algorithms for routing optimization in a hybrid SDN environment. In the H-STE approach, to

achieve minimum MLU, initially, we have utilized the distributed Floyd-Warshal algorithm to expedite the process of determining the most optimal paths between every pair of source and destination nodes. As with the previous methods, scalability issues were the main concern. We considered the scalability issues for large and complex networks by using the distributed approach, which ensures scalability for large and complex networks. For the splitting ratio of SDN nodes, we utilized the Bellman-Ford shortest path algorithm to determine the shortest path tree for selecting the optimized nodes. The splitting ratio plays a crucial role in the iterative process of the H-STE method. It is used to modify the traffic splitting ratios at the SDN nodes at each iteration, ensuring effective load balancing and traffic distribution over multiple paths. The splitting_ratio algorithm is designed to consider multiple objectives and uses the adaptive approach to fine-tune the traffic flows. This approach ensures that traffic is routed appropriately to minimize MLU and meet service-level requirements for critical applications

## 4. ALGORITHMS

In this part, we provide the algorithms employed in this experiment. First, we introduced algorithm 1, the SDN migration sequence algorithm, to find the best migration sequence of SDN nodes to obtain controllable traffic flows. Next, we introduced a local search heuristic algorithm for optimizing flow routing to minimize MLU in a hybrid SDN. We present the algorithms in detail, explore the approximation ratio, and, at the end, analyze the complexity of the algorithms.

### 4.1 Algorithm 1 for the migration sequence of SDN nodes

**Algorithm 1**: SDN migration sequence

**Input:** $G = (V, E), T, k$, *traffic weight*
**Output:** $X$
$H(v1) = 0, \forall v1 \in L, H(v2) = 1, \forall v2 \in C, X \leftarrow \emptyset$
**for** iteration $i = 1$: $k$ do
  $f_m \leftarrow 0$
  **for** every node $\alpha \in L$**do**
    $H(\alpha) \leftarrow 1, T \leftarrow$ every controllable flow, $f_a \leftarrow \sum_{t\in T}\sum_{\tau\in T} L^t \cdot \tau$
    **if** $f_a > f_m$*then*
      $x_i \leftarrow a, f_m \leftarrow f_a$
    **end if**
  **end for**
$\boldsymbol{X} \leftarrow \boldsymbol{X} \cup \{\boldsymbol{x_i}\}$
$\boldsymbol{L} \leftarrow \boldsymbol{L} - \{\boldsymbol{x_i}\}, \boldsymbol{C} \leftarrow \boldsymbol{C} \cap \{\boldsymbol{x_i}\}$
**end for**

Algorithm 1 describes the SDN migration sequence problem; it might take time to determine the ideal migration order. The precise deployment and placement of SDN nodes can influence the data flow control routes. The migration sequence's primary goal is to maximize programmable or controllable traffic under the SDN controller's direction. Searching for an optimized migration sequence is computationally demanding because the solution space is N! We can solve these problems using linear optimizer programs such as CPLEX [32] and Gourbi [33]. However, these procedures may consume too much time, so we introduced a simple greedy algorithm for the migration problem, as shown in Algorithm 1, that selects the node from the legacy network nod set to maximize the amount of controllable traffic in SDN. The time complexity for this greedy solution for the migration of the SDN device is $O(N)$ where $N = |V|$ which is reasonable and acceptable. This method uses the traffic weight data processed through the SDN migration sequence section 3.5, as mentioned and presented above, as an input for this algorithm. It scans each node in the legacy node set $L$ at each iteration, selects a node that maximizes controlled traffic, and then adds it to SDN node set $C$. In the end, an updated list is created after $k$ iterations. After completing the $k$ iteration, the nodes with the maximum controllable traffic will be selected for the migration to the SDN node set in set X. The selection process of nodes continues until an updated list of nodes has been properly chosen for their ability to manage and direct the traffic efficiently.

### 4.2 Algorithm 2: H-STE

After migrating to a hybrid SDN environment, we introduced the heuristic approach H-STE, described in Algorithm 2, for the routing optimization in the hybrid SDN environment. The algorithm's main objective is to optimize traffic engineering in a hybrid SDN environment; it optimizes the splitting ratio of SDN nodes and the OSPF weight setting in a hybrid SDN. The H-STE method helps bridge the gap between both networks by optimizing the whole network performance while adhering to the limitations imposed by traditional routing protocols. This method adjusts the splitting ratio of SDN nodes to reduce network congestion while having a multi-object focus, mainly minimizes MLU, and considers other network performance metrics. As it optimizes both, the optimization process is divided into two phases: the optimization of the OSPF setting and the optimization of the Splitting ratio.

The algorithm is described below: this local search heuristic approach is designed to optimize the splitting ratio of SDN nodes and the OSPF weight setting to minimize MLU in a hybrid SDN environment. To achieve

**Algorithm 2**: H-STE

**Input:** $G = (V, E)$, $V_s$, $C(e)$, $D$,initial_step_size
**Output** : $U$
1 Initial weight setting matrix $\omega$;
2 $U_{current}$ = distributed_floyd_warshall (G, $\omega$, D);
3 $U_{best} = U_{current}$;
4 $\omega_{best} = \omega$;
5 Set Adaptive_step_size = initial_step_size;
6 while iteration_times $\leq$ max_iter_time:do
7 *Ucurrent* = *splitting_ratio*($\omega_{best}$, *Vs*, *D*, *C*(*e*), *G*, daptive_step_size)
8 If Ucurrent $<$ $U_{best}$ then
9 Update $U_{best}$= Ucurrent;
10 Update $\omega_{best} = \omega$;
11 $U = U_{best}$;

minimum MLU, we initially utilized the distributed-Floyd-Warshal algorithm to find the most efficient routes between each combination of source and destination nodes faster. The distributed approach ensures scalability for large and complex networks. Following that, we distribute traffic demands along these optimal shortest paths (lines 1-4). We go through a series of iterations to find the best routing configuration (lines 5-11). The current MLU, denoted as Ucurrent, is calculated through the splitting ratio function; using the established OSPF weight configuration, this method computes the most efficient traffic distribution ratio for the specified set of $k$ SDN devices (line 8). After identifying the current MLU, we revise the optimal OSPF weight configuration and MLU by evaluating $U_{current}$ against the current best MLU, $U_{best}$ (lines 9-11). The iterations end upon reaching a predetermined maximum iteration limit. In this algorithm, we have used the distributed-Floyd-warshall algorithm to find the shortest possible path between each combination of source and destination node faster, and the time complexity of this algorithm is $O(n^3)$, as n is the number of nodes in the topology.

### 4.3 Splitting ratio optimizer

In Algorithm 3, we demonstrate the information for optimizing the splitting ratio of SDN devices.

**Algorithm 3**: Splitting ratio with Bellman-Ford

**Input:** optimal OSPF weight setting $\omega_{curren}$, $V_s$, *D*, *C(e)*, *V*, *E*
**Output:** *U*
1 **for each** $v \in V$**:**
2 DAG = Bellman_ford_Shortest_Path ($\omega_{current}$, *v*, *E*);
3 for each $i \in V_s$, $e_{ij}$ in outgoing_links of *i* **do**
4 Add link eij to ADG where i $\neq$ j and i, j are incomparable;
5 if check_loop (ADG) = = 1 ADG(*i*,*j*) = 0;
6 end if
7 end for
8 foreach *v* in Topological_sort(*V*) **do**
9 Route_flow_adaptively (*G*, *v*, *D*);
11 expr = Multi_commodity (E, V);
12 U = cplex_solver(expr);
13 Return U;

We present a thorough explanation of the splitting ratio in algorithm 3. The main objective of this algorithm is to optimize the traffic splitting ratio of SDN nodes when OSPF weight is given, and the time complexity of the splitting ratio function is $O(n^2)$, as n is the number of nodes in the topology. Traffic demand in a hybrid SDN can be acquired through network monitoring approaches such as sFlow [34].

The splitting ratio plays a crucial role in the iterative process of the H-STE method. It is used to modify the traffic splitting ratios at the SDN nodes at each iteration, ensuring effective load balancing and traffic distribution over multiple paths. The splitting_ratio algorithm is designed to consider multiple objectives and uses the adaptive approach to fine-tune the traffic flows. This approach ensures that traffic is routed to minimize MLU and meet service-level requirements for critical applications. Initially, we form a Directed Acyclic Graph (DAG) by considering each node as a destination node (line 2). DAG in this algorithm refers to the matrix containing all the feasible routes from a node to a particular node $v$. DAG $(i, j) = = 0$ when in a graph, there is no link between the nodes, $i$ and $j$, DAG$(i,j) = = 1$ when a link exists between the nodes $i$ and $j$, in the graph.

The DAG is refined based on heuristic criteria (lines 3-4). For instance, when constructing the DAG, outbound connections with high capacity or low utilization are given priority. The algorithm prevents circular dependencies while ensuring that the nodes in the graph are incomparable. In creating a DAG, we begin by picking a node and identifying all the shortest paths to the node using the Bellman-Ford Shortest Path algorithm, ensuring no loops in the shortest routes tree. We add all the other edges representing the SDN node outbound links into the DAG (line 4). Once the link is added, we will look for the loop; if there is a loop, we will remove that link to ensure no loop exists (line 5). After the construction of DAG, we can easily direct the traffic needs in the route flow function. We select a node based on the topological sorting sequence, such that all the flows traveling through the current node are computed before being allocated to the next hop; the flows are then routed to all outbound links in the DAG (lines 8-9). The splitting flows are applied to the subsequent hops if the specified node is an SDN node, and the exact destination flows are randomly pushed into the subsequent hops. To split the flow, we may utilize the hash functions to map the flows to distinct next hops; if the selected node is a legacy node, the flows are uniformly distributed throughout the shortest pathways and directed toward the following hops. This guarantees that all the flows from the selected node to the

node *v* are correctly routed (line 9). We repeat the process this way and route the flow of each node in a topological sort order to obtain the linear constraints. And finally, we employ the linear programming solver CPLEX [32] to address the optimization problem (line 12). Solver Cplex uses the dual simplex approach to solve linear programming problems, and its time complexity is in polynomial time. Considering the above algorithms, the overall time complexity of our algorithm is $O(n^3)$.

**Analysis of the algorithm**

The splitting ratio with the Bellman-Ford algorithm provides an adaptive and scalable approach for traffic engineering in the hybrid SDN environment. Two main factors shape the algorithm's performance and make it the finest method for the hybrid SDN: the first one is using the Bellman-Ford algorithm to compute the shortest paths, which ensures efficiency and can handle large-scale and complex networks. The Bellman-Ford algorithm performs excellently in network scenarios with dynamically changing traffic demands, and it is the perfect algorithm for handling dynamic changes in link weights. The second one is the heuristic-based DAG construction, which helps the algorithm focus on minimum link utilization and high network capacity, ensuring the traffic is routed via the most efficient pathways, minimizing network congestion, and improving overall network performance. Finally, using an optimization solver such as Cplex ensures that the algorithm finds a globally optimized solution for the network.

## 5. EXPERIMENTS AND ANALYSIS

In this part, we demonstrate the performance of our proposed method on different real network datasets and topologies used in our experimental analysis; we execute the algorithms mentioned in the Algorithm Section and analyze their performance on different real network topologies. First, we will examine the environment setting, simulation topologies, and traffic datasets. Following that, we present visual representations of the migration sequence algorithm FCM method with the controllable traffic weight; after that, we present the effectiveness of the proposed method H-STE on minimizing MLU under different deployment ratios of SDN nodes, and at the end, we show the average time taken by our method for various network topologies.

### 5.1 Environment Setting

All the simulation experiments are simulated on a high-performance personal computer with a 3.20 GHz Intel Core i7 with 16 GB of RAM, and the Mininet Simulator simulates the network environments [35]. First, we created a traditional network topology, then incrementally deployed SDN nodes using the Mininet tool; for configuring legacy routers, we used FRR (Free Range Routing) using VTY Shell to configure network protocols. The simulation experiments are coded with the C++ program to model the traffic routing procedure and to determine the splitting ratio of SDN nodes; we employ the CPLEX program in our C++ program. To connect different networks in the conventional network and to integrate the legacy network with SDN AS (Autonomous Systems), we have configured BGP (Border Gateway Protocol) inside all the legacy routers, as it is necessary to have a peering system that makes the BGP protocol work. and the integration between the legacy network and the SDN network is carried out through the SDN-IP. This program operates on top of the ONOS controller and allows SDN nodes to communicate with legacy routers using BGP protocols.

Selecting the controller in the hybrid SDN is another issue that needs to be considered; in this simulation experiment, the ONOS [36] controller is chosen to be the SDN controller, and the integration between the legacy network and the SDN network is carried out through the SDN-IP. This program operates on top of the ONOS controller. The SDN-IP application allows SDN nodes to communicate with legacy routers using BGP protocols.

To demonstrate the capability of our proposed method for optimizing routing and minimizing MLU with minimum SDN deployment ratios while considering load balance, we have evaluated our experiment on different real network topologies (Abilene, Cernet, Nobel-Germany, and Geant) and compared it to the following methodologies. Compared to alternative routing methods, we demonstrate that H-STE performs better at minimizing MLU with minimum SDN deployment ratios.

OSPF [13]: The OSPF technique is a routing protocol primarily used in conventional legacy networks; it supports routing the flows based on the shortest path from the source to the destination to get load balanced. The shortest pathways are determined based on the link weights. To compare it with the heuristic algorithm H-STE, we set all the parameters the same as [13]; if there are multiple shortest paths, the flows are eventually split into the shortest paths using ECMP.

OSRO [16]: To thoroughly investigate and contrast the efficiency of SDN nodes optimization, we compare it with the optimization of only splitting ratio of SDN nodes for

**Table 2: Network topologies information**

| Network | Nodes | Links | Time | Time-range |
|---|---|---|---|---|
| Abilene | 12 | 30 | 5 min | Six months |
| Cernet | 14 | 32 | 5 min | One month |
| Nobel-Germany | 17 | 26 | 5 min | One day |
| Geant | 23 | 74 | 15 min | Four months |

TE performance; we compare and conduct experiments with static OSPF and optimization of only the splitting ratio of SDN nodes.

H-STE: We conducted experiments on different real network topologies to optimize the splitting ratio of SDN nodes and OSPF weight setting. All of the initial weight settings of the network are set to 1. The CDF curves of the MLU are shown in Figure 5. As shown in Figure 5, compared to OSPF and OSRO under different real network topologies, our method achieves minimum MLU; with a deployment ratio of 30% SDN nodes, we can get close to the optimal result and get minimum MLU in the network.

### 5.2 Dataset and Topologies

The dataset and network topologies used in this experiment are shown in Table 2; the network topologies datasets are available on SDNlib [37]. The Abilene topology, part of the American Research and Education Network, comprises 12 nodes and 30 links. The Cernet topology, belonging to the China education and research network, consists of 14 nodes and 32 links. The Noble-Germany network topology consists of 17 nodes and 26 links. Lastly, the GEANT topology, part of the European research and education network, consists of 23 nodes and 74 links. The traffic datasets required in the experiment analysis are real network datasets measured or created in the specific topologies every five or fifteen minutes.

For the Abilene topology, the total traffic provided by the TOTEM project [38] was measured from the first of March 2004 to September 2004, and every 24 hours, snapshots were created at 5-minute intervals. For the CERNET topology, the datasets are provided by Zhang et al. [39]. They were measured every five minutes from February 19, 2013, to March 26, 2013. For the GEANT topology datasets provided by Uhling [40], the traffic matrices were measured every 15 minutes from 04-May-2005 to 31-Aug-2005.

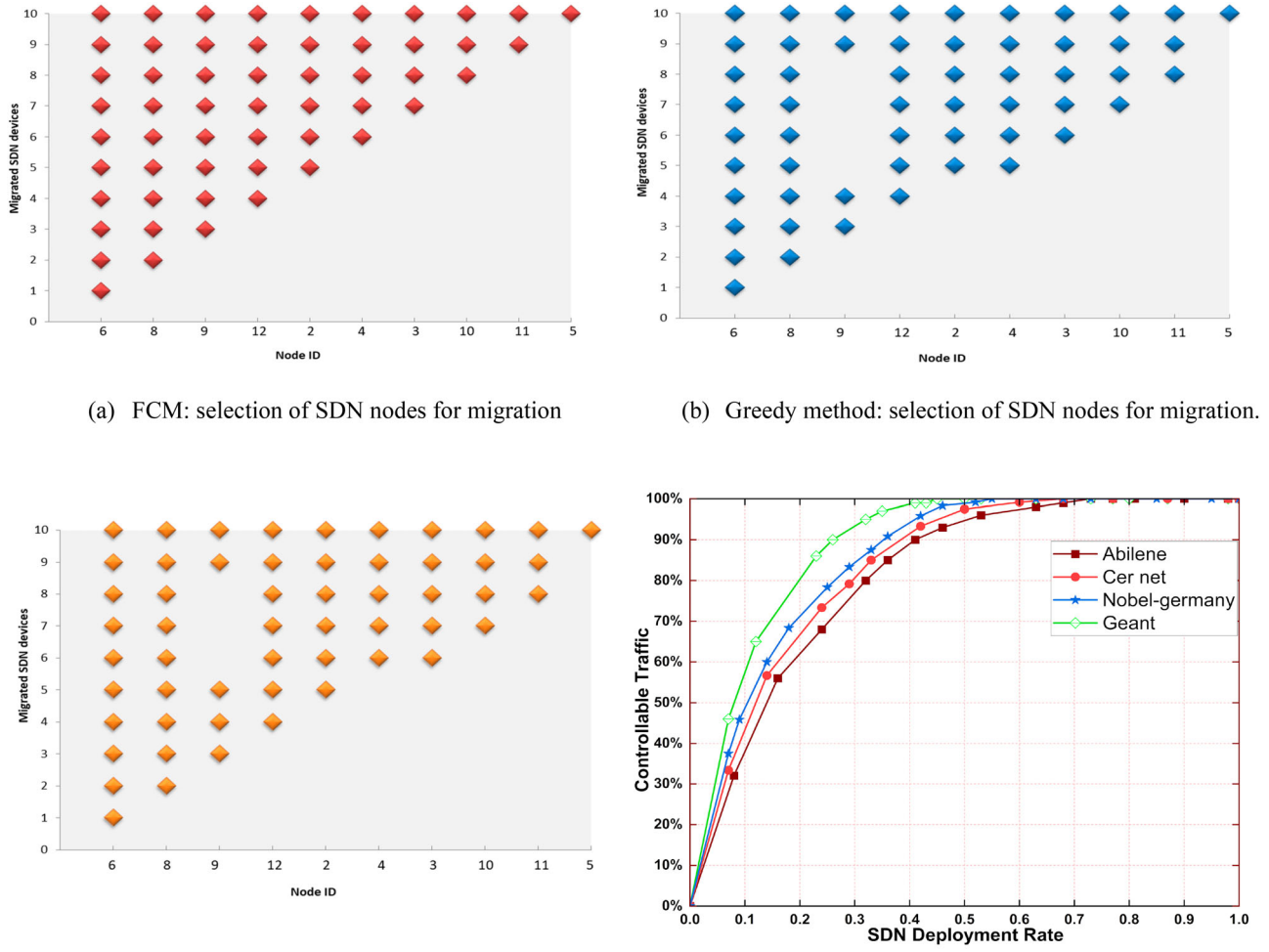
### 5.3 Migration sequence evaluation

Much work evaluates the migration sequence and incremental deployment in hybrid SDN with different strategies and terms; in terms of traffic control, we compare the suggested migration sequence solution, known as FCM (Flow Control Method), to other modern methods that focus on controllable traffic in SDN upgrading strategies presented in Figure 3. This method intends to incrementally upgrade SDN nodes to maximize the weight of controllable traffic.

The Modified-Greedy strategy [41] prioritizes nodes with the highest traffic while factoring in the migration cost. On the other hand, the local search strategy [41] focuses on maximizing the number of controllable paths. Deg [19] prioritizes upgrading nodes with the highest degree, whereas Vol [42] focuses on upgrading nodes with the most traffic passing through them.

Different experiments are carried out and evaluated within different real network topologies, as presented in Table 2. When considering previous deployment solutions, the findings indicate that the FCM method achieves approximately equal performance to modified-greedy and local-search, which do not rely on prior deployment solutions. Furthermore, the experiment shows that this method outperforms Deg and Vol in different topologies. Figure 3(a,b,c) demonstrates the node selection process in each round of migration of SDN nodes using FCM, Modified-Greedy, and local search methods. The horizontal axis denotes the device ID, the vertical axis indicates the number of selected SDN nodes for the migration, and the diamond symbols indicate those chosen nodes.

In comparison to the prior strategy for SDN nodes deployment ratio, the FCM method considers previous deployment ratios; for example, as shown in Figure 3(a) whenever the total number of SDN devices to be migrated is 2, the node of 6 and node of 8 will be selected, whenever the total number of SDN devices to be migrated is three nodes of 6, 8 and node 9 will be chosen, same whenever the total number of SDN devices to be migrated is six then node 6, 8,9,12,2 and node 4 will be selected for the migration as the previous deployment ratios will be considered. In contrast, if we consider the Greedy method for the selection of SDN devices, as shown in Figure 3(b), when the number of the migrated devices is Five, then nodes 6,8,12,2, and node 4 will be selected; the same if we consider Local search method for the migration of SDN devices, as shown in Figure 3(c), when the number of migrated node is six then nodes of 6,8,12,2,4 and node 3 will be selected for the migration as in these two methods, the previous deployment ratio of SDN devices will not be considered. Figure 3(d) shows the controlled traffic weight with the SDN nodes




(a) FCM: selection of SDN nodes for migration

(b) Greedy method: selection of SDN nodes for migration.

(c) Local search: selection of SDN nodes for migration.

(d) SDN deployment with controllable traffic weight

**Figure 3:** Evaluation of SDN Node Migration: (A) FCM: selection of SDN nodes for migration (B) Greedy method: selection of SDN nodes for migration (C) Local search: selection of SDN nodes for migration (D) SDN deployment with controllable traffic weight

deployment ratio under different network topologies outlined in Table 2. There is a direct link between increasing the number of SDN nodes and the increase in controllable traffic; with more SDN nodes, we experience significant improvements in traffic management and overall network performance. As we can see, the findings show that with a proper 30% SDN deployment ratio, we can control over 70% of the network traffic and get the maximum benefit in managing controllable traffic, which improves the overall network performance from a traffic engineering perspective. This empirical data highlights the importance of deploying SDN nodes in shaping and optimizing network traffic management strategies.

In the FCM method, we use a simple greedy algorithm that selects the node with the maximum amount of controllable traffic to SDN and aims to minimize MLU while considering network load balance in the hybrid SDN environment. Compared to other alternative approaches, our method performs well in reducing MLU and considering load balance in the hybrid SDN environment. The FCM method, which considers previous deployment ratios, outperforms the VOL and DEG methods and achieves comparable performance to the modified and local search methods, which do not rely on previous deployment ratios.

## 5.4 Evaluation of Network performance in terms of MLU

Our main objective is to minimize the MLU with the minimum number of SDN nodes while maintaining the network's load balance. To achieve minimum MLU, we increase the deployment ratios, and we can see the improvement of MLU when we deploy 0.3; the MLU of different network topologies begins to stabilize.

In a hybrid SDN network, the SDN deployment ratio is defined as the ratio of SDN nodes throughout the network. Minimizing the network's MLU is an essential

parameter in the hybrid SDN when considering network load balance. Maintaining network load balance in a hybrid SDN refers to distributing the traffic equally to all the network paths to ensure efficient usage of network resources and minimize congestion on any single link. This balance helps to prevent the situation in which some links exceed their capacity (overloaded), leading to congestion and inability to handle the traffic efficiently, while some other links are not utilized at any capacity or are underutilized. By balancing the load, the network can provide consistent performance, minimize latency, and improve reliability.

The MLU of each link can be calculated as the maximum ratio of the link load divided by the link capacity. Our main objective is to minimize the MLU with the minimum number of SDN nodes while maintaining the network's load balance to improve network performance. The value of MLU varies and directly depends on the increase in SDN deployment ratios; we plot the curves to show that the value of MLU becomes smaller and decreases with increases in the SDN node deployment ratio under different network topologies. Figure 4 depicts the effect of the SDN nodes deployment ratio on improving the MLU under different network topologies. The result shows that the improvement of MLU directly depends on the increase of SDN deployment ratios; we can see a rapid decrease in the value of MLU with a shift from 0 to 0.3 in the SDN deployment ratios. As soon as the deployment ratio surpasses 0.3, the MLU begins to stabilize, and the curve of H-STE becomes flat. Minimizing the MLU and optimizing routing in a hybrid SDN environment is challenging, to minimize the MLU, we increase the deployment ratios, and we can see that when we deploy 0.3, the MLU of different topologies begins to stabilize. The findings indicate that increasing the deployment ratios of SDN can obtain near-optimal outcomes in optimizing MLU comprehensively.

Increasing the number of SDN nodes can boost the ability to split and direct the traffic flows in a hybrid SDN, as the main goal is to reduce the MLU to prevent network congestion and evenly distribute the network load to enhance the overall network performance. Different methods and metrics can evaluate the load balance in the hybrid SDN environment, such as load variance, link utilization distribution, MLU, failover capacity, and path diversity. The most crucial metric for evaluating load balance in the hybrid SDN is MLU; in this study, we evaluate the network load balance by considering the MLU. It represents the utilization percentage of the link with the most capacity or overloaded, and with zero capacity or underutilized links in the network. A high MLU indicates that particular links are reaching their maximum capacity, which causes network congestion and an inability to handle the traffic efficiently. Minimizing MLU is critical for load balancing because it helps prevent overload on any single link; with minimum MLU, we can distribute the traffic load more efficiently across the network. In Figure 4, we can see that the OSPF curve remains steady and flat with decreasing or increasing SDN deployment ratios; the reason is that OSPF routes traffic based on the shortest path from the source to the destination.

We consider and set the SDN deployment ratios to 30% based on the experimental analysis of different network topologies. As the MLU starts to stabilize, we can see that the curves are getting near-optimal, and we can get near-optimal results. To show the priority and the efficiency of our proposed strategy for routing optimization and enhancing TE performance, we compared our suggested H-STE strategy with different strategies for minimizing MLU under different real network topologies, as shown in Table 2, to compare it with only optimization of OSPF weight setting [13], we set all the parameters the same as [13]. We compared it with only splitting ratio of SDN nodes with fixed link weight setting [16], performed many experiments with different traffic datasets, and drew the average MLU result in CDF graphs under different topologies in Figure 5. As shown in Figure 5, when the deployment ratio increases, the value of MLU becomes smaller. Compared to other methods, the H-STE method is close to the optimal result, as this method optimizes both the SDN node splitting ratio and OSPF weight setting. In contrast, the OSPF method only optimizes the OSPF weight setting, and the OSRO method optimizes only the splitting ratio of SDN nodes with a fixed link weight.

In hybrid SDN, reducing the MLU and optimizing routing present significant challenges. To minimize the MLU, we increase the deployment ratios, and we can see that with a 30% deployment ratio of SDN, the MLU of different topologies begins to stabilize. The findings indicate that increasing the deployment ratios of SDN can obtain near-optimal outcomes in optimizing MLU comprehensively. More SDN nodes can bring more flexibility to split and forward the traffic flows in a hybrid SDN, as our common goal is to minimize the MLU to avoid network congestion and to balance the network load to enhance the overall network performance. As a result, compared to other optimization methods, the H-STE method obtains a much lower MLU under different network topologies.

This finding is insightful and a baseline for balancing the advantage of SDN control with resource costs in a

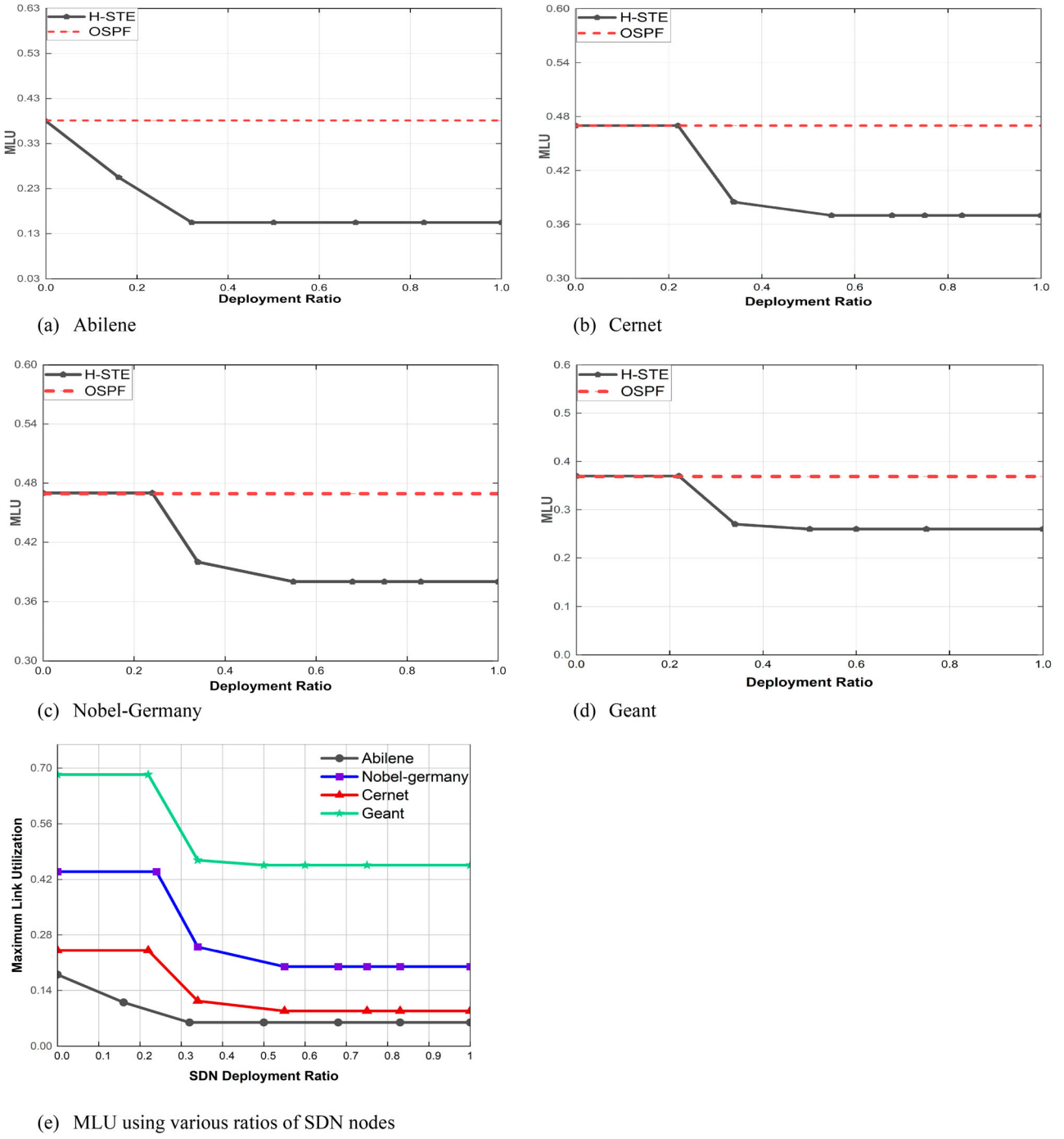



**Figure 4:** MLU with various SDN deployment ratios in real network topologies.: (A) Abilene (B)Cernet (C) Nobel-Germany (D) Geant (E) MLU using various ratios of SDN nodes

typical hybrid SDN network; it suggests a balance point where minimum SDN control is applied to control traffic without heavily depending on SDN resources. However, the generalizability of this result can vary under different network configurations and conditions. Adjusting the SDN deployment ratio may improve performance for more specialized or extreme scenarios, such as networks with high dynamic traffic, strict resiliency requirements, or limited path redundancy. This ratio can be a starting point, but may need to be adjusted to suit specific network characteristics and operational demands.

## 5.5 CPU or Execution time

We conducted the simulation experiments on a personal computer with a 3.20 GHz Intel Core i7 and 16 GB of RAM. To analyze the process's execution times, we repeated the experiments multiple times and measured

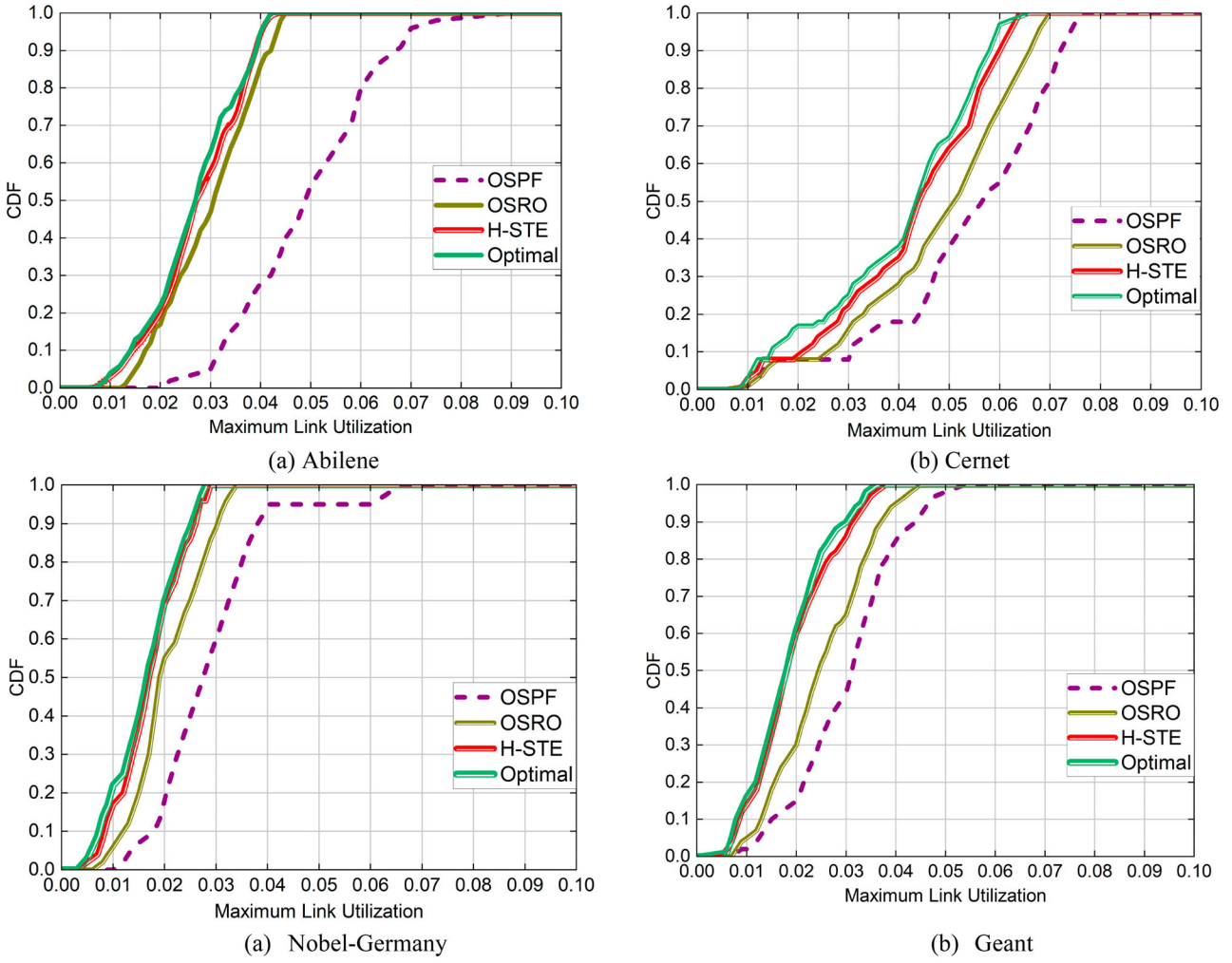



**Figure 5:** MLU CDF curve under real network topologies: (a) Abilene (b) Cernet (C) Nobel-Germany (D) Geant

**Table 3: Execution time**

| Network topologies | Nodes | links | H-STE time |
|---|---|---|---|
| Abilene | 12 | 30 | 8.1s |
| Carnet | 14 | 32 | 8.5s |
| Noble-Germany | 17 | 26 | 10.2s |
| GEANT | 22 | 74 | 13.8s |

the method's execution times on different real network topologies. We calculated and took the average CPU times for various topologies and listed the times in Table 3.

As shown in Table 3, our method executes quickly with small-scale network topologies. We implemented our method on small-scale networks with fewer than 50 nodes, and it executed very quickly. In the future, we will implement our method on large-scale networks, synthetic topologies with more connectivity and complex systems, and a testbed network to examine the performance and check the execution time of our method in different scenarios.

## 6. CONCLUSION

In recent years, academia and industry have demonstrated significant interest in improving and optimizing TE; however, due to network complexity, research on traffic engineering optimization remains an active area of investigation. This study adds to the growing body of knowledge on TE and route optimization in hybrid SDN. we studied the migration sequence from a conventional network to a hybrid SDN to maximize controllable traffic weight. We provide a heuristic H-STE technique to improve traffic engineering in the hybrid SDN. This technique focuses on optimizing the OSPF weight setting and the splitting ratio of traffic across SDN nodes. We have conducted several experiments on real network datasets, and we find that with a proper 30% SDN node deployment, we can control more than 70% of the network traffic; this leads to a decrease and stabilization of the network MLU, resulting in an almost ideal network performance. In the future, our plans involve implementing our approach on large-scale networks and

a real testbed network environment; this will allow us to assess the execution time, dependability, and scalability of our method in different scenarios. The future of HSDN systems holds enormous potential, especially when considering long-term sustainability, adaptability, and integration with fast-changing technologies such as IoT and 5 G. Future research could concentrate on integrating hybrid SDN with edge computing, in which hybrid SDN solutions route data intelligently to the local edge nodes for real-time processing while only sending critical data to the data centers or central cloud. As IoT and 5G networks generate massive amounts of data, this technique would improve network latency, minimize congestion, and enhance overall response in IoT and 5G applications.

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

## REFERENCES


1. U. Cisco. Cisco Annual Internet Report (2018-2023) White Paper, 2022.

2. A. Mendiola, et al., "A survey on the contributions of software-defined networking to traffic engineering," *IEEE Commun. Surv. Tutor.*, Vol. 19, no. 2, pp. 918–53, 2016. DOI: 10.1109/COMST.2016.2633579

3. C.-Y. Hong, et al. "Achieving high utilization with software-driven WAN." Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM. 2013.

4. S. Jain, et al., "B4: experience with a globally-deployed software-defined WAN," *ACM SIGCOMM Comput. Commun. Rev.*, Vol. 43, no. 4, pp. 3–14, 2013. DOI: 10.1145/2534169.2486019

5. R. Amin, M. Reisslein, and N. Shah, "Hybrid SDN networks: A survey of existing approaches," *IEEE Commun. Surv. Tutor.*, Vol. 20, no. 4, pp. 3259–306, 2018. DOI: 10.1109/COMST.2018.2837161

6. Y. Sinha, and K. Haribabu, "A survey: hybrid sdn," *J. Netw. Comput. Applic.*, Vol. 100, pp. 35–55, 2017. DOI: 10.1016/j.jnca.2017.10.003

7. B. Fortz, and M. Thorup, "Optimizing OSPF/IS-IS weights in a changing world," *IEEE J. Select. Areas Commun.*, Vol. 20, no. 4, pp. 756–67, 2002. DOI: 10.1109/JSAC.2002.1003042

8. Y. Guo, et al., "Incremental deployment for traffic engineering in hybrid SDN network," in *2015 IEEE 34th international performance computing and communications conference (IPCCC)*, Nanjing, China: IEEE, 2015, pp. 1–8.

9. B. Kar, E. H.-K. Wu, and Y.-D. Lin, "The budgeted maximum coverage problem in partially deployed software defined networks," *IEEE Trans. Netw. Service Manage.*, Vol. 13, no. 3, pp. 394–406, 2016. DOI: 10.1109/TNSM.2016.2598549

10. X. Jia, J. Yong, and G. Zehua, "Incremental switch deployment for hybrid software-defined networks," in *2016 IEEE 41st conference on local computer networks (LCN)*, Dubai, United Arab Emirates: IEEE, 2016, pp. 571–574.

11. Z. Yang, and K. L. Yeung, "SDN candidate selection in hybrid IP/SDN networks for single link failure protection," *IEEE/ACM Trans. Netw.*, Vol. 28, no. 1, pp. 312–21, 2020. DOI: 10.1109/TNET.2019.2959588

12. H. Xu, et al., "Incremental deployment and throughput maximization routing for a hybrid SDN," *IEEE/ACM Trans. Netw.*, Vol. 25, no. 3, pp. 1861–75, 2017. DOI: 10.1109/TNET.2017.2657643

13. B. Fortz, and M. Thorup, "Internet traffic engineering by optimizing OSPF weights," in *Proceedings IEEE INFOCOM 2000 Conference on computer communications. Nineteenth annual joint conference of the IEEE computer and communications societies (Cat. No. 00CH37064)*, Tel Aviv, Israel: IEEE, 2000, Vol. 2, pp. 519–528.

14. M. Ericsson, M. G. C. Resende, and P. M. Pardalos, "A genetic algorithm for the weight setting problem in OSPF routing," *J. Combinator. Optimiz.*, Vol. 6, pp. 299–333, 2002. DOI: 10.1023/A:1014852026591

15. S. Srivastava, et al., "Determining link weight system under various objectives for OSPF networks using a Lagrangian relaxation-based approach," *IEEE Trans. Netw. Service Manage.*, Vol. 2, no. 1, pp. 9–18, 2005. DOI: 10.1109/TNSM.2005.4798297

16. S. Agarwal, M. Kodialam, and T. V. Lakshman, "Traffic engineering in software-defined networks," in *2013 proceedings IEEE INFOCOM*, Turin, Italy: IEEE, 2013, pp. 2211–2219.

17. D. K. Hong, et al. "Incremental deployment of SDN in hybrid enterprise and ISP networks." Proceedings of the Symposium on SDN Research. 2016.

18. M. Caria, et al., "Divide and conquer: partitioning OSPF networks with SDN," in *2015 IFIP/IEEE international symposium on integrated network management (IM)*, Ottawa, ON, Canada: IEEE, 2015, pp. 467–474.

19. J. He, and W. Song, "Achieving near-optimal traffic engineering in hybrid software-defined networks," in *2015 IFIP networking conference (IFIP networking)*, Toulouse, France: IEEE, 2015, pp. 1–9.

20. Y. Guo, et al., "SOTE: traffic engineering in hybrid software-defined networks," *Comput. Netw.*, Vol. 154, pp. 60–72, 2019. DOI: 10.1016/j.comnet.2019.03.008

21. H. Xu, et al., "Scalable software-defined networking through hybrid switching," in*IEEE INFOCOM 2017-IEEE*

*conference on computer communications*, Atlanta, GA, USA: IEEE, 2017, pp. 1–9.

22. Y. Guo, et al., "Traffic engineering in hybrid software defined network via reinforcement learning," *J. Netw. Computer Applic.*, Vol. 189, pp. 103–116, 2021. DOI: 10.1016/j.jnca.2021.103116

23. Z. Xu, et al. "Teal: Learning-accelerated optimization of wan traffic engineering." Proceedings of the ACM SIGCOMM 2023 Conference. 2023.

24. C. Liu, et al., "Scalable deep reinforcement learning-based online routing for multi-type service requirements," *IEEE Trans. Parallel Distrib. Syst.*, Vol. 34, no. 8, pp. 2337–51, 2023. DOI: 10.1109/TPDS.2023.3284651

25. B. Lin, et al., "TITE: A transformer-based deep reinforcement learning approach for traffic engineering in hybrid SDN with dynamic traffic," *Fut. Gener. Computer Syst.*, Vol. 161, pp. 95–105, 2024. DOI: 10.1016/j.future.2024.07.006

26. Y. Guo, et al., "MATE: A multi-agent reinforcement learning approach for traffic engineering in hybrid software defined networks," *J. Netw. Computer Applic.*, Vol. 231, pp. 103981, 2024. DOI: 10.1016/j.jnca.2024.103981

27. Y. Guo, et al., "Traffic engineering in hybrid SDN networks with multiple traffic matrices," *Computer Netw.*, Vol. 126, pp. 187–99, 2017. DOI: 10.1016/j.comnet.2017.07.008

28. Y. Guo, et al., "Routing optimization with path cardinality constraints in a hybrid SDN," *Computer Commun.*, Vol. 165, pp. 112–21, 2021. DOI: 10.1016/j.comcom.2020.11.004

29. M. M. Tajiki, et al., "SDN-based resource allocation in MPLS networks: A hybrid approach," *Concurr. Comput. Pract. Exper.*, Vol. 31, no. 8, pp. e4728, 2019. DOI: 10.1002/cpe.4728

30. N. McKeown, et al., "Openflow: enabling innovation in campus networks," *ACM SIGCOMM Comput. Commun. Rev.*, Vol. 38, no. 2, pp. 69–74, 2008. DOI: 10.1145/1355734.1355746

31. D. Sanvito, et al., "Adaptive robust traffic engineering in software-defined networks," in *2018 IFIP networking conference (IFIP networking) and workshops*, Zurich, Switzerland,: IEEE, 2018, pp. 145–153.

32. I. B. M. I. L. O. G. Cplex, "V12. 1: user's manual for CPLEX," *Intern. Business Mach. Corp.*, Vol. 46, no. 53, pp. 157, 2009.

33. G.optimization, Inc. "Gurobi optimizer reference manual." 2020.

34. S. Panchen, N. McKee, and P. Phaal, "Sflow: A method for monitoring traffic in switched and routed networks," *Internet Eng. Task Force* (2001.

35. B. Lantz, B. Heller, and N. McKeown. "A network in a laptop: rapid prototyping for software-defined networks." Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, 2010.

36. https://opennetworking.org/onos/ accessed 13/02/2024.

37. S. Orlowski, R. Wessäly, M. Pióro, and A. Tomaszewski, "Sndlib 1.0—survivable network design library," *Networks An Int. J*, Vol. 55, no. 3, pp. 276–86, 2010.

38. S. Balon, and G. Monfort. "The traffic matrices and topology of the Abilene network". (http://totem.run.montefiore.ulg.ac.be/datatools.html).

39. B. Zhang, et al., "Cte: cost-effective intra-domain traffic engineering," *ACM SIGCOMM Computer Commun. Rev.*, Vol. 44, no. 4, pp. 115–6, 2014. DOI: 10.1145/2740070.2631438

40. S. Uhlig, et al., "Providing public intradomain traffic matrices to the research community," *ACM SIGCOMM Computer Commun. Rev.*, Vol. 36, no. 1, pp. 83–6, 2006. DOI: 10.1145/1111322.1111341

41. K. Poularakis, et al., "One step at a time: optimizing SDN upgrades in ISP networks," in *IEEE INFOCOM 2017-IEEE conference on computer communications*, Atlanta, GA, USA: IEEE, 2017, pp. 1–9.

42. D. Levin, et al. "Panopticon: reaping the Benefits of incremental SDN deployment in enterprise networks." 2014 USENIX Annual Technical Conference (USENIX ATC 14). 2014.

## AUTHORS



**Samiullah Mehraban** is pursuing PhD in the Department of Computer Science & Engineering at Delhi Technological University, Delhi, India. He completed his Btech in computer science at Kabul Education University, Kabul, Afghanistan, and ME at Gujarat Technological University, Gujarat, India. His research areas are computer networks, software-defined networks, hybrid SDN networks, network security, and wireless sensor networks. He worked as an academic and professional member of the teacher education directory at the Ministry of Education in Afghanistan.

**Corresponding author. Email:** mehraban780@gmail.com.



**RajeshKumarYadav** works as an associate professor in the Department of Computer Science and Engineering at Delhi Technological University, New Delhi, India. He received his BE (CSE) from G. B. Pant Engineering College Pauri (H.N.B Garahwal University), Mtech (CS) from UPTU, Lucknow, and PhD (computer engineering) from Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India. His research interests include mobile ad hoc networks, SDN networks, mobile computing, computer networks, theory of computation, PCA, and wireless sensor networks. He has published many research papers in Conferences and reputed international journals.

**Email:** rkyadav@dtu.ec.in.

Journal of Electronic Materials
https://doi.org/10.1007/s11664-025-11957-w

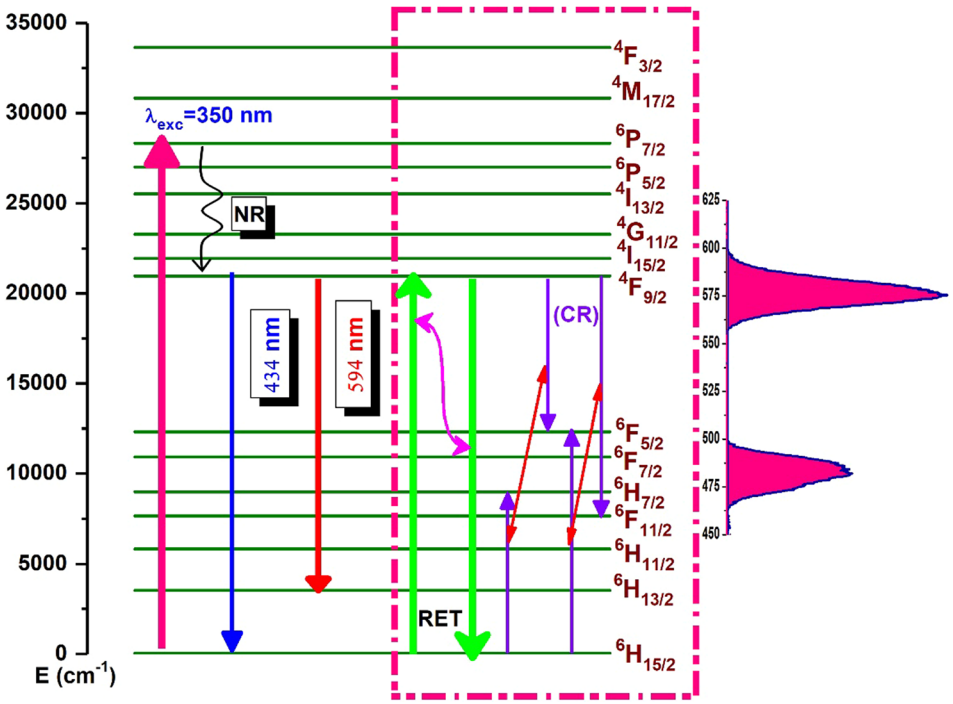
ORIGINAL RESEARCH ARTICLE

# Structural, Judd–Ofelt and Photoluminescence Properties of $Dy^{3+}$ Ions Doped SrO-MgO-$Na_2O$-$Al_2O_3$-$B_2O_3$ Glasses for Visible Laser and W-LED Applications

G. Muralidhar[1] · Mahamuda Shaik[1] · P. Sailaja[2] · K. Swapna[1] · M. Venkateswarlu[1] · G. Dedeepya[1] · A. S. Rao[3]

Received: 23 July 2024 / Accepted: 27 March 2025
© The Minerals, Metals & Materials Society 2025

## Graphical Abstract




**Keywords** $Dy^{3+}$ ions · XRD · FT-IR · photoluminescence · laser · W-LED

## Introduction

Glasses based on a combination of sodium and borate are attracting the attention of researchers, as they are a sturdy host material for doping heavy metal oxides (HMO) and transition metal oxides (TMO).[1,2] Nowadays, glasses doped with rare earth (RE) ions play an important role in industry and science because of their excellent optical, physical, and chemical features. Optical amplifiers and optical fiber waveguide lasers which are optimal for Q-switching and tuning are made possible with glasses doped with RE ions.[3] From

✉ Mahamuda Shaik
mahamuda.ss@gmail.com; mahamuda.ss@kluniversity.in

1 Department of Physics, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, AP 522 502, India

2 Department of Physics, Vardhaman College of Engineering, Shamshabad, Telangana 501 218, India

3 Department of Applied Physics, Delhi Technological University, Bawana Road, New Delhi 110 042, India

Published online: 07 May 2025

the literature it is evident that the glasses are good optical materials when compared to phosphors and ceramics, as they have some exceptional features including thermal stability and the ability to adsorb high concentrations of RE ions. It is also known that the glass fabrication process is cost-effective and easy.[4] The major applications of glasses doped with RE ions include magneto-optical devices, medical lasers, optical memory devices, laser material preparation, flat panel displays, fluorescent lamps, and eye-safe lasers.[5–12] Other applications of RE-doped glasses include channel waveguides, infrared (IR)-to-visible up-converters, white light-emitting diodes (W-LEDs), solid-state lasers, light converters, and nonlinear optical devices.[4]

The optical properties of RE ions are mainly affected by the surrounding ligand field and the phonon energies of the host glass matrix. Importantly, a host glass with relatively low phonon energy exhibits high quantum efficiency, which can be useful when designing optical fiber amplifiers and lasers.[13] Borate glasses are host materials with good RE ion solubility that are popularly known for their low melting point, high transparency, and high thermal stability.[14] $B_2O_3$ is a good glass former but is less popular due to its high phonon energy.[15] Apart from this drawback, the borate glass system has many unique and interesting properties that are desirable for doping RE ions. The inclusion of heavy metal oxide in a borate glass reduces the high phonon energy and enhances the intensity of fluorescence in the visible region. These systems are promising for applications such as electro-optic modulators, electro-optic switches, and nonlinear parametric communications.[16,17]. The non-radiative properties of some glasses are increased by the relatively high phonon energies of some glasses. Borate, silicate, phosphate, and tellurite glasses typically exhibit high phonon energies of 1400 $cm^{-1}$, 1100 $cm^{-1}$, 1200 $cm^{-1}$, and 700 $cm^{-1}$, respectively.[18,19]

To enhance the photonic features of a glass system, that system must possess phonon energy that is as low as possible. Therefore, for the present work, borate glass systems with chemical constituents $Al_2O_3$, $Na_2O$, MgO, and $Sr_2CO_3$ were chosen as network modifiers and intermediates. Among different glass intermediates, $Al_2O_3$was chosen as it can boost the fluorescence of activated ions in host glass materials. The presence of $Al_2O_3$ in a glass increases the thermal stability and mechanical strength and it also prevents devitrification.[20] $Al^{3+}$ions improve the optical response of the material by inhabiting octahedral and tetrahedral sites in the host glass system.[21] The presence of $Al_2O_3$ can suppress crystallization and improve the stability of the glass structure.[5] It also modifies the host glass network structure, enhancing the optical performance of metal ions in the host glass network.[17] Hence, these glasses exhibit attractive properties such as high electrical resistivity, large Young's modulus, high hardness, low density, and low coefficient of thermal expansion.[17,22,23] Aluminum trioxide is an intermediate that is added to a glassy network to enhance fluorescence and it also allows the uniform distribution of RE ions, minimizing the transition temperature of glass.[24] The coordination state of boron oxide and alumina in the glass structure changes with the composition of the glass, which affects the luminescence properties of rare earth ions.[25]

To decrease the viscosity of the molten melt and melting temperature of the glass, $Na_2O$ can be doped into the glass composition. The crystallization tendency can also be controlled by the addition of $Na_2O$, which increases the integrity of the network.[3] The inclusion of $Na_2O$ in any glass network will not only increase the refractive index and density of the glass, but also improve the glass forming ability, physical characteristics, luminescence intensity, and resistance to moisture. The structure of the host glass can also be modified by changing the triangular $BO_3$ structural units to tetrahedral $BO_4$ units, leading to the formation of non-bridging oxygens (NBOs).[26] Doping of alkaline earth metals such as calcium oxide (CaO) and magnesium oxide (MgO) into the glass network will enhance its chemical stability and resistance to acid.[27] The MgO-$B_2O_3$ binary system presents unique microwave dielectric properties due to the crystallization of the magnesium borate ($Mg_2B_2O_5$) crystalline phase.[28]

RE ions are considered notable photonic centers for host materials such as glasses and phosphors due to their wide range of energy levels in the ultraviolet (UV)–visible–near-infrared (NIR) regions, which are associated with 4f-5d and 4f-4f electronic transitions.[29] Among the other RE ions, dysprosium ($Dy^{3+}$) ions are promising candidates due to their favorable luminescent characteristics.[30] $Dy^{3+}$ ions are also considered the best choice of activating material in the preparation of phosphor materials, which are crucial in the preparation of white light-emitting phosphors. Glasses doped with $Dy^{3+}$ ions are also useful in fabricating solid-state lasers in the NIR and visible regions. On the whole, $Dy^{3+}$ion-doped glasses have gained enormous recognition in the evolution of W-LEDs because of their superior properties such as good thermal stability, easy fabrication, good transparency, and low cost of synthesis compared to phosphors.[31] Furthermore, the host glass matrix activated with dysprosium ($Dy^{3+}$) ions have excellent luminescent centers, exhibiting blue ($^4F_{9/2} \rightarrow {}^6H_{15/2}$), yellow ($^4F_{9/2} \rightarrow {}^6H_{13/2}$), and red ($^4F_{9/2} \rightarrow {}^6H_{11/2}$) characteristic emissions from the visible to NIR regions.[32] Moreover, it is easy to pump glasses doped with $Dy^{3+}$ions. $Dy^{3+}$ ion-activated glassy systems exhibit two strong bands, blue (~ 470–500 nm) and yellow (~ 570–600 nm).[3] Between these two transitions, the one positioned in the blue region is known as a magnetic dipole transition and that in the yellow region is an electric

dipole transition, also referred to as a hypersensitive transition, as it is highly sensitive to the local environment.[4]

Though many $Dy^{3+}$ ion-doped glasses have been synthesized and their optical properties have been studied, there is still a need for a suitable glass host with relatively low phonon energy to obtain improved optical properties. In order to achieve better luminescence efficiency, the aforementioned chemical constituents (SrO-MgO-$Na_2O$-$Al_2O_3$-$B_2O_3$) were selected based on their key features, and alkaline earth aluminoborate (AEAlB) glassy systems doped with different concentrations of $Dy^{3+}$ ions were synthesized. All the optical properties were studied to optimize the luminescent material for visible laser and W-LED applications.

## Experimental Details

Different concentrations ranging from 0.1 mol.% to 2.5 mol.% of $Dy^{3+}$ ions doped AEAlB glassy systems were prepared with high-purity (99.99%) analytical-grade chemicals including $B_2O_3$, $Al_2O_3$, MgO, $Sr_2CO_3$, and $Dy_2O_3$ using the traditional melt quenching method. The prepared glasses are designated as follows.

AEAlBDy01:19.9SrO-10MgO-5$Na_2O$-5$Al_2O_3$-60$B_2O_3$-0.1$Dy_2O_3$
AEAlBDy05:19.5SrO-10MgO-5$Na_2O$-5$Al_2O_3$-60$B_2O_3$-0.5$Dy_2O_3$
AEAlBDy10:19.0SrO-10MgO-5$Na_2O$-5$Al_2O_3$-60$B_2O_3$-1.0$Dy_2O_3$
AEAlBDy15:18.5SrO-10MgO-5$Na_2O$-5$Al_2O_3$-60$B_2O_3$-1.5$Dy_2O_3$
AEAlBDy20:18.0SrO-10MgO-5$Na_2O$-5$Al_2O_3$-60$B_2O_3$-2.0$Dy_2O_3$
AEAlBDy25:17.5SrO-10MgO-5$Na_2O$-5$Al_2O_3$-60$B_2O_3$-2.5$Dy_2O_3$

According to the chemical composition, all chemicals were weighed using an electrical balance and placed into an agate mortar. The chemicals were crushed well and the whole mixture was poured into a silica crucible. The silica crucibles were placed in an electric furnace operated at 1000°C for 40 min until a molten melt formed. The obtained mixture was then immediately transferred into a preheated brass plate and pressed with another brass plate to obtain circular glass samples with moderate thickness. The glass samples thus formed were annealed at 200°C to eliminate internal stress and remove air bubbles during quenching.

### Characterization Techniques

X-ray diffraction (XRD) patterns were recorded on an X'Pert Pro diffractometer using Cu $K_\alpha$ ($\lambda = 1.54059$ Å) radiation at a current of 30 mA and voltage of 40 kV with a step size 0.01°. The Fourier transform infrared (FT-IR) spectrum of the undoped AEAlB glass was recorded in the range 400–3600 $cm^{-1}$ using a PerkinElmer spectrometer with a spectral resolution of ±1 $cm^{-1}$. The densities and refractive indices of the synthesized glass samples were measured using the Archimedes and Brewster angle methods. A Jasco V-670 UV–Vis–NIR spectrophotometer was used to record the absorption spectra of the synthesized glass samples in the range of 340–2000 nm with 0.1 nm resolution. A Hitachi F7000 fluorescence spectrophotometer was used for excitation, photoluminescence (PL), and PL decay spectral measurements with a spectral resolution

**Table I** Physical properties of $Dy^{3+}$ ion-doped AEAlB glasses

| Physical property | AEAlB | | | | | |
|---|---|---|---|---|---|---|
| | Dy0.1 | Dy0.5 | Dy1.0 | Dy1.5 | Dy2.0 | Dy2.5 |
| Refractive index ($n_d$) | 2.461 | 2.478 | 2.480 | 2.490 | 2.501 | 2.528 |
| Density, $d$ (g/$cm^3$) | 2.851 | 2.868 | 2.889 | 2.911 | 2.932 | 2.954 |
| Average molecular weight, $M$ (g) | 79.07 | 79.97 | 81.10 | 82.23 | 83.35 | 84.48 |
| $Dy^{3+}$ ion concentration, N ($10^{21}$ ions/$cm^3$) | 2.171 | 10.79 | 21.45 | 31.98 | 42.37 | 52.64 |
| Mean atomic volume, $v_m$ (g/$cm^3$/atm) | 27.73 | 27.88 | 28.07 | 28.24 | 28.43 | 28.60 |
| Molar refraction, $R_m$ ($cm^{-3}$) | 17.40 | 17.60 | 17.74 | 17.91 | 18.101 | 18.37 |
| Dielectric constant ($\varepsilon$) | 6.058 | 6.140 | 6.150 | 6.204 | 6.257 | 6.394 |
| Optical dielectric constant ($\varepsilon$-1) | 5.058 | 5.140 | 5.150 | 5.204 | 5.257 | 5.394 |
| Reflection losses $R$ (%) | 0.178 | 0.180 | 0.180 | 0.182 | 0.1838 | 0.187 |
| Polaron radius ($r_p$) (Å) | 3.112 | 1.823 | 1.450 | 1.269 | 1.156 | 1.075 |
| Interatomic distance ($r_i$) (Å) | 7.722 | 4.524 | 3.598 | 3.150 | 2.868 | 2.668 |
| Molar polarizability ($\alpha_m$) ($10^{-25}$) | 24.02 | 24.35 | 24.35 | 24.60 | 24.81 | 25.36 |
| Field strength, $F$ ($10^{15}$ $cm^{-2}$) | 0.309 | 0.902 | 1.426 | 1.860 | 2.244 | 2.594 |

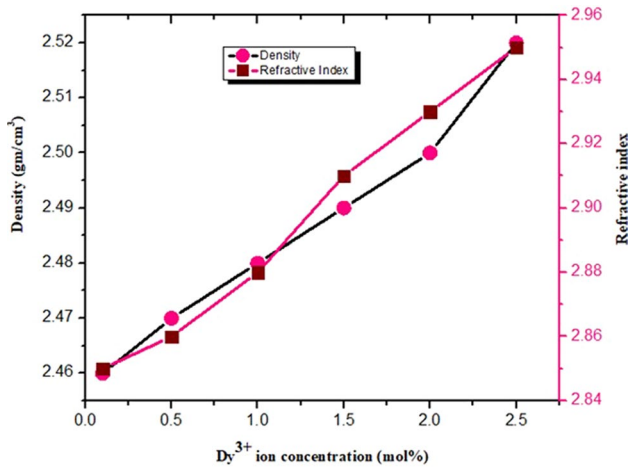



**Fig. 1** Graphical relation between density and refractive index of AEAlBDy glass samples.

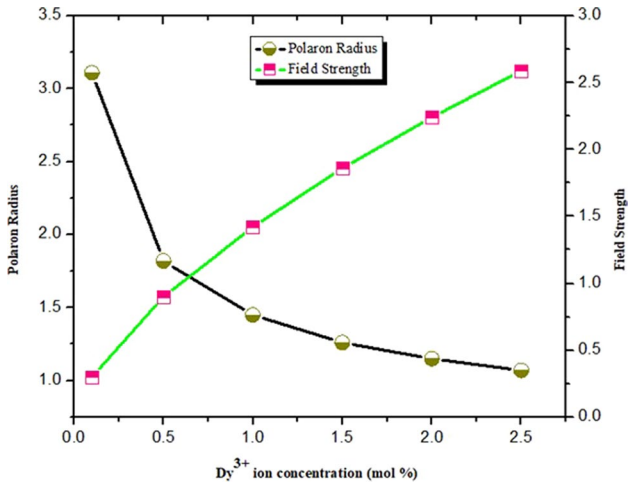



**Fig. 2** Graphical relation between polaron radius and field strength of AEAlBDy glass samples.

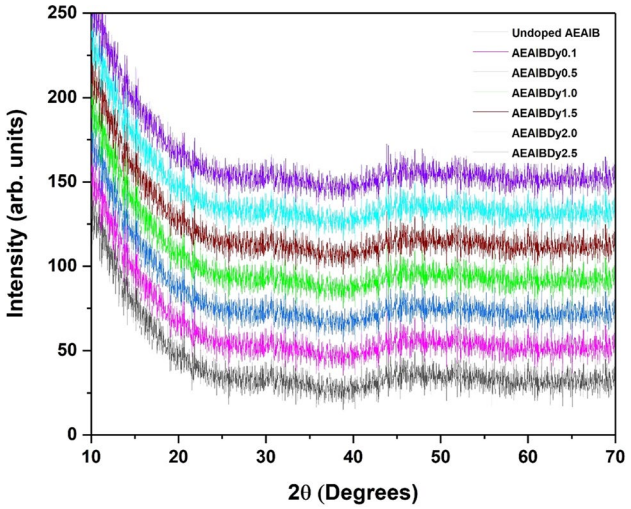



**Fig. 3** XRD patterns of undoped AEAlB glass and AEAlBDy glasses.

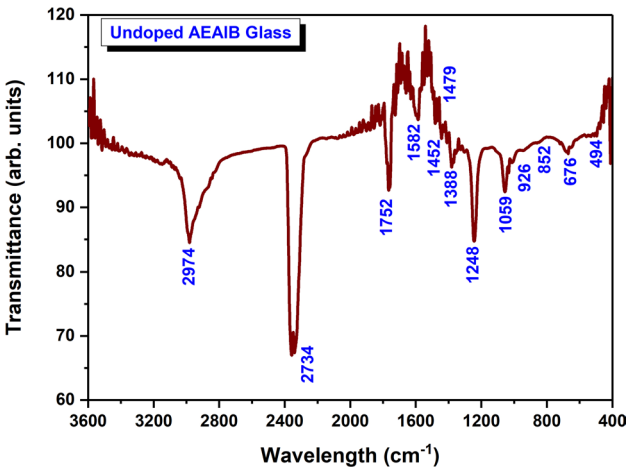



**Fig. 4** FT-IR spectrum of undoped AEAlB glass sample.

of ±1 nm, and a 150 W xenon lamp was used as the excitation source.

## Results and Discussion

### Physical Parameters

The densities of the prepared glasses were determined using the Archimedes method with water as the immersion liquid. Brewster's angle method (He–Ne laser with a wavelength 650 nm line) was employed to measure the refractive indices of the glasses. Employing the measured values of density and refractive index, several physical parameters were evaluated[33] and are shown in Table I. Table I displays various parameters evaluated form the density and refractive index such as average molecular weight, molar refraction, dielectric constant, optical dielectric constant, reflection loss, polaron radius, interatomic distance, molar polarizability, and field strength. The density and refractive index of AEAlBDy glasses exhibit a linear relationship, and its graphical representation is shown in Fig. 1. The polaron radius tends to decrease, whereas the field strength tends to increase with $Dy^{3+}$ ion content ranging from 0.1 mol.% to 2.5 mol.%. The graphical relation between polaron radius

and field strength with $Dy^{3+}$ concentration is shown in Fig. 2.

## XRD Spectral Analysis

Figure 3 shows the XRD patterns of undoped AEAlB and $Dy^{3+}$ ion-doped AEAlB glass samples. The pattern shows a broad hump with no sharp peaks. Hence, it can be confirmed that the fabricated AEAlB glass systems are noncrystalline in nature.

## Fourier Transform Infrared (FT-IR) Spectroscopy

FT-IR analysis is mainly used to note the various vibrational groups at different wavelengths. The FT-IR spectrum of undoped AEAlB glass sample is shown in Fig. 4.

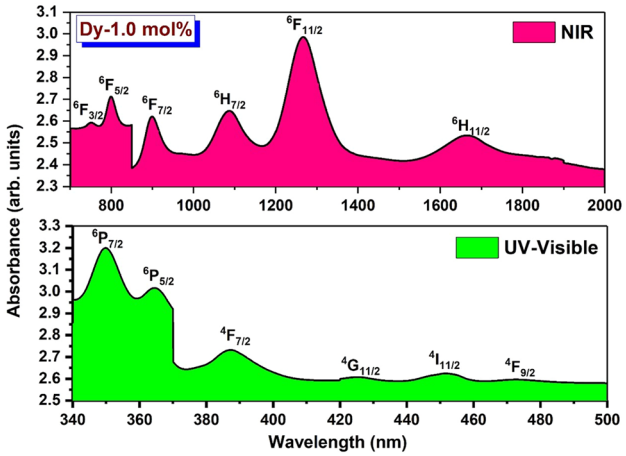



**Fig. 5** Absorption spectrum of 1 mol.% $Dy^{3+}$ ion-activated AEAlB glasses.

The spectrum was recorded in the wavelength range of 400–3600 $cm^{-1}$ at room temperature, and it is composed of 13 bands, at 494 $cm^{-1}$, 676 $cm^{-1}$, 852 $cm^{-1}$, 926 $cm^{-1}$, 1059 $cm^{-1}$, 1248 $cm^{-1}$, 1388 $cm^{-1}$, 1452 $cm^{-1}$, 1479 $cm^{-1}$, 1582 $cm^{-1}$, 1752 $cm^{-1}$, 2734 $cm^{-1}$, and 2974 $cm^{-1}$. The band at 494 $cm^{-1}$ is due to metal cation ($Sr^{2+}$, $Ba^{2+}$ ) $AlO_6$ octahedral stretching vibrations.[34] The band positioned at 676 $cm^{-1}$ is due to bending vibrations of B-O-B bonds occurring from pentaborate groups.[35] The bands at 852 and 926 $cm^{-1}$ are due to symmetrical stretching vibrations by $BO_4$ units. The bands between 1200 $cm^{-1}$ and 1500 $cm^{-1}$ are due to B–O stretching vibrations of trigonal $BO_3$ units.[35] The bands located at 2734 $cm^{-1}$ and 2974 $cm^{-1}$ are due to hydrogen bonding.[36]

## Absorption Spectral Analysis

Absorption is the primary process that is involved in the lasing action. The absorption spectrum of the AEAlB glass doped with 1 mol.% of $Dy^{3+}$ ions was recorded in the UV–Vis and NIR regions, as shown in Fig. 5, in the wavelength ranges of 340–500 nm and 800–2000 nm. The recorded absorption spectrum in the UV–Vis and NIR regions each consist of six transitions starting from the ground state $^6H_{15/2}$ to the various excited states, namely $^6P_{7/2}$, $^6P_{5/2}$, $^4F_{7/2}$, $^4I_{11/2}$, $^4G_{11/2}$, and $^6F_{3/2}$ for the UV–Vis region, and $^6F_{5/2}$, $^6F_{7/2}$,$^6H_{7/2}$,, $^6F_{11/2}$, and $^6H_{11/2}$ for the NIR region. The corresponding wavelengths are 349 nm, 364 nm, 387 nm, 425 nm, 452 nm, 472 nm, 753 nm, 799 nm, 901 nm, 1085 nm, 1266 nm, and 1673 nm, respectively. All these transitions were assigned based on the paper by Carnall et al.[37] From the spectra, it is observed that the transition $^6H_{15/2} \rightarrow {}^6F_{11/2}$ with a wavelength of 1266 nm is very sensitive to the environment surrounding the rare earth ion in

**Table II** Experimental ($f_{exp} \times 10^{-6}$) and calculated ($f_{cal} \times 10^{-6}$) oscillator strengths and root-mean-square deviation ($\delta_{rms}$) of $Dy^{3+}$ ion-doped AEAlB glasses

| Transition from $^4P_{5/2}$ | AEAlBDy0.1 | | AEAlBDy0.5 | | AEAlBDy1.0 | | AEAlBDy1.5 | | AEAlBDy2.0 | | AEAlBDy 2.5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $f_{exp}$ | $f_{cal}$ | $f_{exp}$ | $f_{cal}$ | $f_{exp}$ | $f_{cal}$ | $f_{exp}$ | $f_{cal}$ | $f_{exp}$ | $f_{cal}$ | $f_{exp}$ | $f_{cal}$ |
| $^6P_{7/2}$ | 0.609 | 0.013 | 0.727 | 0.366 | 0.901 | 0.518 | 0.858 | 0.364 | 0.717 | 0.295 | 0.637 | 0.504 |
| $^4F_{7/2}$ | 0.778 | 0.402 | 0.815 | 0.076 | 1.295 | 0.551 | 1.166 | 0.483 | 0.985 | 0.444 | 0.846 | 0.408 |
| $^4G_{11/2}$ | 0.377 | 0.011 | 0.950 | 0.023 | 1.383 | 0.030 | 1.326 | 0.025 | 1.029 | 0.020 | 0.931 | 0.024 |
| $^4I_{11/2}$ | 0.742 | 1.136 | 0.775 | 1.253 | 1.075 | 1.444 | 1.062 | 1.305 | 0.948 | 1.179 | 0.862 | 1.015 |
| $^4F_{9/2}$ | 0.811 | 0.401 | 0.849 | 0.427 | 1.018 | 0.482 | 1.001 | 0.434 | 0.914 | 0.403 | 0.830 | 0.342 |
| $^6F_{3/2}$ | 0.632 | 0.526 | 0.652 | 0.556 | 0.761 | 0.628 | 0.657 | 0.565 | 0.608 | 0.526 | 0.507 | 0.444 |
| $^6F_{5/2}$ | 2.093 | 2.783 | 2.394 | 2.945 | 3.737 | 3.323 | 3.042 | 2.992 | 2.491 | 2.785 | 2.191 | 2.352 |
| $^6F_{7/2}$ | 4.391 | 5.106 | 4.970 | 5.443 | 5.433 | 6.151 | 4.749 | 5.529 | 4.530 | 5.141 | 4.088 | 4.368 |
| $^6H_{7/2}$ | 1.286 | 0.232 | 1.853 | 0.246 | 2.560 | 0.277 | 2.487 | 0.250 | 1.942 | 0.232 | 1.567 | 0.196 |
| $^6F_{11/2}$ | 10.304 | 10.377 | 13.387 | 13.434 | 16.602 | 16.630 | 15.095 | 15.141 | 12.402 | 12.450 | 11.467 | 11.486 |
| $^6H_{11/2}$ | 3.764 | 3.185 | 3921 | 3.548 | 4.340 | 4.111 | 4.082 | 3.717 | 3.712 | 3.337 | 3.037 | 2.886 |
| $\delta_{rms(\times 10^{-6})}$ | ±0.559 | | ±0.658 | | ±0.901 | | ±0.881 | | ±0.696 | | ±0.548 | |

the host network, and it obeys the selection rule $|\Delta S| = 0$, $|\Delta L| \leq 2, |\Delta J| \leq 2$. Thus, it is noted as a hypersensitive transition. The intensity of the absorption bands is referred to as oscillator strength and can be measured using the following expression.[3]

$$f_{\exp} = 4.318 \times 10^{-9} \int \varepsilon(\nu) \mathrm{d}\nu, \tag{1}$$

where $f_{\exp}$ represents the experimental oscillator strength, which can be calculated by blending the areas under the absorption bands, and $\varepsilon(\nu)$ represents the coefficient of molar extinction at a wavenumber $\nu$ (in $\mathrm{cm}^{-1}$). The oscillator strength of all the AEAlB glasses was calculated using Eq. (1), and the values are listed in Table II. Based on J–O theory, the measured oscillator strength can be approximated using the following expression.[38,39]

$$f_{\mathrm{cal}} = \left(\frac{8\pi^2 mc\nu}{3h(2J+1)}\right)\left(\frac{\left(n^2+2\right)^2}{n}\right), \tag{2}$$

where $m$ is the electron mass, $\nu$ is the energy transition ($\mathrm{cm}^{-1}$), $n$ is the refractive index, $c$ is the velocity of light, $h$ is the Planck constant, $2J+1$ is the spin multiplicity, and $(n^2+n)/9n$ is the Lorentz field correction factor.

By adopting the J–O theory, the intensity parameters $\Omega_\lambda = 2$, 4, 6 can be evaluated. The J–O intensity parameters are primarily used to find various ligand effects. There are short-range effects between $RE^{3+}$ and $RE^{3+}$ ions, such as the presence of an oxygen ligand and covalency, which can be described by $\Omega_2$, and it also shows considerable influence on the local environment around the $RE^{3+}$ ions. Thus, $\Omega_2$ is firmly influenced by hypersensitive transition. Long-range effects such as rigidity and viscosity of the host glass matrix can be understood by $\Omega_4$, $\Omega_6$ parameters. The measured intensity parameters are shown in Table III. The prepared AEAlBDy glass samples show a trend of $\Omega_2 > \Omega_6 > \Omega_4$. By this result, AEAlBDy glasses are covalent in nature and are in good agreement with some reported glasses.[40–44] Among all the AEAlBDy glass samples, AEAlBDy10 glass turns out to be a good candidate for visible light applications due to its relatively high $\Omega_2$ intensity parameters.

## Bandgap Analysis

In materials, the optical bandgap plays a major role in identifying the optical properties. In the field of photonics, the investigation of optical spectra over a large range of photon energy is one of the most productive tools for understanding and developing the electronic structure theory in crystalline and amorphous material. To study the structural disorder of amorphous materials in a glassy system, the optical bandgap energy plays a vital role. In general, the electronic band structure of crystalline materials is well defined, i.e., the edges of the conduction and valance bands are sharply defined, which leads to an explicit bandgap. But for amorphous materials, the determination of the optical bandgap is difficult because the edge of the tail of localized states extends into the forbidden region, which will make it difficult to describe the true optical bandgap. The occupancy of the localized states specifies the presence of vacancies, broken bonds, impurities, and dangling bonds. Because of the inter-band transition, noncrystalline materials show a large absorption region.[45,46] Mott and Davis derived the relation between the absorption coefficient $\alpha(\nu)$ and optical bandgap ($E_g$), expressed as follows.[47]

$$(\alpha h\nu)^n = B\left(h\nu - E_g\right), \tag{3}$$

where $n$ indicates the nature of the transition effect during the absorption process, $B$ indicates the band tailing constant parameter, and $h\nu$ indicates the energy of the incident photon.

By employing the absorption spectral data and with help of above equation, we find $n = 1/2$ for direct allowed transitions and $n = 2$ for indirect allowed transitions. The plots between $(\alpha h\nu)^{1/2}$ and $h\nu$(eV) for direct allowed transition,

**Table III** Judd–Ofelt parameters ($\Omega_\lambda \times 10^{-21}$ $\mathrm{cm}^2$) of $Dy^{3+}$ ion-doped AEAlB glasses with other reported $Dy^{3+}$ ion-doped glasses

| Glass name | $\Omega_2$ | $\Omega_4$ | $\Omega_6$ | Trend | References |
|---|---|---|---|---|---|
| AEAlBDy0.1 | 6.413 ± 0.004 | 0.819 ± 0.005 | 3.2360 ± 0.0001 | $\Omega_2 > \Omega_6 > \Omega_4$ | Present work |
| AEAlBDy0.5 | 8.225 ± 0.0001 | 0.412 ± 0.002 | 3.379 ± 0.001 | $\Omega_2 > \Omega_6 > \Omega_4$ | Present work |
| AEAlBDy1.0 | 10.220 ± 0.001 | 0.814 ± 0.001 | 3.806 ± 0.001 | $\Omega_2 > \Omega_6 > \Omega_4$ | Present work |
| AEAlBDy1.5 | 9.2620 ± 0.0001 | 0.388 ± 0.002 | 3.3980 ± 0.0001 | $\Omega_2 > \Omega_6 > \Omega_4$ | Present work |
| AEAlBDy2.0 | 7.4790 ± 0.0001 | 0.217 ± 0.005 | 3.1350 ± 0.0001 | $\Omega_2 > \Omega_6 > \Omega_4$ | Present work |
| AEAlBDy2.5 | 6.7300 ± 0.0001 | 0.995 ± 0.003 | 2.589 ± 0.001 | $\Omega_2 > \Omega_6 > \Omega_4$ | Present work |
| $MgB_20_3Dy0.7$ | 2.302 | 1.286 | 1.217 | $\Omega_2 > \Omega_4 > \Omega_6$ | 40 |
| BaBPDy0.4 | 4.71 | 1.27 | 0.99 | $\Omega_2 > \Omega_4 > \Omega_6$ | 41 |
| NbFSDy10 | 10.05 | 1.37 | 2.16 | $\Omega_2 > \Omega_6 > \Omega_4$ | 42 |
| K2GdF5:$Dy^{3+}$ | 2.51 | 0.94 | 2.12 | $\Omega_2 > \Omega_6 > \Omega_4$ | 43 |
| NaBiSrP:1.0$Dy^{3+}$ | 2.48 | 0.82 | 0.61 | $\Omega_2 > \Omega_4 > \Omega_6$ | 44 |

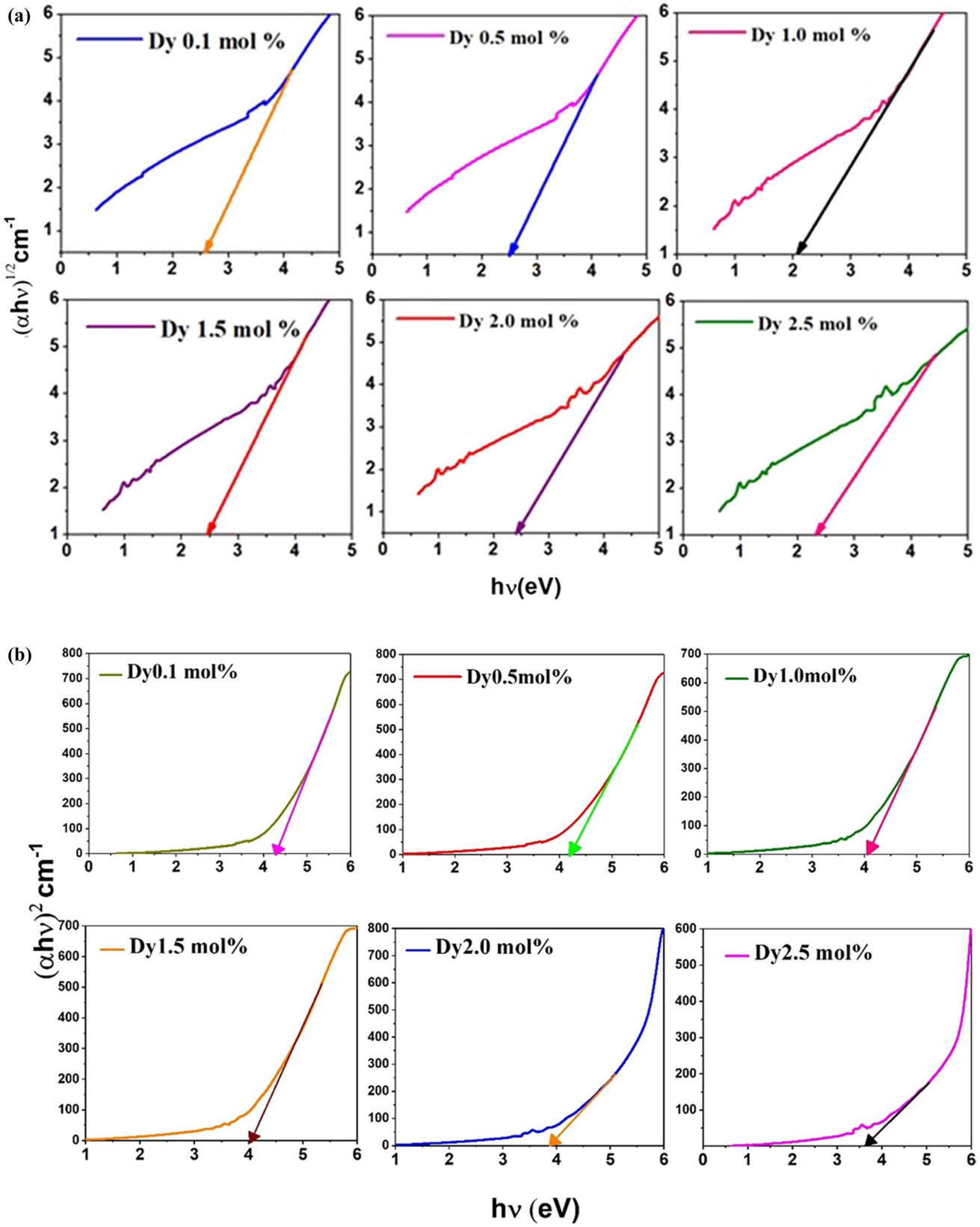



**Fig. 6** (a) Direct allowed bandgap of AEAlBDy glass samples. (b) Indirect allowed bandgap of AEAlBDy glass samples.

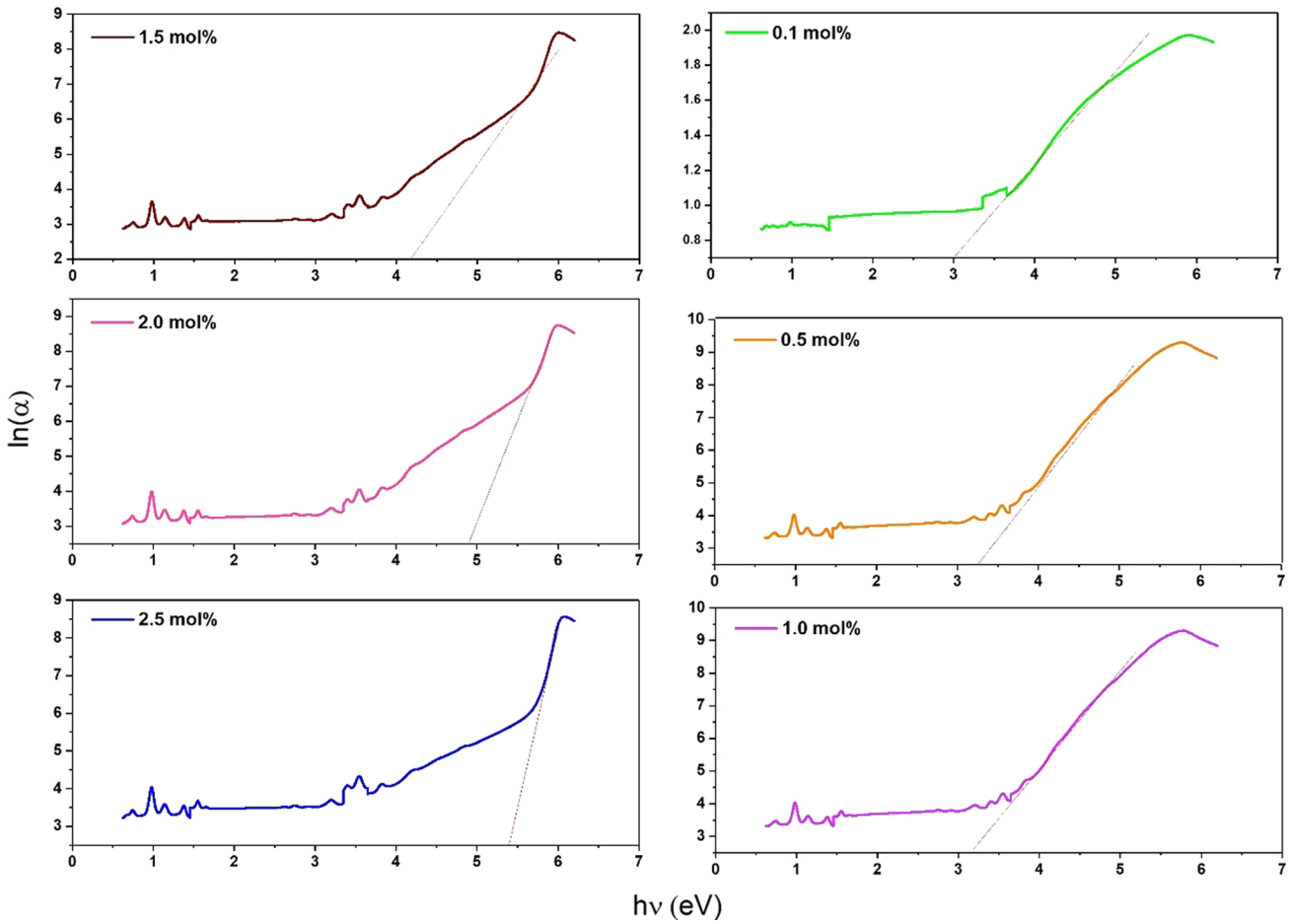



**Fig. 7** Urbach energies of AEAlBDy glass samples.

$(\alpha h\nu)^2$ and $h\nu$(ev) for indirect allowed transition are shown in Fig. 6a, b. The optical bandgap can be found by drawing a tangent line on $(\alpha h\nu) = 0$. The obtained optical bandgap values are $2.56 \pm 0.001$ eV, $2.50 \pm 0.001$ eV, $2.22 \pm 0.002$ eV, $2.48 \pm 0.001$ eV, $2.47 \pm 0.002$ eV, and $2.38 \pm 0.001$ eV for direct allowed transitions and $4.32 \pm 0.001$ eV, $4.22 \pm 0.001$ eV, $4.14 \pm 0.001$ eV, $4.05 \pm 0.002$ eV, $3.96 \pm 0.001$ eV, and $3.78 \pm 0.002$ eV for indirect allowed transitions of AEAlBDy samples ranging from 0.1 mol.% to 2.5 mol.%, respectively. For glassy materials doped with rare earth ions, both direct and indirect optical bandgap measurements are important to understand the optical absorption as well as emission properties. Introduction of rare earth ions in a glass matrix can influence the electronic structure and optical transitions in complex ways. Rare earth ions can contribute to direct transitions, which are related to the energy levels of the RE ion itself, and indirect transitions, which involve the interaction between host matrix states and RE ion states. Both types of transitions must be considered to understand the overall optical performance of RE ion-doped glasses.[48] The absorption coefficient changes exponentially

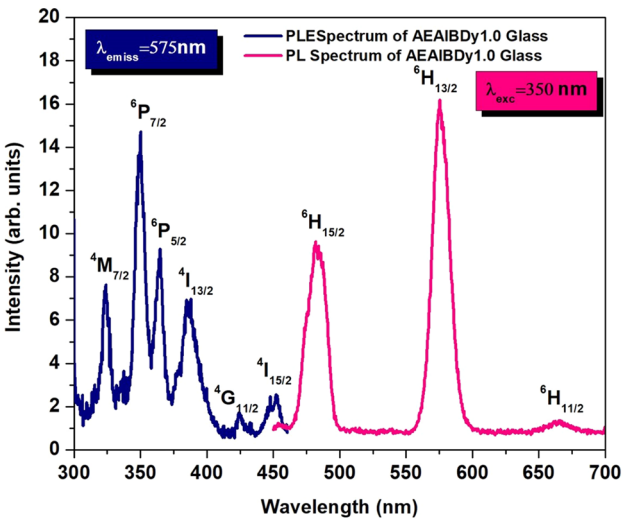



**Fig. 8** Excitation and emission spectra of 1 mol.% $Dy^{3+}$ ion-activated AEAlB glass.

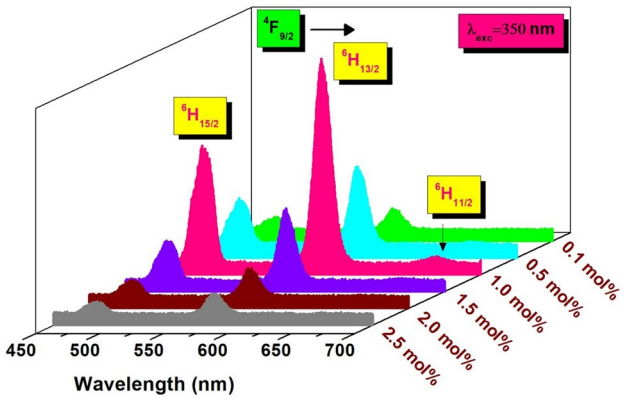



**Fig. 9** Emission spectra of $Dy^{3+}$ ion-activated AEAlB glasses.

with the photon's energy below the absorption edge, clearly indicating the presence of an Urbach tail. The Urbach tail is due to the presence of disorder in the prepared AEAlBDy glasses. The Urbach energies are evaluated using the expressions available in the literature[49,50] and are found to decrease as 0.31 eV, 0.271 eV, 0.263 eV, 0.251 eV, 0.202 eV, and 0.192 eV (Fig. 7). Hence, in the present series of glasses, Urbach energies decrease with the increase of $Dy^{3+}$ ion concentration, indicating the decrease of disorder in the glasses.

## Photoluminescence Analysis

Figure 8 shows the excitation and emission spectrum of the AEAlB glass sample doped with 1 mol.% of $Dy^{3+}$ ions. The excitation spectrum is recorded in the spectral range of 300–450 nm with an emission wavelength of 575 nm. Six excitation bands are observed positioned at 324 nm, 350 nm, 365 nm, 386 nm, 425 nm, and 453 nm, with energies of 30864, 28653, 27397, 25906, 23529, and 22075 $cm^{-1}$, respectively, whose electronic transitions are $^4M_{7/2}$, $^6P_{7/2}$,$^6P_{5/2}$,$^4I_{13/2}$,$^4G_{11/2}$, and$^4I_{15/2}$ respectively. The electronic transition $^6P_{7/2}$ positioned at 350 nm shows increased intensity when compared to all other excitation bands. Hence, it is used to record the photoluminescence spectrum of all AEAlBDy glass samples. Excitation is carried out primarily to pump a greater number of atoms from the ground state to an excited state to achieve population inversion, which is an essential parameter for lasing action. Various pumping methods have been widely used, including optical, electrical, chemical, and gas dynamic pumping. For the current work, optical pumping was employed.

The photoluminescence spectrum of the AEAlBDy1.0 glass sample was recorded at an excitation wavelength of 350 nm in the wavelength range 450–700 nm, as shown in Fig. 8 and as shown for all the glasses in Fig. 9. The spectra are composed of two bands positioned at 481 nm and 575 nm with energies of 20,790 and 17,391 $cm^{-1}$, bearing electronic transitions $^4F_{9/2} \rightarrow {}^6H_{15/2}$ and $^4F_{9/2} \rightarrow {}^6H_{13/2}$, respectively. The$^4F_{9/2} \rightarrow {}^6H_{15/2}$ electronic transition dominates over the $^4F_{9/2} \rightarrow {}^6H_{13/2}$ transition, even though both are present in the visible region of the electromagnetic spectrum. All these emission transitions arise in the PL spectrum far above the optical bandgap due to the direct-bandgap recombination mechanism and also the existence of electronic and magnetic dipole transitions.[51–53] Figure 9 shows the PL spectra of the AEAlBDy glass samples, which exhibit the same profile with no shift in emission bands. The $^4F_{9/2} \rightarrow {}^6H_{15/2}$ emission transition observed in the blue region arises due to the magnetic dipole transition, and its intensity is independent of the local environment of the host glass matrix. On the other hand, the $^4F_{9/2} \rightarrow {}^6H_{13/2}$ transition, present in the yellow region, emerges due to the electric dipole transition, and its intensity is highly dependent on the local environment of the host glass matrix.[45] It can be seen from Fig. 9 that the intensity of the$^4F_{9/2} \rightarrow {}^6H_{13/2}$ transition (yellow) increases up to 1 mol.% and then decreases. This may be due to the energy transfer that takes place between $Dy^{3+}$-$Dy^{3+}$ ions through cross-relaxation channels after 1 mol.%, a phenomenon known as concentration quenching. The Y/B intensities were also measured from the PL spectra of the AEAlBDy glass samples and were found to be 1.67, 1.72, 1.57, 1.87, 1.82, and 1.60. The Y/B intensity ratios provide the nature of covalency and asymmetry effects.[3] This Y/B intensity ratio can also be used to realize white light emission by adjusting the ratio of yellow-to-blue (Y/B) emission intensity.[3] In addition, by tuning the $Dy^{3+}$ ion concentration in OCBT glasses, the Y/B emission intensity ratio changes, which in turn leads to white light emission. The partial energy level diagram of the AEAlBDy glass samples is displayed in Fig. 10.

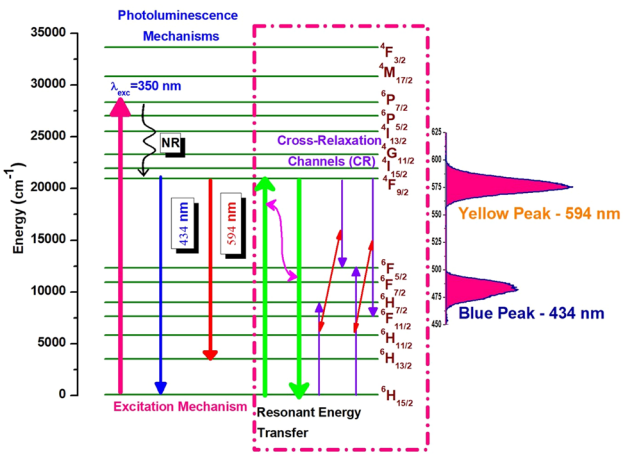



**Fig. 10** Partial energy level diagram of AEAlBDy glass system.

## Radiative Properties

To estimate the photoluminescence properties of the prepared glass samples, various radiative parameters including radiative transition probability ($A_R$), total radiative transition probability ($A_\tau$), branching ratio ($\beta_R$), and radiative lifetime ($\tau_R$) were evaluated. All these parameters were evaluated using J–O intensity parameters measured from the absorption spectra. By estimating the full width at half maximum ($\Delta\lambda_P$), other radiative properties including the stimulated emission cross-section ($\sigma_{sec}$), gain bandwidth ($\sigma_{sec} \times \Delta\lambda_P$), and optical gain ($\sigma_{sec} \times \tau_R$) were also evaluated using the formulas available in the literature.[54] Any host material doped with RE ions with high gain bandwidth and optical gain is considered highly suitable for optical fibers, and even indicates the high stability of the host material.[55] The stimulated emission cross-section ($\sigma_{sec}$) is another key parameter that plays a pivotal role in estimating the suitability of the prepared glass samples for lasing applications, and it can be evaluated by the following expression.[56]

$$\sigma_{sec}(aj, bj') = \left[\frac{\lambda_p^4}{8\pi c n^2 \Delta\lambda_p}\right] A_R(aj, aj'). \quad (4)$$

Here, $\lambda_P$ indicates the emission peak wavelength, $n$ is the refractive index of the glass sample, and $\Delta\lambda_p$ denotes the full width at half maximum or effective bandwidth. In addition to these key parameters, the branching ratio ($\beta_R$) also plays a vital role in the fabrication of laser devices and is very helpful in achieving the stimulated emission cross-section. The expression used for measuring the branching ratio is as follows.

**Table IV** Radiative transition probability ($A_R$) ($S^{-1}$), branching ratios ($\beta_R$), radiative lifetime ($\tau_R$) (ms), and total radiative transition probability ($A\tau$) for the observed emission transitions of $Dy^{3+}$ ion-doped AEAlB glasses

| Glass Name | Transition ($^4F_{9/2}$) | $A_R$ | $\beta_R$ | $A\tau$ | $\tau_R$ |
|---|---|---|---|---|---|
| AEAlBDy0.1 | $^6H_{15/2}$ | 374 ± 1 | 0.0650 ± 0.0002 | 5751 ± 14 | 173 ± 1 |
| | $^6H_{13/2}$ | 3612 ± 8 | 0.6280 ± 0.0004 | | |
| | $^6H_{11/2}$ | 1191 ± 8 | 0.2070 ± 0.0001 | | |
| AEAlBDy0.5 | $^6H_{15/2}$ | 467 ± 1 | 0.6820 ± 0.0001 | 6847 ± 10 | 146 ± 2 |
| | $^6H_{13/2}$ | 4449 ± 4 | 0.6490 ± 0.0002 | | |
| | $^6H_{11/2}$ | 1285 ± 6 | 0.1870 ± 0.0003 | | |
| AEAlBDy1.0 | $^6H_{15/2}$ | 1268 ± 1 | 0.0430 ± 0.0001 | 2890 ± 12 | 134 ± 1 |
| | $^6H_{13/2}$ | 15406 ± 5 | 0.5320 ± 0.0004 | | |
| | $^6H_{11/2}$ | 6435 ± 6 | 0.2220 ± 0.0003 | | |
| AEAlBDy1.5 | $^6H_{15/2}$ | 524 ± 1 | 0.0700 ± 0.0003 | 7474 ± 10 | 133 ± 2 |
| | $^6H_{13/2}$ | 4943 ± 7 | 0.6610 ± 0.0001 | | |
| | $^6H_{11/2}$ | 1319 ± 5 | 0.1760 ± 0.0002 | | |
| AEAlBDy2.0 | $^6H_{15/2}$ | 445 ± 1 | 0.0680 ± 0.0002 | 6544 ± 12 | 152 ± 3 |
| | $^6H_{13/2}$ | 4226 ± 7 | 0.6450 ± 0.0001 | | |
| | $^6H_{11/2}$ | 1238 ± 4 | 0.1890 ± 0.0001 | | |
| AEAlBDy2.5 | $^6H_{15/2}$ | 419 ± 1 | 0.0700 ± 0.0001 | 5975 ± 14 | 167 ± 1 |
| | $^6H_{13/2}$ | 3874 ± 5 | 0.6480 ± 0.0001 | | |
| | $^6H_{11/2}$ | 1070 ± 3 | 0.1790 ± 0.0002 | | |

**Table V** Stimulated emission cross-section ($\sigma_{sec} \times 10^{-21}$) ($cm^2$), gain bandwidth ($\sigma_{sec} \times \Delta_{\lambda P}$) ($10^{-27}$ $cm^3$), and optical gain ($\sigma_{sec} \times \tau_R$) ($10^{18} cm^2 s$) parameters evaluated for $Dy^{3+}$ ion-doped AEAlB glasses

| Spectral parameters | AEAlBDy0.1 | AEAlBDy0.5 | AEAlBDy1.0 | AEAlBDy1.5 | AEAlBDy 2.0 | AEAlBDy 2.5 |
|---|---|---|---|---|---|---|
| $^4F_{9/2} \rightarrow {}^6H_{15/2}$ | | | | | | |
| $\sigma_{sec}$ | 1.150 ± 0.001 | 0.940 ± 0.002 | 2.880 ± 0.003 | 0.850 ± 0.001 | 0.930 ± 0.002 | 0.813 ± 0.002 |
| $\sigma_{sec} \times \Delta_{\lambda p}$ | 0.800 ± 0.003 | 0.900 ± 0.001 | 2.680 ± 0.001 | 1.100 ± 0.002 | 0.920 ± 0.001 | 0.850 ± 0.001 |
| $\sigma_{sec} \times \tau_R$ | 6.590 ± 0.001 | 6.460 ± 0.001 | 8.310 ± 0.002 | 7.050 ± 0.001 | 6.500 ± 0.002 | 4.860 ± 0.001 |
| $^4F_{9/2} \rightarrow {}^6H_{13/2}$ | | | | | | |
| $\sigma_{sec}$ | 22.600 ± 0.003 | 23.600 ± 0.001 | 81.500 ± 0.001 | 20.200 ± 0.001 | 25.600 ± 0.003 | 15.300 ± 0.002 |
| $\sigma_{sec} \times \Delta_{\lambda p}$ | 15.800 ± 0.002 | 19.200 ± 0.001 | 66.300 ± 0.003 | 21.100 ± 0.002 | 17.900 ± 0.001 | 16.000 ± 0.001 |
| $\sigma_{sec} \times \tau_R$ | 130.000 ± 0.001 | 161.000 ± 0.001 | 236.000 ± 0.001 | 151.000 ± 0.001 | 168.000 ± 0.002 | 91.600 ± 0.001 |

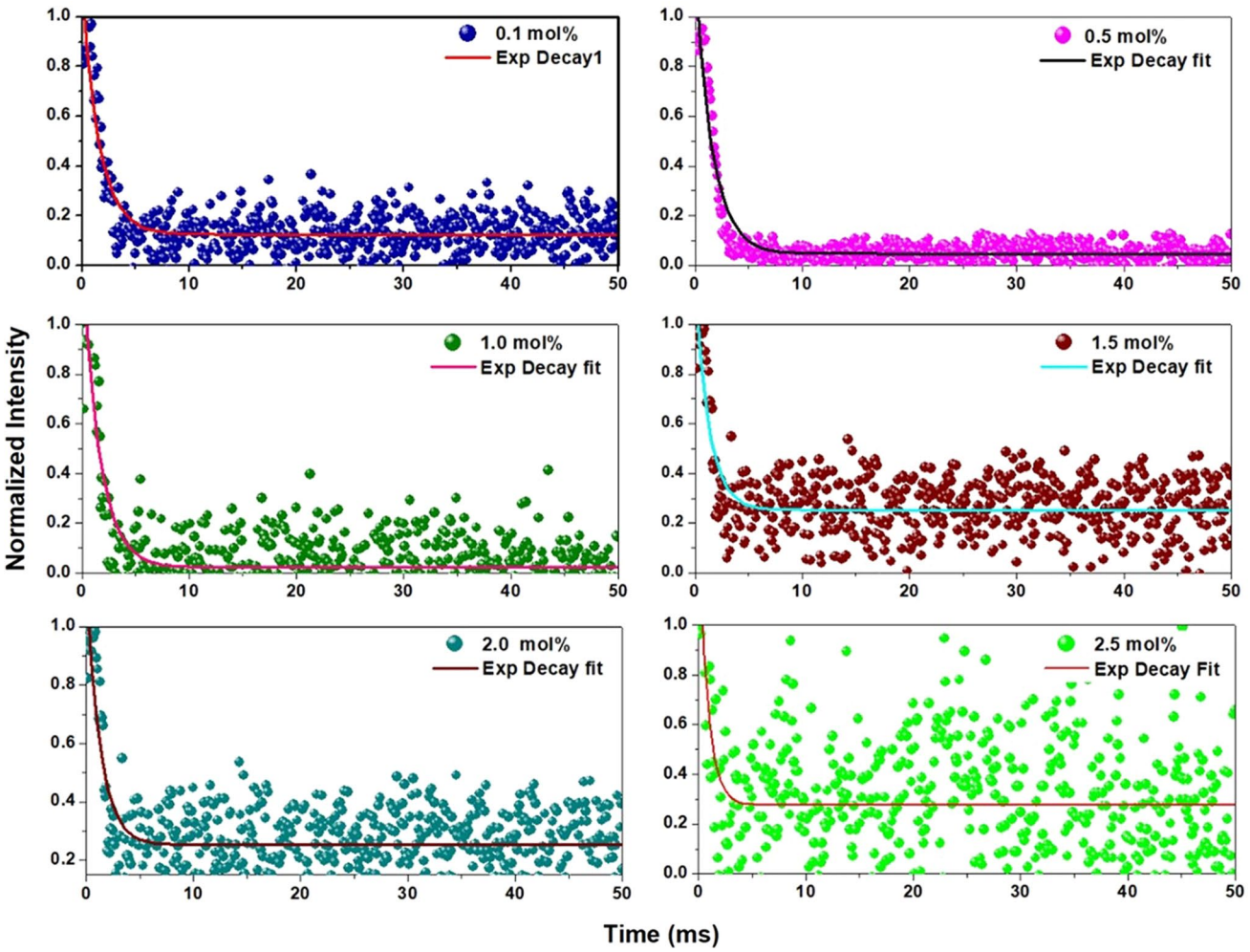



**Fig. 11** Decay curves of $Dy^{3+}$ ion-doped AEAlB glasses.

**Table VI** Experimental ($\tau_{exp}$) and radiative lifetime ($\tau_R$), quantum efficiency ($\eta$), non-radiative transition rate ($W_{NR}$), Y/B intensity ratio, and color coordinates of $Dy^{3+}$ ion-doped AEAlB glasses

| Glass name | $\tau_R$ (ms) | $\tau_{exp}$ (ms) | $\eta$ (%) | $W_{NR}$ ($ms^{-1}$) | Y/B ratio | ($x$, $y$) | CCT |
|---|---|---|---|---|---|---|---|
| AEAlBDy0.1 | 1.7700 ± 0.0001 | 1.2340 ± 0.0009 | 69 | 0.2463 | 1.67 | (0.34, 0.35) | 4982 |
| AEAlBDy0.5 | 1.4600 ± 0.0003 | 1.2170 ± 0.0001 | 83 | 0.1367 | 1.72 | (0.36, 0.38) | 4473 |
| AEAlBDy1.0 | 1.3400 ± 0.0002 | 1.2090 ± 0.0003 | 90 | 0.0808 | 1.57 | (0.35, 0.36) | 4737 |
| AEAlBDy1.5 | 1.3300 ± 0.0001 | 1.0080 ± 0.0001 | 75 | 0.2401 | 1.87 | (0.33, 0.33) | 5273 |
| AEAlBDy2.0 | 1.5200 ± 0.0002 | 0.9260 ± 0.0003 | 60 | 0.4220 | 1.82 | (0.33, 0.34) | 5200 |
| AEAlBDy2.5 | 1.6700 ± 0.0001 | 0.8340 ± 0.0001 | 49 | 0.6001 | 1.60 | (0.36, 0.40) | 4464 |

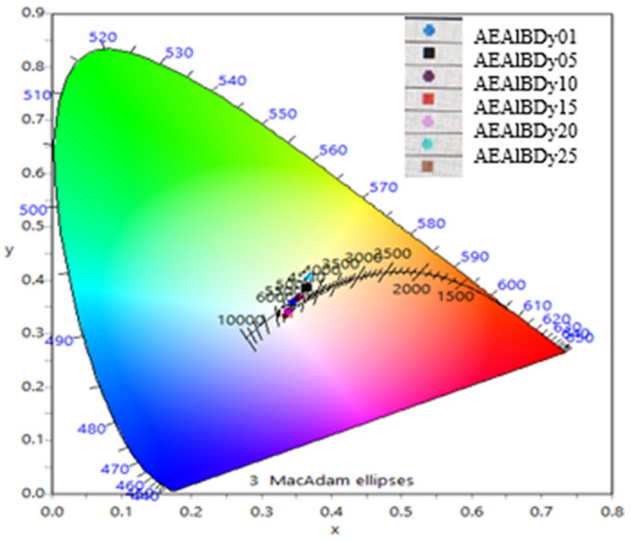



**Fig. 12** Color chromaticity diagram of AEAlBDy glass samples.

$$\beta_R = \frac{A_R}{A_T}, \tag{5}$$

where $A_R$, $A_\tau$, $\beta_R$, and $\tau_R$ values measured for the emission transitions of $Dy^{3+}$ ions activated AEAlB glasses are given in Table IV. The emission peak wavelength ($\lambda_P$), emission peak bandwidth ($\Delta\lambda_P$), gain bandwidth ($\sigma_{sec} \times \Delta\lambda_P$), and optical gain ($\sigma_{sec} \times \tau_R$) for the two emission transitions are shown in Table V. Among the two emission transitions, the intensified transition for 1 mol.% $Dy^{3+}$ ions doped AEAlB glasses exhibit the highest radiative lifetime ($\tau_R$), which ultimately leads to a relatively high stimulated emission cross-section required for better lasing action. From Table V, it can be seen 1 mol.% $Dy^{3+}$ ions activated AEAlB glasses have the highest value of $\sigma_{sec}$, $\sigma_{sec} \times \lambda_P$ and $\sigma_{sec} \times \tau_R$ which confirms that the prepared glass samples can be regarded as outstanding optical materials for photoluminescence devices.

## PL Lifetime Profile, Partial Energy Level Diagram, and Quantum Efficiency

The PL decay lifetime spectra of $Dy^{3+}$ ion-doped AEAlB glasses were recorded at an excitation wavelength of 350 nm and are shown in Fig. 11. The experimental lifetime ($\tau_{exp}$) of AEAlBDy glass samples were measured using the expression.

$$\tau_{exp} = \frac{\int tI(t)dt}{\int I(t)dt} \tag{6}$$

The obtained experimental lifetimes ($\tau_{exp}$) along with radiative lifetimes ($\tau_R$) were tabulated in Table VI. From Table VI it is noticed that the experimental lifetime ($\tau_{exp}$) of $Dy^{3+}$ ions in AEAlB glass shows a decrease with an increase in $Dy^{3+}$ ion content. The decrease in experimental lifetime ($\tau_{exp}$) is due to non-radiative transitions, which can be expressed as

$$\frac{1}{\tau_{exp}} = \frac{1}{\tau_R} + W_{NR}. \tag{7}$$

Here, $W_{NR}$ indicates non radiative relaxation probability, which is possible due to the energy transfer probability ($W_{ET}$) and multi-phonon relaxation rate ($W_{MP}$). Energy transfer probability mainly depends on $Dy^{3+}$ ion concentration, whereas multi-phonon relaxation depends on the host glass matrix. The energy transfer that takes place between $Dy^{3+}$-$Dy^{3+}$ ions through cross-relaxation channels is mainly responsible for the concentration quenching mechanism. The partial energy level diagram of AEAlBDy glass samples is shown in Fig. 10, and

**Table VII** Comparison of Judd–Ofelt parameter [$\Omega_2$ ($10^{-20}$ $cm^2$)], branching ratio ($\beta_R$), stimulated emission cross-section ($\sigma_{sec}$) ($10^{-21}$), quantum efficiency ($\eta$ (%)) (%), non-radiative transition rates ($W_{NR}$), and CIE chromaticity coordinates for the prominent emission transitions of $Dy^{3+}$ (1 mol.%)-doped OCBT glass with other reported $Dy^{3+}$þ ion-doped glasses

| Glass name | $\Omega_2$ | $\beta_R$ | $\sigma_{sec}$ | $\eta$ (%) | $W_{NR}$ ($ms^{-1}$) | (x, y) | References |
|---|---|---|---|---|---|---|---|
| AEAlBDy1.0 | 10.200 ± 0.001 | 0.5320 ± 0.0004 | 81.500 ± 0.001 | 90 | 808 | (0.35, 0.36) | Present work |
| ZPABDy0.5 | 5.71 | 0.2445 | 23.4 | 79.6 | 1688 | – | 59 |
| OCBTDy10 | 120 | – | 46.4 | 91 | – | (0.34, 0.33) | 4 |
| 0.5DZTFB | 82.8 | – | 5.61 | 45 | – | (0.34, 0.40) | 60 |
| BPB0.5D | 69.9 | – | 0.42 | 61 | – | (0.40, 0.42) | 61 |
| BaBPDy0.4 | 47.1 | 0.60 | 2.13 | – | – | (0.37, 0.42) | 62 |

clearly demonstrates all the possible energy transfer pathways made by $Dy^{3+}$ ions when excited by 350 nm. The $Dy^{3+}$ ions have been promoted from the ground state $^6H_{15/2}$ to various excited states and are responsible for occurrence of various mechanisms such as absorption, excitation, and spontaneous and stimulated emission transitions, as clearly depicted in Fig. 10. The quantum efficiency ($\eta$ %) of AEAlBDy glass samples was evaluated and is presented in Table VI. The quantum efficiency describes the capability of optical material for lasing action, and it is determined by taking the ratio of experimental lifetime to the radiative lifetime. The expression used for estimating quantum efficiency ($\eta$%) is as follows,[57,58]

$$\eta(\%) = \frac{\tau_{exp}}{\tau_R} \times 100. \quad (8)$$

Among all the AEAlBDy glass samples, AEAlBDy1.0 glass has relatively high quantum efficiency and low non-radiative transition rate as visualized in Table VI. Hence, AEAlBDy1.0 glass is considered the most suitable candidate for visible light lasers.

### CIE Chromaticity Coordinates

The color chromaticity diagram has been plotted for the prepared glass samples to estimate the color of light flourished by the glassy materials prepared in the present work. The CIE chromaticity coordinates and color correlated temperature (CCT) for all the prepared glass samples were determined from the emission spectra measured at an excitation wavelength of 350 nm and are listed in Table VI. Figure 12 shows the CIE plot with color coordinates for all the glass samples. Among all the glass samples, 1.0 mol.% $Dy^{3+}$ ion-activated AEAlB glass shows chromaticity coordinates very close to the white region in the CIE chromaticity diagram. This allows us to determine the suitability of the 1.0 mol.% $Dy^{3+}$ ion-doped AEAlB glass for the fabrication of W-LEDs. Table VII shows a comparison of crucial key parameters of 1 mol.% $Dy^{3+}$ ion-doped AEAlB glass with other reported glasses.[4,59–62]

## Conclusion

AEAlB glasses doped with $Dy^{3+}$ ions were fabricated by adopting the conventional melt quenching technique. XRD and FT-IR analysis confirmed the amorphous nature and various vibrational and stretching modes present in the glasses, respectively. The J–O parameters derived from the absorption spectral features follow the trend $\Omega_2 > \Omega_6 > \Omega_4$ for all the AEAlBDy glasses. Among all the glasses, AEAlBDy1.0 glass has the highest $\Omega_2$ parameter, confirming the highest covalency. The estimated stimulated emission cross-section and quantum efficiency confirm that AEAlBDy1.0 glass can be considered as a good optical material for lasing action in the visible region and also for W-LED applications. These results are also consistent with the CIE color coordinates of these glasses.

**Acknowledgments** The authors from Koneru Lakshmaiah Education Foundation (KLEF) are grateful to the management of KLEF. The authors Sk. Mahamuda and K. Swapna are also thankful to DST & DST-SERB for sanctioning FIST (SR/FST/PS-1/2018/35), PURSE Project (SR/PURSE/2023/196) and CRG Project (CRG/2022/003549).

**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. M. Ahmad, K. Aly, Y.B. Saddeek, and A. Dahshan, Glass transition and crystallization kinetics of $Na_2O$-$B_2O_3$-$Nb_2O_5$-$Bi_2O_3$ ceramic glasses. *J. Non Cryst. Solids* 546, 12026 (2020). https://doi.org/10.1016/j.jnoncrysol.2020.120260.
2. M.K. El-Mansy, H.M. Gomaa, N. Hendawy, and A.S. Morsy, Effect of exchange of $Bi^{+3}$ by $Nb^{+5}$ on the structural and optical properties of some $(BBiNb)_2O_7CaO$ oxide glasses. *J. Non-Cryst. Solids* 485, 42 (2018). https://doi.org/10.1016/j.jnoncrysol.2018.01.036.
3. V.M. Krishna, Sk. Mahamuda, R.A. Talewar, K. Swapna, M. Venkateswarlu, and A.S. Rao, $Dy^{3+}$ ions doped oxy-fluoro boro tellurite glasses for the prospective optoelectronic device applications. *J. Alloys Compd.* 762, 814 (2018). https://doi.org/10.1016/j.jallcom.2018.05.191.
4. P. Sailaja, Sk. Mahamuda, R.A. Talewar, K. Swapna, and A.S. Rao, Spectroscopic investigations of dysprosium ions doped oxy chloro boro tellurite glasses for visible photonic device applications. *J. Alloys Compd.* 789, 744 (2019). https://doi.org/10.1016/j.jallcom.2019.03.148.
5. Sk. Mahamuda, K. Swapna, M. Venkateswarlu, A.S. Rao, S. Shakya, and G.V. Prakash, Spectral characterisation of $Sm^{3+}$ ions doped Oxy-fluoroborate glasses for visible orange luminescent applications. *J. Lumin.* 154, 410 (2014). https://doi.org/10.1016/j.jlumin.2014.05.017.
6. R. Sharma, A.S. Rao, N. Deopa, M. Venkateswarlu, M. Jayasimhadri, D. Haranath, and G.V. Prakash, Spectroscopic study of $Pr^{3+}$ ions doped Zinc Lead Tungsten Tellurite glasses for visible photonic device applications. *Opt. Mater.* 78, 457 (2018). https://doi.org/10.1016/j.optmat.2018.02.054.
7. R.A. Talewar, Sk. Mahamuda, K. Swapna, M. Venkateswarlu, and A.S. Rao, Spectroscopic studies of $Sm^{3+}$ ions doped alkaline-earth chloro borate glasses for visible photonic applications. *Mater. Res. Bull.* 105, 45 (2018). https://doi.org/10.1016/j.materresbull.2018.04.033.
8. 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.* 194, 56 (2018). https://doi.org/10.1016/j.jlumin.2017.09.057.
9. R. Mi, J. Chen, Y. Liu, L. Mei, J. Yuan, Y. Xia, Z. Huang, and M. Fang, Crystal structure and luminescence property of a novel single-phase white light emission phosphor $KCaBi(PO_4)_2$:$Dy^{3+}$. *Mater. Res. Bull.* 86, 146 (2017). https://doi.org/10.1016/j.materresbull.2016.10.005.
10. M. Ding, Y. Li, D. Chen, H. Lu, J. Xi, and Z. Ji, Hexagonal crown-capped $NaYF_4$:$Ce^{3+}$/$Gd^{3+}$/$Dy^{3+}$ microrods: formation mechanism,

energy transfer and luminescence properties. *J. Alloys Compd.* 658, 952 (2016). https://doi.org/10.1016/j.jallcom.2015.10.273.
11. M. Ding, M. Xu, and D. Chen, A new non-contact self-calibrated optical thermometer based on $Ce^{3+}$→$Tb^{3+}$→$Eu^{3+}$ energy transfer process. *J. Alloys Compd.* 713, 236 (2017). https://doi.org/10.1016/j.jallcom.2017.04.202.
12. C.R. Kesavulu, H.J. Kim, S.W. Lee, J. Kaewkhao, N. Chanthima, and Y. Tariwong, Physical, vibrational, optical and luminescence investigations of $Dy^{3+}$-doped yttrium calcium silicoborate glasses for cool white LED applications. *J. Alloy. Compd.* 726, 1062 (2017). https://doi.org/10.1016/j.jallcom.2017.08.091.
13. Y.G. Choi and J. Heo, 1.3 μm emission and multiphonon relaxation phenomena in PbO- $Bi_2O_3$- $Ga_2O_3$ glasses doped with rare-earths. *J. Non-Cryst. Solids* 217, 199 (1997). https://doi.org/10.1016/S0022-3093(97)00189-0.
14. Sk. Mahamuda, K. Swapna, A. Srinivasa Rao, T. Sasikala, and L. Rama Moorthy, Reddish-orange emission from $Pr^{3+}$ doped zinc alumino bismuth borate glasses. *Physica B. Cond. Matter.* 428, 36 (2013). https://doi.org/10.1016/j.physb.2013.07.010.
15. S.M. Kaczmarek, $Li_2B_4O_7$ glasses doped with Cr Co, Eu and Dy. *Opt. Mater.* 19, 189 (2002). https://doi.org/10.1016/S0925-3467(01)00218-X.
16. P. Sailaja, Sk. Mahamuda, K. Swapna, M. Venkateswarlu, and A.S. Rao, Optical properties of Sm3+ions doped 10SrO-(10–x) $Al_2O_3$-10$BaCl_2$-60$B_2O_3$-10$TeO_2$ glasses for reddish orange laser applications. *Mater. Sci. Eng. B* 270, 115198 (2021). https://doi.org/10.1016/j.mseb.2021.115198.
17. G. Krishna Priya, S. Yusab, A. Ramesh Babu, N. Sree Ram, K. Rama Krishna, and V. Aruna, The eminence of copper ions on optical, electrical properties and morphology of $B_2O_3$-$Bi_2O_3$-$Al_2O_3$-MgO glasses. *J. Non-Cryst. Solids* 564, 120844 (2021). https://doi.org/10.1016/j.jnoncrysol.2021.120844.
18. M. Yamane and Y. Asahara, *Glasses of Photonics* (Cambridge: Cambridge University Press, 2000). https://doi.org/10.1017/CBO9780511541308.
19. M. Djamal, L. Yuliantini, R. Hidayat, K. Boonin, P. Yasaka, and J. Kaewkhao, Development of optical material based on glass doped rare earth for photonic devices. *Mater. Today Proc.* 43, 2531 (2021). https://doi.org/10.1016/j.matpr.2020.04.613.
20. Sk. Mahamuda, K. Swapna, P. Packiyaraj, A. Srinivasa Rao, and G. Vijaya Prakash, Optical studies of $Sm^{3+}$ ions doped zinc alumino bismuth borate glasses. *Opt. Mater.* 36, 362 (2013). https://doi.org/10.1016/j.saa.2014.01.025.
21. N. Deopa, A.S. Rao, Sk. Mahamuda, M. Gupta, M. Jayasimhadri, D. Haranath, and G. Vijaya Prakash, Spectroscopic studies of $Pr^{3+}$ doped lithium lead alumino borate glasses for visible reddish orange luminescent device applications. *J. Alloys Compd.* 708, 911 (2017). https://doi.org/10.1016/j.jallcom.2017.03.020.
22. K. Swapna, Sk. Mahamuda, A. Srinivasa Rao, M. Jayasimhadri, S. Shakya, and G. Vijaya Prakash, $Tb^{3+}$ doped Zinc Alumino Bismuth Borate glasses for green emitting luminescent devices. *J. Lumin.* 156, 180 (2014). https://doi.org/10.1016/j.jlumin.2014.08.019.
23. A.V. Lalitha Phani, B. Srinivas, A. Hameed, M. Narasimha Chary, J.L. Rao, and M. Shareefuddin, Comparative studies on physical and spectroscopic properties of alumino bismuth borate glasses containing Pb, Zn & Cd ions. *Chin. J. Phys.* 58, 303 (2019). https://doi.org/10.1016/j.cjph.2018.12.021.
24. P. Rekha Rani, M. Venkateswarlu, K. Swapna, Sk. Mahamuda, R.A. Talewar, Ch.B. Annapurna-Devi, and A.S. Rao, NIR photoluminescence studies of $Nd^{3+}$ doped $B_2O_3$–$BaF_2$–$PbF_2$–$Al_2O_3$ glasses for 1063 μm laser applications. *J. Lumin.* 229, 117701 (2021). https://doi.org/10.1016/j.jlumin.2020.117701.
25. G. Hu, Y. Zhou, R. Liu, C. Li, L. Liu, H. Lin, F. Zeng, and Su. Zhongmin, Regulation of luminescence properties of SBGNA:$Eu^{3+}$ glass by the content of $B_2O_3$ and $Al_2O_3$. *Opt. Mater.* 106, 110025 (2020). https://doi.org/10.1016/j.optmat.2020.110025.
26. L. Shamshad, G. Rooh, K. Kirdsiri, N. Srisittipokakun, B. Damdee, H.J. Kim, and J. Kaewkhao, Effect of alkaline earth oxides on the physical and spectroscopic properties of $Dy^{3+}$- doped $Li_2O$-$B_2O_3$ glasses for white emitting material application. *Opt. Mater.* 64, 268 (2017). https://doi.org/10.1016/j.optmat.2016.12.027.
27. K. Siva Rama Krishna Reddy, K. Swapna, M. Venkateswarlu, Sk. Mahamuda, and A.S. Rao, Thermal, up-conversion and near-infrared luminescence studies of erbium ions doped alkaline-earth Boro Tellurite glasses. *Solid State Sci.* 97, 106016 (2019). https://doi.org/10.1016/j.solidstatesciences.2019.106016.
28. S. Banijamali and T. Ebadzadeh, Glass-forming ability, sintercrystallization behavior and microwave dielectric properties of MgO-$B_2O_3$-($Al_2O_3$)-$SiO_2$ glass-ceramics. *J. Non-Cryst. Solids* 441, 34 (2016). https://doi.org/10.1016/j.jnoncrysol.2016.03.014.
29. J. Kaewkhao, N. Wantana, S. Kaewjaeng, S. Kothan, and H.J. Kim, Luminescence characteristics of $Dy^{3+}$ doped $Gd_2O_3$-CaO-$SiO_2$-$B_2O_3$ scintillating glasses. *J. Rare Earths* 34, 583 (2016). https://doi.org/10.1016/S1002-0721(16)60065-0.
30. Sk. Mahamuda, F. Syed, Ch.B. Annapurna Devi, K. Swapna, M.V.V.K.S. Prasad, M. Venkateswarlu, and A.S. Rao, Spectral characterization of $Dy^{3+}$ ions doped phosphate glasses for yellow laser applications. *J. Non-Crystal. Solids* 555, 120538 (2021). https://doi.org/10.1016/j.jnoncrysol.2020.120538.
31. K. Siva Rama Krishna Reddy, K. Swapna, S. Mahamuda, M. Venkateswarlu, A.S. Rao, and G. Vijaya Prakash, Investigation on structural and luminescence features of $Dy^{3+}$ ions doped alkaline-earth boro tellurite glasses for optoelectronic devices. *Opt. Mater.* 85, 200 (2018). https://doi.org/10.1016/j.optmat.2018.08.057.
32. C. Sumalatha, R. Doddoji, M. Venkateswarlu, P.R. Rani, K. Swapna, Sk. Mahamuda, and A.S. Rao, White light emission from $Dy^{3+}$-doped ZnO + $Bi_2O_3$ + $BaF_2$ + $B_2O_3$ + $TeO_2$ glasses: Structural and spectroscopic properties. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 240, 118568 (2020). https://doi.org/10.1016/j.saa.2020.118568.
33. A.S. Rao, Y.N. Ahammed, R.R. Reddy, and T.V.R. Rao, Spectroscopic studies of $Nd^{3+}$ -doped alkali fluoroborophosphate glasses. *Opt. Mater.* 10, 245 (1998).
34. Y.B. Saddeek, M.A. Kaid, and M.R. Ebeid, FT-IR and physical features of $Al_2O_3$- $La_2O_3$-$P_2O_5$-PbO glasses. *J. Non-Cryst. Solids* 387, 30 (2014). https://doi.org/10.1016/j.jnoncrysol.2013.12.029.
35. R.S.E.S. Dawaud, S. Hashim, Y.S.M. Alajerami, M.H.A. Mhareb, and N. Tamchek, Optical and structural properties of lithium sodium borate glasses doped $Dy^{3+}$ ions. *J. Mol. Struct.* 1075, 113 (2014). https://doi.org/10.1016/j.molstruc.2014.06.032.
36. C. Gautam, A.K. Yadav, and A.K. Singh, A Reviewon infrared spectroscopy of borate glasses with effects of different additives. *ISRN Ceram.* 2012, 1 (2012). https://doi.org/10.5402/2012/428497.
37. W.T. Carnall, P.R. Fields, and K. Rajnak, Spectral intensities of the trivalent lanthanides and actinides in solution. II. $Pm^{3+}$, $Sm^{3+}$, $Eu^{3+}$, $Gd^{3+}$, $Tb^{3+}$, $Dy^{3+}$, and $Ho^{3+}$. *J. Chem. Phys.* 49, 4424 (1968). https://doi.org/10.1063/1.1669892.
38. R. Manjeet, K.P. Amit, N. Deopa, A. Kumar, and R.P. Chahal, Optimization of dysprosium ions doped borate glasses for photoluminescence applications. *Mater. Lett. X* 19(2023), 100208 (2023). https://doi.org/10.1016/j.mlblux.2023.100208.
39. B.R. Judd, Optical absorption intensities of rare-earth ions. *Phys. Rev.* 127, 750 (1962). https://doi.org/10.1103/PhysRev.127.750.

40. G.S. Ofelt, Intensities of crystal spectra of rare-earth ions. *J. Chem. Phys.* 37, 511 (1962). https://doi.org/10.1063/1.1701366.
41. A. Ichoja, S. Hashim, S.K. Ghoshal, I.H. Hashim, and R.S. Omar, Physical, structural and optical studies on magnesium borate glasses doped with dysprosium ion. *J. Rare Earths* 36, 1264 (2018). https://doi.org/10.1016/j.jre.2018.0.
42. S. Karki, C.R. Kesavulu, H.J. Kim, J. Kaewkhao, N. Chanthim, S. Kothan, and S. Kaewjaeng, Physical, optical and luminescence properties of the $Dy^{3+}$doped barium borophosphate glasses. *J. Non-Cryst. Solids* 521, 119483 (2019). https://doi.org/10.1016/j.jnoncrysol.2019.119483.
43. K.V. Krishnaiah, K.U. Kumar, and C.K. Jayasankar, Spectroscopic properties of $Dy^{3+}$ doped oxyfluoride glasses for white light emitting diodes. *Mater. Express* 3, 61 (2013). https://doi.org/10.1166/mex.2013.1094.
44. C.R. Kesavulu, H.J. Kim, S.W. Lee, J. Kaewkhao, N. Chanthima, and Y. Tariwong, Physical, vibrational, optical and luminescence investigations of $Dy^{3+}$-doped yttrium calcium silicoborate glasses for cool white LED applications. *J. Alloys Compd.* 726, 1062 (2017). https://doi.org/10.1016/j.jallcom.2017.08.091.
45. H. George, N. Deopa, S. Kaur, A. Prasad, M. Sreenivasulu, R. Jayasimhadri, and A.S. Rao, Judd-Ofelt parametrization and radiative analysis of $Dy^{3+}$ ions doped Sodium Bismuth Strontium Phosphate glasses. *J. Lumin.* 215, 116693 (2019). https://doi.org/10.1016/j.jlumin.2019.116693.
46. H. Ebendorff-Heidepriem and D. Ehrt, Effect of glass composition on Judd-Ofelt parameters and radiative decay rates of $Er^{3+}$ in fluoride phosphate and phosphate glasses. *Glass Sci. Technol.* 240, 289 (1998). https://doi.org/10.1016/S0022-3093(98)00706-6.
47. I.Z. Hager, Optical properties of lithium barium haloborate glasses. *J. Phys. Chem. Solid* 70, 210 (2009). https://doi.org/10.1016/j.jpcs.2008.10.007.
48. S.F. Hathot, B.M. Al Dabbagh, and H. Aboud, Structural and spectroscopic correlation in barium-boro-tellurite glass hosts: effects of $Dy_2O_3$ doping. *Chalcogenide Lett.* 21, 201 (2024). https://doi.org/10.15251/CL.2024.212.201.
49. M. Behera and R. Naik, Optical properties change in laser-induced $Te/As_2Se_3$ chalcogenide thin films. *Appl. Phys. A* 122, 913 (2016). https://doi.org/10.1007/s00339-016-0451-7.
50. S. Giri, P. Priyadarshini, D. Alagarasan, R. Ganesan, and R. Naik, Annealing-induced phase transformation in In10Se70Te20 thin films and its structural, optical and morphological changes for optoelectronic applications. *RSC Adv.* 13(26), 24955 (2023). https://doi.org/10.1039/D3RA03731B.
51. S. Das, S. Supriya, D. Alagarasan, R. Ganesan, and R. Naik, Investigating the temperature-dependent Raman spectroscopy of $Se/Bi_2Te_3$ thin films and its enhanced photoresponse for optoelectronic applications. *J. Appl. Phys.* 136, 065302 (2024). https://doi.org/10.1063/5.0216795.
52. S. Supriya, P.C. Kumar, M. Pradhan, and R. Naik, Hydrothermally synthesized bi-doped CdTe nanoparticles with tunable optical and photoresponse properties for visible photodetectors. *ACS Appl. Mater. Interfaces* 23, 33806 (2024). https://doi.org/10.1021/acsami.4c06201.
53. P.C. Kumar, G.K. Pradhan, S. Senapati, and R. Naik, Two-dimensional $Bi_2O_5Te$ nanosheets for sensitive and fast response broadband visible light photodetectors. *ACS Appl. Electron. Mater.* 6, 3311 (2024). https://doi.org/10.1021/acsaelm.4c00179.
54. R. Sharma and A.S. Rao, Photoluminescence study of $Sm^{3+}$ doped Zinc Lead Tungsten Tellurite glasses for reddish-orange photonic device applications. *J. Optmat.* 84, 375 (2018). https://doi.org/10.1016/j.optmat.2018.07.035.
55. P. You, G. Yin, X. Chen, Bo. Yue, Z. Huang, X. Liao, and Y. Yao, Luminescence properties of $Dy^{3+}$-doped $Li_2SrSiO_4$ for NUV-excited white LEDs. *Opt. Mater.* 33, 1808 (2011). https://doi.org/10.1016/j.optmat.2011.06.018.
56. D.V.R. Murthy, B.C. Jamalaiah, T. Sasikala, L. Rama Moorthy, M. Jayasimhadri, K. Jang, H.S. Lee, S. SooYi, and J.H. Jeong, Optical absorption and emission characteristics of $Pr^{3+}$doped RTP glasses. *Physica B* 405, 1095 (2010). https://doi.org/10.1016/j.physb.2009.11.012.
57. M. Liao, Z. Duan, L. Hu, Y. Fang, and L. Wen, Spectroscopic properties of $Er^{3+}/Yb^{3+}$ codoped fluorophosphate glasses. *J. Lumin.* 126, 139 (2007). https://doi.org/10.1016/j.jlumin.2006.06.009.
58. G. Lakshminarayana, K.M. Kaky, S.O. Baki, S. Ye, A. Lira, I.V. Kityk, and M.A. Mahdi, Concentration dependent structural, thermal and optical features of $Pr^{3+}$- doped multi component tellurite glasses. *J. Alloys Compd.* 686, 769 (2016). https://doi.org/10.1016/j.jallcom.2016.06.069.
59. N. Deopa, S. Saini, S. Kaur, A. Prasad, and A.S. Rao, Spectroscopic investigationson $Dy^{3+}$ions doped zinc lead alumino borate glasses for photonic device application. *J. Rare Earths* 37, 52 (2018). https://doi.org/10.1016/j.jre.2018.04.013.
60. P. Suthanthirakumar and K. Marimuthu, Investigations on spectroscopic propertiesof $Dy^{3+}$ doped zinc telluro-fluoroborate glasses for laser and white LED applications. *J. Mol. Struct.* 1125, 443 (2016). https://doi.org/10.1016/j.molstruc.2016.06.080.
61. M. Vijayakumar and K. Marimuthu, Structural and luminescence properties of $Dy^{3+}$ doped oxyfluoro-borophosphate glasses for lasing materials and white LEDs. *J. Alloys Compd.* 629, 230 (2015). https://doi.org/10.1016/j.jallcom.2014.12.214.
62. S. Karki, C.R. Kesavulu, H.J. Kim, J. Kaewkhao, N. Chanthima, S. Kothan, and S. Kaewjaeng, Physical, optical and luminescence properties of the $Dy^{3+}$doped barium borophosphate glasses. *J. Non-Cryst. Solids* 521, 119483 (2019). https://doi.org/10.1016/j.jnoncrysol.2019.119483.


**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.

2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3315-2169-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/IATMSI64286.2025.10984992

# Sum Rate Maximization and Performance Analysis of NOMA-Based IRS-Aided Uplink Visible Light Communication System

Vishal Kumar
*Delhi Technological University, India*
vishalkumar_2k21phdec07@dtu.ac.in

Rohit Kumar
*Delhi Technological University, India*
rohit.kumar@dtu.ac.in

Shankar Prakriya
*Indian Institute of Technology, Delhi, India*
shankar@ee.iitd.ac.in

***Abstract*—Achieving high data rates in uplink visible light communication (UVLC) systems while ensuring the safety of human eyes presents a significant challenge. This study investigates a UVLC system leveraging non-orthogonal multiple access (NOMA) and intelligent reflecting surfaces (IRS) to enhance the sum rate. The rate maximization problem is formulated based on the IRS attribute association matrix, and an optimization framework is developed to solve it. The proposed solution is validated through comprehensive simulations, demonstrating its effectiveness in improving system performance. Moreover, the impact of variation in transmit power and field of view on the sum rate with different numbers of IRS units is extensively studied. And the results demonstrate the effectiveness of the IRS over a system without the IRS.**

***Index Terms*—Uplink visible light communication, optimization, intelligent reflecting surfaces, sum rate, non-orthogonal multiple access (NOMA).**

## I. Introduction

THE rapid expansion of diverse internet-driven activities has fueled an unprecedented demand for unlimited data in wireless communication. This increasing reliance on connectivity necessitates access to broader spectrum resources and significantly higher data rates. However, the traditional radio frequency (RF) spectrum, which has been the foundation of communication systems for decades, is becoming increasingly saturated, posing challenges to meeting these escalating requirements.

The optimization of intelligent reflecting surface (IRS) configurations has been extensively explored to improve system performance [1], [2]. In visible light communication (VLC) systems, the susceptibility of light to blockages makes the IRS's ability to leverage reflected signals essential in overcoming these limitations. However, the role of IRS in VLC systems differs markedly from its application in RF communication systems due to the unique properties of visible light, including its distinct propagation behavior and non-negative, real-valued signal characteristics [3].

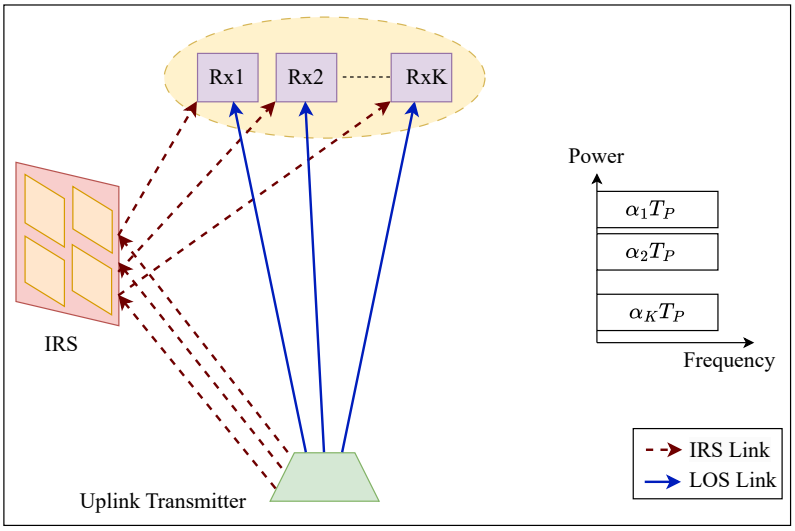
979-8-3315-2169-1/25/$31.00 ©2025 IEEE

To ensure safety in uplink visible light communication (VLC) systems, the optical power must be carefully controlled to prevent harmful irradiance to the human eye [4]. In contrast, downlink VLC systems typically utilize ceiling-mounted optical sources (LEDs), which are positioned far enough from human eyes to minimize visual discomfort. However, in uplink VLC, optical sources are often located much closer to the eyes, making the issue of irradiance more critical. While VLC is well-suited for delivering moderate data rates, achieving high data rates in the uplink is particularly challenging due to limitations in transmit power and the sensitivity of photodetectors. Reduced transmission power often results in weaker signals at the photodetector, lowering the signal-to-noise ratio (SNR). To address these challenges, researchers have proposed bidirectional systems combining VLC for downlink and RF for uplink communication [5]. However, such hybrid systems are unsuitable in environments like hospitals or aircraft, where RF usage is restricted. An alternative approach involves integrating infrared communication systems, as explored in [6]. Another promising solution is the development of low-power uplink VLC systems, which take advantage of the existing LED infrastructure while optimizing the inherent benefits of VLC technology.

While numerous non-orthogonal multiple access (NOMA) techniques have been developed as pivotal solutions for massive machine-type communication (mMTC) across various applications, this study emphasizes the commonly implemented power-domain NOMA. This method allows multiple users to simultaneously access the same spectrum by employing superposition coding (SC) at the transmitter, where transmission power is distributed according to the specific channel conditions of each user. At the receiver, users decode their signals using successive interference cancellation (SIC) [7]. Research has demonstrated that incorporating NOMA can substantially enhance the capacity of visible light communication (VLC) systems [8], [9].

In visible light communication (VLC) systems, non-line-of-sight (NLOS) signals often undergo significant attenuation due to reflections and diffractions caused by surfaces and obsta-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:02:27 UTC from IEEE Xplore. Restrictions apply.




(a)

Fig. 1: (a) System model for single transmitter multi-receiver uplink VLC system with NOMA

cles in indoor environments. Consequently, the contribution of NLOS paths to the overall received signal is typically minimal and can often be disregarded. Intelligent reflecting surface (IRS) technology offers a transformative solution by enhancing NLOS signal strength, effectively redirecting and focusing scattered light toward the receiver. Motivated by the potential of IRS, this study investigates its application in uplink lightwave communication to optimize system performance and address these challenges.

A genetic algorithm (GA) is employed in [10] to optimize the decoding order in an IRS-assisted NOMA VLC system. However, the study in [10] focuses explicitly on a downlink VLC scenario using a multiple-input single-output (MISO) configuration. Motivated by this, GA is utilized in this work to maximize the sum rate by putting constraints on total transmit power to prevent human eye irradiance. Our focus in this work is on using the state-of-the-art technology of the IRS and applying optimization to examine and analyze the NOMA-based IRS-aided UVLC system. To the extent of our understanding, this work is the first of its kind. The notable contributions of our work include:

- A NOMA-based on power domain IRS-aided single transmitter and multi-receiver uplink VLC system is proposed.
- The optimization problem is framed and analytically solved to maximize the sum rate by putting constraints on total transmit power, enabling high data rate uplink transmission while adhering to eye safety standards.
- To validate the analytically solved optimization problem, GA is utilized.
- The impact of variation in transmit power and field of view on the sum rate with different numbers of IRS units are analyzed.

## II. System and Channel Model

A single transmitter multi-receiver NOMA-based IRS-aided UVLC system is proposed in Fig. 1 (a), where $K$ receivers are served by a single transmitter aided by $N$ IRS units fixed on a wall of the room.

### A. VLC LOS Link

In general, channel gain of the direct LOS follows the Lambertian model in the VLC system [11], so that the expression of the gain from the transmitter to the $k$-th receiver can be expressed as

$$h_k^{los} = \frac{A(m+1)R_p}{2\pi d_k^2}\cos^m(\theta_k)U(\beta_k)G(\beta_k)\cos(\beta_k) \quad (1)$$

where, $U(\beta_k)$, $G(\beta_k)$ are optical filter gain and optical concentrator gain for $k$-th receiver respectively, $A$ is the physical area of the PD, $R_p$ is responsivity of the PD, $d_k$ is the euclidean distance between $k$-th receiver and the transmitter. The angle of illumination and incidence angle for the LOS path from $k$-th receiver are denoted by $\theta_k$ and $\beta_k$, respectively (with $h_k^{los} = 0$ for $\beta_k > \psi_c$). Moreover, the field of view (FOV) at the concentrator's half-angle is denoted by $\psi_c$. Additionally, $\phi_{1/2}$ denotes the LED's half power semi-angle, while $m$, the Lambertian emission order, is determined as, $m = \frac{-\ln 2}{\ln(\cos(\phi_{1/2}))}$. With $f$ as an internal refractive index, the optical concentrator gain is given by [11],

$$G(\beta_k) = \begin{cases} \frac{f^2}{\sin^2(\psi_c)} & \text{for } 0 \le \beta_k \le \psi_c \\ 0 & \text{for } \beta_k > \psi_c \end{cases} \quad (2)$$

### B. IRS Reflected Link

The reflected signals in the IRS-aided VLC system include diffuse and specular reflections. However, the impact of diffuse

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:02:27 UTC from IEEE Xplore. Restrictions apply.

reflection can be ignored due to its low level compared to the LOS channel gain [12]. For specular reflection, various methods exist for determining channel gain through estimation [13]. An approximate expression for the channel gain reflected by the IRS can be used under the point assumption, treating the specular IRS reflection as a signal emitted from an image source, with a coefficient factor $\rho_{irs}$ representing the energy loss due to reflection [3]. Additionally, reference [3] demonstrates that the formula adheres to an "additive" model, allowing the calculation of the IRS-reflected gain between the transmitter and the $k$-th receiver via the $n$-th IRS unit.

$$h_k^{irs,n} = \frac{A(m+1)R_p}{2\pi(d_{t,n}+d_{n,k})^2}\rho_{irs}\cos^m(\theta_k^n)U(\beta_k)G(\beta_k)\cos(\beta_k^n) \tag{3}$$

where $d_{t,n}$ and $d_{n,k}$ represents the separation between the transmitter and the $n$-th IRS and between the $n$-th IRS and the $k$-th receiver's PD respectively. Then, $\theta_k^n$ and $\beta_k^n$ denote the angles of irradiance and incidence of the IRS-reflected path, respectively. Moreover, a matrix $\boldsymbol{H_k^{irs}} = [h_k^{irs,1}, h_k^{irs,2}, \ldots, h_k^{irs,N}]^T \in \mathbb{R}_+^{N\times 1}$ is defined for simplicity.

### C. Combined Link

In the proposed IRS-assisted UVLC system, the receivers are sparsely distributed in space, and the specular reflection direction is highly dependent on the geometric configuration of the IRS units [1]. Consequently, each IRS unit is capable of serving only one receiver, eliminating interference between specular IRS-reflected paths under the point-source assumption. Building on this characteristic, a discrete matrix $\boldsymbol{A} = [a_1, a_2, \ldots, s_K] \in \mathbb{R}_+^{N\times K}$ is introduced to represent the assignment of IRS units to receivers. Each element $a_{n,k}$ is a binary variable, where $a_{n,k} = 1$ indicates that the n-th IRS unit is assigned to the k-th receiver, and $a_{n,k} = 0$ indicates otherwise.

Thus, the combined link's channel gain for $k$-th receiver is expressed as

$$h_k = h_k^{los} + (\boldsymbol{H_k^{irs}})^T a_k \tag{4}$$

where $(.)^T$ denotes transpose of the matrix.

## III. Problem Formulation and Performance Analysis

### A. NOMA-based UVLC system with IRS aided

In the proposed system, the transmitter operates over the same bandwidth to simultaneously serve all receivers. Adhering to the NOMA principle, transmission power is allocated based on the channel conditions of each receiver, which in turn determines the decoding order. However, the assignment of IRS units can alter the decoding sequence of the receivers. To address this, the indices of the receivers are reordered according to their combined channel quality, ensuring that $h_1 \geq h_2 \geq \cdots \geq h_K$. After ranking the receivers by channel quality, the allocated transmission power must satisfy $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_K$ and $\sum_{i=1}^K \alpha_i = 1$. For simplicity, it is assumed that each receiver is allocated a fixed transmission power. Under these conditions, considering $T_P$ as total transmit power, $\alpha_i$ power allocation coefficient of the i-th receiver, the received signal at the k-th receiver can be expressed as follows in (5).

$$y_k = h_k \sum_{i=1}^{K} \sqrt{\alpha_i T_P} x_i + \omega_k \tag{5}$$

where $x_i$ denotes the transmitted symbol and $\omega_k$ is Gaussian noise with zero mean and variance $\sigma$ at the k-th receiver.

### B. Sum Rate Maximization Problem

In the VLC systems, the traditional Shannon capacity formula is not directly applicable due to the distinct properties of visible light compared to RF signals. Instead, a more practical approach involves utilizing a tight lower bound on the achievable rate, which is commonly employed in dimmable VLC systems to account for these unique characteristics [14]. Considering $B$ as bandwidth and $\gamma_k$ denoting signal to interference plus noise ratio (SINR) of the k-th receiver, the rate of k-th receiver is expressed as

$$R_k = \frac{B}{2}\log_2\left(1 + \frac{e}{2\pi}\gamma_k\right) \tag{6}$$

where $e = \exp(1)$. With successive interference cancellation (SIC) implemented at the receiver, signals with higher power are decoded first, while lower-power signals are regarded as noise during the process. Based on the allocated transmission power, decoding proceeds in reverse order, starting from the last user and progressing to the first. Moreover, with $\rho = T_P/\sigma^2$ SINR of the k-th receiver is given as

$$\gamma_k = \begin{cases} \frac{h_k^2\alpha_k}{\frac{1}{\rho}+h_k^2\sum_i=1^{k-1}\alpha_i}, & \text{for } 2 \leq k \leq K \\ \rho h_1^2\alpha_1, & \text{for } k = 1 \end{cases} \tag{7}$$

In order to simplify the discussion, sets $\mathcal{K}$ and $\mathcal{N}$ are defined to represent the index of receivers and IRS units, respectively. To maximize the overall sum rate, the optimization problem is expressed as

$$\mathbf{P}: \max_{\boldsymbol{A}} R_{sum} = \max_{\boldsymbol{A}} \sum_{k}^{K} R_k \tag{8}$$

$$\text{s.t. } R_k \geq \tilde{R} \tag{9}$$

$$\sum_{k=1}^{K} a_{n,k} = 1, \forall k \in \mathcal{K} \tag{10}$$

$$a_{n,k} \in \{0,1\}, \forall n \in \mathcal{N} \tag{11}$$

The objective in (8) is to maximize the overall sum rate, and the restriction in (9) is to guarantee that the achievable rate by each receiver must be greater than a certain threshold value $\tilde{R}$ and (10) and (11) comes from the association matrix $\boldsymbol{A}$.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:02:27 UTC from IEEE Xplore. Restrictions apply.

The objective function in **P** is non-convex due to the term $\gamma$, which contains a quadratic term within a logarithmic function. Therefore, we used an iterative optimization algorithm to turn this non-convex problem into a convex problem [15]. After using this we denote $a_k^{(t)}$ association vector at $t$-th iteration. Then, $a_k^{(t-1)}$ is utilized to turn the **P** in $t$-th iteration into convex problem. That results in modified expression for SINR as

$$\gamma_k = \begin{cases} \frac{(\mathcal{D})\alpha_k}{\frac{1}{\rho}+(h_k^{los}+(H_k^{irs})^T a_k^{(t-1)})^2 \sum_{i=1}^{k-1}\alpha_i}, & \text{for } 2 \le k \le K \\ \rho \mathcal{D}_1 \alpha_1, & \text{for } k = 1 \end{cases} \tag{12}$$

where $\mathcal{D} = (h_k^{los})^2 + 2h_k^{los}(\boldsymbol{H_k^{irs}})^T a_k^{(t)} + (\boldsymbol{H_k^{irs}})^T a_k^{(t-1)} (\boldsymbol{H_k^{irs}})^T a_k^{(t)}$ and $\mathcal{D}_1 = (h_1^{los})^2 + 2h_1^{los}(\boldsymbol{H_1^{irs}})^T a_1^{(t)} + (\boldsymbol{H_1^{irs}})^T a_1^{(t-1)} (\boldsymbol{H_1^{irs}})^T a_1^{(t)}$. Now, the Lagrangian function of **P** after turning it into a convex problem with $\bar{\lambda}$ as Lagrangian multiplier

$$\mathcal{L}(a_{n_1,k}, a_{n_2,k}, \bar{\lambda}) = -\sum_{k}^{K} R_k + \bar{\lambda}(R_k - \tilde{R}) \tag{13}$$

The Karush-Kuhn-Tucker (KKT) conditions of **P** are

$$\frac{\partial \mathcal{L}(a_{n_1,k}, a_{n_2,k}, \bar{\lambda})}{\partial \bar{\lambda}} = R_k - \tilde{R} = 0 \tag{14}$$

$$\begin{aligned} &\frac{\partial \mathcal{L}(a_{n_1,k}, a_{n_2,k}, \bar{\lambda})}{\partial a_{n_1,k}} \\ &= \frac{\frac{-e}{2\pi}\left(2\sum_{k=1}^{K}(h_k^{los} + (\boldsymbol{H_k^{irs}})^T)\alpha_k h_k^{irs,n1}\right)}{(\mathcal{E})(\mathcal{F})} \\ &+ \bar{\lambda}\frac{c_k^{(t)}(2h_k^{los} + (\boldsymbol{H_k^{irs}})^T a_k^{(t-1)})\alpha_k h_k^{irs,n1}}{\frac{1}{\rho} + (h_k^{los} + (\boldsymbol{H_k^{irs}})^T a_k^{(t-1)})^2 \sum_{i=1}^{k-1}\alpha_i} = 0 \end{aligned} \tag{15}$$

$$\begin{aligned} &\frac{\partial \mathcal{L}(a_{n_1,k}, a_{n_2,k}, \bar{\lambda})}{\partial a_{n_2,k}} \\ &= \frac{\frac{-e}{2\pi}\left(2\sum_{k=1}^{K}(h_k^{los} + (\boldsymbol{H_k^{irs}})^T)\alpha_k h_k^{irs,n2}\right)}{(\mathcal{E})(\mathcal{F})} \\ &+ \bar{\lambda}\frac{c_k^{(t)}(2h_k^{los} + (\boldsymbol{H_k^{irs}})^T a_k^{(t-1)})\alpha_k h_k^{irs,n2}}{\frac{1}{\rho} + (h_k^{los} + (\boldsymbol{H_k^{irs}})^T a_k^{(t-1)})^2 \sum_{i=1}^{k-1}\alpha_i} = 0 \end{aligned} \tag{16}$$

where

$$\mathcal{E} = \frac{K}{\rho} + \sum_{k=1}^{K}(h_k^{los} + (\boldsymbol{H_k^{irs}})^T)^2 \sum_{i=1}^{k-1}\alpha_i \tag{17}$$

$$\mathcal{F} = K + \frac{e}{2\pi}\sum_{k=1}^{K}\frac{h_k^{los} + 2h_k^{los}(\boldsymbol{H_k^{irs}})^T + ((H_k^{irs})^T)^2\alpha_k}{\frac{1}{\rho} + \sum_{k=1}^{K}(h_k^{los} + (\boldsymbol{H_k^{irs}})^T)^2 \sum_{i=1}^{k-1}\alpha_i} \tag{18}$$

After solving (14)-(16), we obtained optimal values for $\bar{\lambda}$, $a_{n1,k}^{(t)}$ and $a_{n2,k}^{(t)}$ in (19), (20) and (21) respectively.

$$\bar{\lambda}^* = \frac{\frac{e}{2\pi}\left(2\sum_{k=1}^{K}(h_k^{los} + (\boldsymbol{H_k^{irs}})^T)\alpha_k h_k^{irs,n1}\right)}{\frac{(2h_k^{los}+\boldsymbol{H_k^{irs}})^T a_k^{(t-1)})\alpha_k h_k^{irs,n1}(\mathcal{E})(\mathcal{F})}{\frac{1}{\rho}+(h_k^{los}+(\boldsymbol{H_k^{irs}})^T a_k^{(t-1)})^2 \sum_{i=1}^{k-1}\alpha_i}} \tag{19}$$

$$a_{n1,k}^{(t)\ *} = \frac{\mathcal{M} - h_k^{irs,n2}}{h_k^{irs,n1} - h_k^{irs,n2}} \tag{20}$$

$$a_{n2,k}^{(t)\ *} = 1 - a_{n1,k}^{(t)\ *} \tag{21}$$

where with $\mathcal{J} = h_k^{los} + h_k^{irs,n1} a_{n1,k}^{(t-1)} + h_k^{irs,n2} a_{n2,k}^{(t-1)}$, $\mathcal{M}$ is given as (22).

$$\mathcal{M} = \frac{\frac{2\pi}{e\alpha_k} 4^{(\tilde{R}/B-1)}\left(\frac{1}{\rho} + (\mathcal{J})^2 \sum_{i=1}^{k-1}\alpha_i\right) - (h_k^{los})^2}{2h_k^{los} + + h_k^{irs,n1} a_{n1,k}^{(t-1)} + + h_k^{irs,n2} a_{n2,k}^{(t-1)}} \tag{22}$$

## IV. Results and Analysis

Table I: Parameters used for simulation

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $R_p$ | 0.53 A/m | $\rho_{irs}$ | 0.95 |
| $A$ | 1 $cm^2$ | $B$ | 20 MHz |
| $U(.)$ | 1 | $f$ | 1.5 |
| $\psi_c$ | $85^o$ | $\phi_{1/2}$ | $70^o$ |
| $\eta$ | $10^{-21}$ $A^2/Hz$ | $\tilde{R}$ | 6 Mbps |

The numerical and simulation results are discussed in this section. The room is assumed to have dimensions of $(10 \times 10 \times 4)$ $m^3$, featuring a single uplink transmitter and multi-receivers. The IRS units are evenly spaced on the wall near the receivers, in a rectangle with two corners at (1m, 0m, 1.5m) and (7m, 0m, 2.5m). Bandwidth $B$ is considered as 20 MHz, and a minimum rate required, $\tilde{R}$ is 6 Mbps. Moreover, for simulation $K = 2$ is considered with power allocation factors $\alpha_1$ and $\alpha_2$ as 0.35 and 0.65, respectively. Additional simulation parameters, as outlined in [16, 17], are provided in Table I.

In Fig. 2, the Genetic algorithm (GA) is utilized to solve the optimization problem, and theoretical and algorithm-based simulation is verified. GAs are optimization techniques inspired by natural selection. They are used here to allocate resources in a way that maximizes the sum rate. GAs are especially effective in solving complex, non-linear optimization problems like those encountered in VLC systems, where factors like interference, power allocation, and channel conditions must be considered simultaneously. The iterative improvement seen in the GA indicates that it is effectively learning to optimize resource allocation in the VLC system, bringing its performance closer to the theoretical maximum. This suggests that GA can serve as a practical solution for resource management in VLC systems, particularly when achieving theoretical limits directly is infeasible. In another way, we can infer that the theoretical results match the simulation results well. Fig. 2 shows that around 22.5 Mbps of data is achieved in the proposed system

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:02:27 UTC from IEEE Xplore. Restrictions apply.

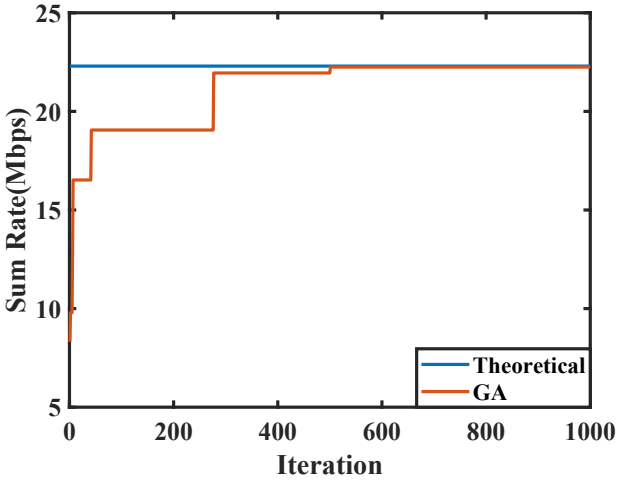



Fig. 2: Sum rate maximization characteristic curve

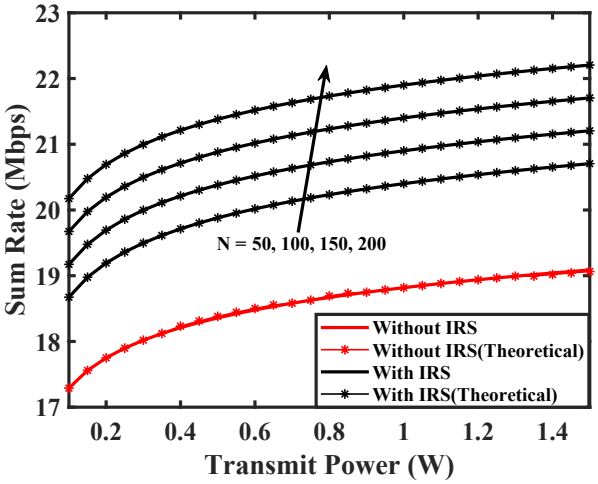



Fig. 3: Sum rate versus transmit power for different N

Fig. 3 illustrate the impact of integrating intelligent reflecting surfaces (IRS) into VLC systems on the achievable sum rate across varying transmit power levels. The inclusion of the IRS significantly enhances system performance compared to scenarios without the IRS, as seen by the higher sum rates achieved. Specifically, the presence of IRS enables better channel quality by reflecting and directing light toward the receivers, mitigating signal degradation and interference. As the transmit power increases, the gap between scenarios with and without the IRS widens, demonstrating the effectiveness of the IRS in efficiently utilizing power to enhance communication rates. This highlights the IRS as a transformative technology for improving the data capacity of VLC systems, particularly in environments where efficient resource allocation and signal optimization are critical. Furthermore, Fig. 3 also shows that a significant data rate is achieved with a lower power of 1.5 W with $N = 200$. Since the transmit power is lower thus, human eye irradiance problem in UVLC is also avoided.

Fig. 4 shows the influence of the field of view (FOV) on the sum rate performance of UVLC systems, comparing scenarios with and without the integration of intelligent reflecting surfaces (IRS). As the FOV increases, the sum rate decreases in both cases due to the widening of the receiver's light acceptance angle, which reduces signal intensity and increases interference from undesired reflections. However, systems with IRS consistently outperform those without IRS, demonstrating the capability of IRS to enhance signal directionality and mitigate interference. The practical performance trends closely follow the theoretical bounds, emphasizing the feasibility of implementing IRS in real-world VLC systems. Furthermore, the improvement in sum rate with IRS becomes more pronounced as N increases, indicating the scalability and adaptability of IRS technology. These findings underscore the potential of IRS to optimize system performance in VLC networks by effectively managing spatial constraints and enhancing signal quality across varying FOVs.

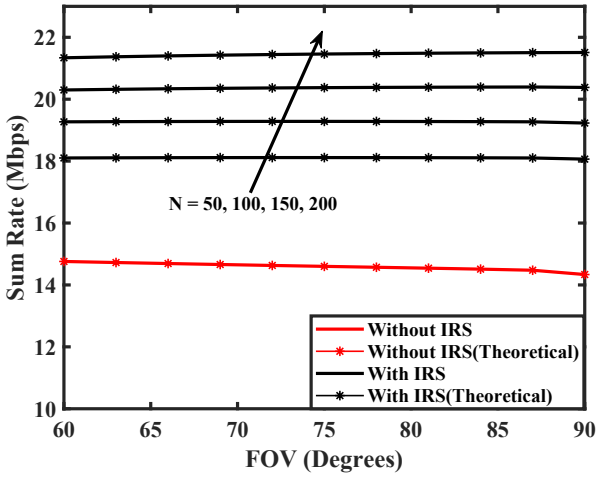



Fig. 4: Sum rate versus field of views of UVLC transmitter for different N

## V. Conclusion

This study addresses the critical challenge of achieving high data rates in uplink visible light communication (UVLC) systems while prioritizing human eye safety. By integrating non-orthogonal multiple access (NOMA) and intelligent reflecting surfaces (IRS), the proposed system demonstrates significant enhancements in sum rate performance. Through the formulation of an optimization framework based on the IRS attribute association matrix and its validation via comprehensive simulations, the solution proves highly effective in optimizing system performance. Furthermore, an in-depth analysis of the effects

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:02:27 UTC from IEEE Xplore. Restrictions apply.

of transmit power, field of view, and varying numbers of IRS units reveals the substantial advantages of incorporating IRS technology compared to systems without IRS. These findings underscore the potential of IRS-aided NOMA-UVLC systems as a promising approach for next-generation high-performance communication networks.

## REFERENCES


[1] S. Sun, F. Yang, and J. Song, "Sum rate maximization for intelligent reflecting surface-aided visible light communications," *IEEE Communications Letters*, vol. 25, no. 11, pp. 3619–3623, 2021.

[2] S. Sun *et al.*, "Joint resource management for intelligent reflecting surface–aided visible light communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 6508–6522, 2022.

[3] A. M. Abdelhady *et al.*, "Visible light communications via intelligent reflecting surfaces: Metasurfaces vs mirror arrays," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1–20, 2020.

[4] *BS EN 62471:2008, Photobiological Safety of Lamps and Lamp Systems, BSI British Standards Std., Sep.*, 2008.

[5] A. Jovicic, J. Li, and T. Richardson, "Visible light communication: opportunities, challenges and the path to market," *IEEE communications magazine*, vol. 51, no. 12, pp. 26–32, 2013.

[6] A. Burton *et al.*, "Experimental demonstration of a 10base-t ethernet visible light communications system using white phosphor light-emitting diodes," *IET Circuits, Devices & Systems*, vol. 8, no. 4, pp. 322–330, 2014.

[7] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users," *IEEE signal processing letters*, vol. 21, no. 12, pp. 1501–1505, 2014.

[8] L. Yin, W. O. Popoola, X. Wu, and H. Haas, "Performance evaluation of non-orthogonal multiple access in visible light communication," *IEEE Transactions on Communications*, vol. 64, no. 12, pp. 5162–5175, 2016.

[9] B. Cao, M. Chen, Z. Yang, M. Zhang, J. Zhao, and M. Chen, "Reflecting the light: Energy efficient visible light communication with reconfigurable intelligent surface," in *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*. IEEE, 2020, pp. 1–5.

[10] Z. Liu *et al.*, "Noma-based miso visible light communication systems with optical intelligent reflecting surface: Joint active and passive beamforming design," *IEEE Internet of Things Journal*, vol. 11, no. 10, pp. 18 753–18 767, 2024.

[11] X. Deng *et al.*, "Modeling and analysis of transmitter performance in visible light communications," *IEEE Tran. on Vehicular Tech.*, vol. 68, no. 3, pp. 2316–2331, 2019.

[12] M. Obeed *et al.*, "On optimizing vlc networks for downlink multi-user transmission: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2947–2976, 2019.

[13] H. Guo and V. K. Lau, "Uplink cascaded channel estimation for intelligent reflecting surface assisted multiuser miso systems," *IEEE Transactions on Signal Processing*, vol. 70, pp. 3964–3977, 2022.

[14] J.-B. Wang, Q.-S. Hu, J. Wang, M. Chen, and J.-Y. Wang, "Tight bounds on channel capacity for dimmable visible light communications," *Journal of Lightwave Technology*, vol. 31, no. 23, pp. 3771–3779, 2013.

[15] Z. Liu, F. Yang, S. Sun, J. Song, and Z. Han, "Sum rate maximization for noma-based vlc with optical intelligent reflecting surface," *IEEE Wireless Communications Letters*, vol. 12, no. 5, pp. 848–852, 2023.

[16] S. Aboagye, T. M. Ngatched, O. A. Dobre, and A. R. Ndjiongue, "Intelligent reflecting surface-aided indoor visible light communication systems," *IEEE Comm. Letters*, vol. 25, no. 12, pp. 3913–3917, 2021.

[17] V. Kumar, R. Kumar, and S. Prakriya, "Performance of an intelligent reflecting mirror-aided uplink lightwave communication system," *IEEE Wireless Communications Letters*, vol. 13, no. 4, pp. 954–958, 2024.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:02:27 UTC from IEEE Xplore. Restrictions apply.

Applied Nanoscience (2025) 15:19
https://doi.org/10.1007/s13204-025-03094-x

ORIGINAL ARTICLE

# Sustained release of epigallocatechin gallate from calcium carbonate nanoparticles: promising therapy for colorectal cancer

**Ritu[1] · Prakash Chandra[1] · Asmita Das[1]**

Received: 16 October 2024 / Accepted: 16 April 2025
© King Abdulaziz City for Science and Technology 2025

## Abstract

Nanotechnology has provided an enabling platform for innovations in site-specific cancer therapy and personalized oncomedicine. One of the most important advances in this area is the creation of nanomedicines that target cancer cells, which facilitates the progress of precise therapeutic interventions. Efficient drug delivery into tumor cells remains one of the critical challenges in cancer therapy. However, cancer-cell-targeted nanomedicines that function within the intricate milieu of the tumor microenvironment have shown potential to improve therapeutic efficiency. Epigallocatechin gallate (EGCG), the major phytochemical in *Phyllanthus emblica* (amla), is known for its anticancer and anti-inflammatory properties. In this study, we developed pH-responsive calcium carbonate nanoparticles (CCNPs) as a nanocarrier for EGCG to enhance its intracellular delivery and therapeutic efficacy. EGCG was physically adsorbed onto the surface of CCNPs. The synthesized nanoparticles were characterized using UV–visible spectroscopy, scanning electron microscopy (SEM), dynamic light scattering (DLS), zeta potential analysis, Fourier transform infrared spectroscopy (FTIR), and X-ray diffraction (XRD). In-vitro drug release studies demonstrated a sustained and more prominent release of EGCG at the acidic pH (4.8) typical of the tumor microenvironment. The biological activity of EGCG-loaded CCNPs was evaluated using the MTT assay and apoptosis analysis. The results showed significant cytotoxicity against colorectal cancer cells (COLO-320 DM), while blank CCNPs exhibited low toxicity and high biocompatibility. Intracellular uptake studies further confirmed the preferential accumulation of nanoparticles within colorectal cancer cells. Flow cytometry-based apoptosis assays revealed that EGCG-conjugated CCNPs induced considerable cancer cell death. These findings suggest that EGCG-loaded calcium carbonate nanoparticles may serve as an effective and biocompatible drug delivery platform for targeted colorectal cancer therapy.

**Keywords** Nanotechnology · Colorectal cancer · Calcium carbonate nanoparticles · Epigallocatechin gallate · Targeted drug delivery · Tumor microenvironment

## Introduction

Cancer represents a group of disorders resulting from uncontrolled cell division and the ability of these malignant cells to invade or metastasize to other parts of the body (Brown et al. 2023). Histologically, cancers are classified based on the presumed cell type of origin of the tumor (Baba and Catoi 2007): carcinoma, sarcoma, lymphoma, leukemia, germ cell tumor, blastoma, etc. Among them, carcinoma is cancer derived from epithelial cells and represents nearly all types of malignancies in the breast, prostate, lung, pancreas, and colon (Cancer Classification 2024; Sawicki et al. 2021). Colorectal cancer is the fourth most commonly occurring cancer in the world, which causes malignant neoplasmatic growth in the inner lining of the colon or the rectum (Knowlton et al. 2024; Douaiher et al. 2017). American Joint Committee on Cancer has classified CRC into five stages based on the development of the cancer cells. In stage 0, complete cure is permissible after surgical excision. For stages I–IIC, surgical resection is the standard modality of treatment, where the 5-year survival rate falls under a wide range, from 37% to 74%. The advanced stages are the ones that include stages IIIA–IV, when adjuvant chemotherapy is provided

✉ Asmita Das
asmitadas1710@dce.ac.in

Ritu
ritu_2k19phdbt13@dtu.ac.in

Prakash Chandra
prakash011@gmail.com; prakashchandra@dtu.ac.in

[1] Department of Biotechnology, Delhi Technological University, Main Bawana Road, Delhi 110042, India

Published online: 15 May 2025

subsequent to surgical resection. In such patients, the risk of metastasis and recurrence slashes the survival rate down to 6% (Edge 2010; Puppa et al. 2010). Colorectal cancer is a very heterogeneous disease at the cellular level, and these heterogeneities have to be formulated against a backdrop of multiple genetic and metabolic alterations (Blanco-Calvo et al. 2015; Burrell et al. 2013). However, once formed, metastasis can be further spread through the circulatory and lymphatic systems to distant organs, hence greatly increasing difficulties in therapies and raising mortality rates (Testa et al. 2018; Dukes and Bussey 1958). Although chemotherapy and radiotherapy are influential treatment approaches that can enhance survival in patients with cancer, sometimes they turn out to be acutely, severely, and long-termly adverse to the healthy organs of the patient (Anand et al. 2023). For example, irinotecan is an FDA-approved chemotherapeutic drug used in the treatment of colorectal cancer, and when used for a long period of time, it resulted in cholinergic syndrome (Ritu and Das 2023; Fujita et al. 2015).

Despite the progress in diagnosis and treatment, cancer still looms large in human health, and mortality rates remain as poor as ever (Siegel et al. 2023). The more traditional approaches in treating cancers—in terms of surgery, radiation, and chemotherapy—hold some hopeful promise but are harmfully laden with the limitation of systemic toxicity, damage to healthy tissues, and limited efficacy in the advanced stages of the disease. Moreover, multidrug resistance occurs due to the overexpression of a number of proteins in the tumor cells, which acts as a very stringent barrier. In this condition, chemotherapy is found to be highly reduced (Ashique et al. 2024). While quite promising, immunotherapies have only reported success rates of 25%–30% against advanced cancers (Grumberg et al. 2023), thus proving the need for such treatments to become more effective and targeted (Debela et al. 2021). The natural products present a promising avenue of cancer therapy. In this regard, botanicals, particularly antioxidant-rich plants, are of great significance. Green tea contains many polyphenols, with catechins being the most abundant and effective antioxidants that exhibit antimutagenic and anticancer properties. Among these compounds, EGCG is the most prevalent and exhibits high efficacy in suppressing cancer development (Du et al. 2012; Min and Kwon 2014; Chen et al. 2020). We previously showed that EGCG enhances the anticancer activities of ginseng compounds in human colorectal cancer cells (Ritu and Das 2023). Delivery of natural compounds to targeted tumor sites is quite challenging for several reasons. These compounds can have poor solubility, hence their poor absorption and bioavailability in the circulatory system of the body. The compounds can also be biodegradable natural products that get degraded and eliminated before getting a chance to reach the target cancer site. Moreover, varying pH levels in the human body may result in loss of effectiveness and stability for natural compounds, thereby losing their therapeutic potency in the process of delivery. For this reason, efficient delivery systems that allow for efficient loading of loading of natural compounds away from such challenges need to be developed in order to improve their potential for cancer treatment.

Nanoparticles are one of the most promising tools in the field of nanobiotechnology, more particularly in drug delivery (Afzal et al. 2022). They could, with good reason, surpass the limitations that conventional cancer-solving methods and formulas face in the market by providing a greater deal of targeted, controlled, and long-term release of therapeutic molecules with minimal systemic side effects (Kashyap et al. 2023; Chehelgerdi et al. 2023; Patra et al. 2018).

Research investigations in progress are oriented to optimize this approach with the aim of minimizing the negative impacts of the conventional techniques. Some of the benefits which can be attained by utilizing NPs as vehicles for drug delivery include: i) overcoming the problems associated with the low solubility and bioavailability of the drug (Patra et al. 2018), ii) increasing permeability of the drug across cancer cell and controlling its release, iii) NPs are non-toxic, biodegradable, highly fluorescent particles of small size, 1–100 nm, on which the cancer drug is effortlessly loaded (Maurya et al. 2019).

With regard to nanoparticles explored in the field of nanobiotechnology, $CaCO_3$ stands out in terms of its biocompatibility, pH responsiveness, safety, and biodegradability (Liang et al. 2023). CCNPs are especially potential candidate for cancer therapy because they have the potential to induce the death of tumor cells, increase intracellular Ca2+ levels, and even cause mitochondrial damage (Xiao et al. 2023). They are normally engineered in a way that makes dissolution, entrapment, or covalent adsorption attach them to drugs, thereby releasing the drugs onto the cancer tissues with remarkable precision and minimizing exposure on healthy cellular levels (Ma and Yang 2016). The specific properties of CCNPs include very high loading capacity and biocompatibility, thereby making them very excellent choices for drug delivery systems (Bai et al. 2022). Figure 1.

The present study is related to CCNPs, the synthesis and characterization of which have been approached as carriers for the anticancer agent epigallocatechin gallate. In the present study, CCNPs are synthesized by the chemical precipitation method and characterized by various techniques such as SEM, XRD, FTIR, TEM, and DLS. The in vitro efficacy of EGCG-loaded CCNPs was evaluated against the colorectal cancer cell line for their potential as a drug delivery system. This study utilizes the pH-sensitivity of CCNPs to explain an overall, controlled, and long-term release of this drug compound that is itself endowed with the enhancement of the anticancer efficacy of EGCG, coupled with a

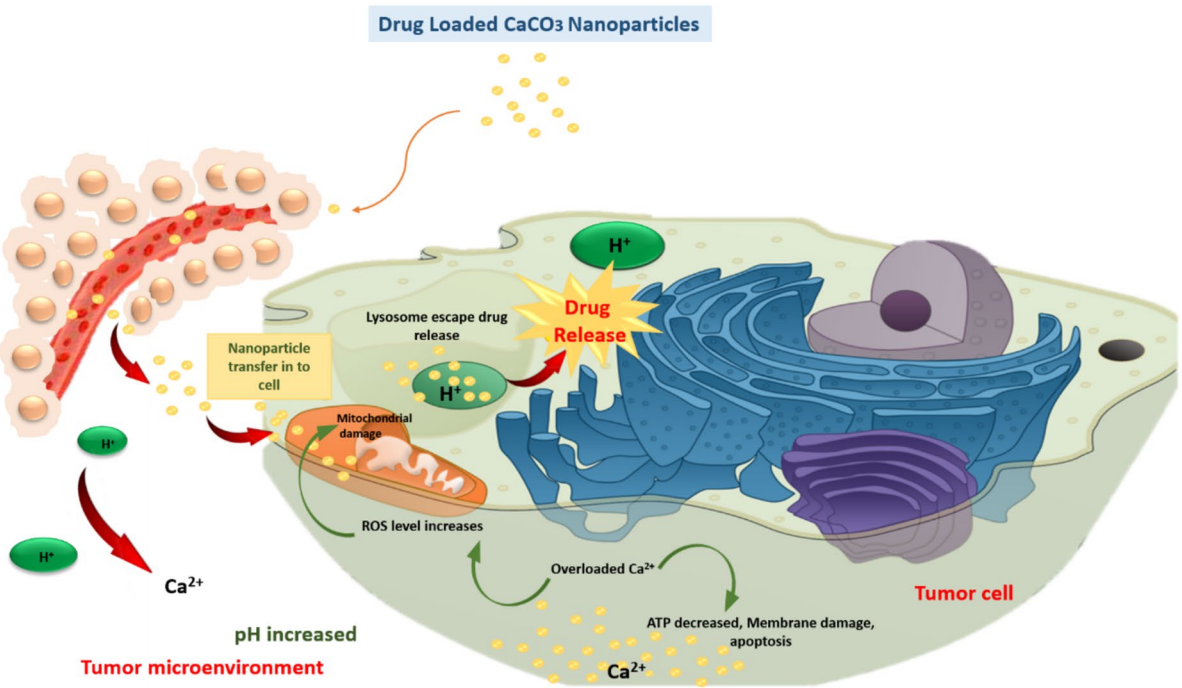



**Fig. 1** Schematic representation of drug-loaded calcium carbonate nanoparticles within cells, illustrating their role in various mechanisms of cell apoptosis. The nanoparticles facilitate targeted drug delivery and induce mitochondrial damage by increasing ROS levels, leading to cell apoptosis

reduction in its systemic toxicity. The outcome is expected to provide valuable insights into the development of efficient nanoparticle-based drug delivery systems for treating cancerous diseases and eventually open up new avenues of research and applicability in the field of nanomedicine.

Cancer comprises a diverse group of disorders characterized by uncontrolled cell proliferation and the capacity of malignant cells to invade surrounding tissues or metastasize to distant organs (Brown et al. 2023). Histologically, tumors are classified according to their presumed cell of origin (Baba and Catoi 2007): for example, carcinoma arises from epithelial cells and accounts for most malignancies of the breast, prostate, lung, pancreas, and colon (Cancer Classification 2024; Sawicki et al. 2021). Colorectal cancer (CRC) is the fourth most commonly diagnosed cancer worldwide, originating from neoplastic growth in the colonic or rectal mucosa (Knowlton et al. 2024; Douaiher et al. 2017). The American Joint Committee on Cancer (AJCC) stages CRC from 0 to IV based on tumor invasion and spread: stage 0 lesions can be cured by surgical excision alone, stages I–IIC rely primarily on surgery (5-year survival: 37–74%), whereas stages IIIA–IV require adjuvant chemotherapy and carry a 5-year survival rate as low as 6% due to high risks of metastasis and recurrence (Edge 2010; Puppa et al. 2010). CRC exhibits extensive intra- and intertumoral heterogeneity driven by genetic and metabolic alterations, complicating treatment strategies (Blanco-Calvo et al. 2015; Burrell et al. 2013). Once metastatic, CRC cells disseminate via the circulatory and lymphatic systems, further increasing therapeutic challenges and mortality (Testa et al. 2018; Dukes and Bussey 1958).

Despite advances in early detection and treatment, cancer remains a major health challenge, and overall mortality rates have shown little improvement (Siegel et al. 2023). Current cancer treatment strategies—surgery, radiation, and chemotherapy—are hampered by systemic toxicity, collateral damage to healthy tissues, and reduced efficacy in advanced disease. Multidrug resistance, often mediated by overexpression of efflux proteins (e.g., P-gp, MRP1), further diminishes chemotherapeutic outcomes (Ashique et al. 2024). For instance, irinotecan, a topoisomerase I inhibitor approved for CRC, also inhibits acetylcholinesterase, leading to cholinergic syndrome (symptoms: diarrhea, abdominal cramping, and excessive salivation) upon prolonged use (Ritu and Das 2023; Fujita et al. 2015). Immunotherapies, while promising, achieve objective response rates of only 25–30% in advanced cancers, underscoring the need for more effective, targeted modalities (Grumberg et al. 2023; Debela et al. 2021).

Natural products, particularly antioxidant-rich botanicals, offer a complementary avenue for cancer therapy. Epigallocatechin gallate (EGCG), the principal catechin in green tea,

exhibits potent anticancer and anti-inflammatory activities (Du et al. 2012; Min and Kwon 2014; Chen et al. 2020). In our previous work, EGCG synergized with ginseng-derived compounds to enhance apoptosis in CRC cells (Ritu and Das 2023). However, EGCG's clinical translation is limited by poor water solubility, low bioavailability, rapid metabolic degradation, and pH-dependent instability, which reduce its therapeutic potency (Afzal et al. 2022).

Nanoparticle-based drug delivery systems can address these limitations by improving solubility, protecting labile compounds, and enabling controlled, site-specific release (Kashyap et al. 2023; Chehelgerdi et al. 2023; Patra et al. 2018). Nanocarriers that overcome low drug solubility and bioavailability (Patra et al. 2018) enhancing cellular uptake and minimizing systemic side effects are under intensive investigation. Advantages of nanoparticle vehicles include their tunable size (1–100 nm), biodegradability, and capacity for surface functionalization to achieve targeted delivery (Maurya et al. 2019).

Among various nanocarriers, calcium carbonate nanoparticles ($CaCO_3$ NPs) stand out for their inherent biocompatibility, pH responsiveness, safety, and biodegradability (Liang et al. 2023). In acidic tumor microenvironments, $CaCO_3$ NPs dissolve, triggering site-specific drug release; they can also elevate intracellular $Ca^{2+}$ levels and induce mitochondrial dysfunction in cancer cells (Xiao et al. 2023). Moreover, $CaCO_3$ NPs offer high drug-loading capacity and can be engineered via dissolution–reprecipitation or adsorption to achieve efficient payload entrapment (Ma and Yang 2016; Bai et al. 2022).

In this study, we report the synthesis of pH-responsive $CaCO_3$ NPs by a chemical precipitation method and their comprehensive characterization by SEM, XRD, FTIR, TEM, and DLS. We physically adsorbed EGCG onto the nanoparticle surface and investigated pH-dependent release at pH 7.4 and 4.8 to mimic physiological and tumor-acidic conditions. The in vitro anticancer efficacy of EGCG-loaded $CaCO_3$ NPs was evaluated in CRC cell lines via cytotoxicity (MTT), apoptosis assays, and cellular uptake studies. Our goal is to elucidate the potential of $CaCO_3$ NPs as an efficient, biocompatible platform for targeted colorectal cancer therapy and to lay the groundwork for future translational investigations.

## Methodology

### Materials and methods

Calcium chloride, sodium bicarbonate salts, fetal bovine serum, RPMI-1640, and trypsin ethanol were purchased from Sigma-Aldrich. Glasswares for cell culture and for nanoparticle synthesis were purchased from Tarson pvt limited and NEST. Colorectal cancer cell lines (COLO-320 DM) were obtained from National Centre for Cell Science (NCCS) Pune. The compound 3-(4,5-Dimethyl-2-thiazolyl)−2,5-diphenyl-2-H-tetrazolium bromide (MTT) was acquired from Krishgen Biosystems Pvt Ltd. All the remaining reagents were of high purity and were acquired from commercially accessible sources.

### Preparation of calcium carbonate nanoparticles

In the present work, nano-sized calcium carbonate has been synthesized through chemical precipitation. For this process, the aqueous solutions of sodium carbonate and calcium chloride were mixed and immediately resulted in precipitates of calcium carbonate. The powder obtained was characterized further in detail for structure and morphology.

### Chemical precipitation

One of the most used procedures in the synthesis of $CaCO_3$ micro- and nanoparticles is the precipitation technique. This technique involves a very simple process without additives and is conducted by the mere mixing of supersaturated solutions of $Na_2CO_3$ and $CaCl_2$. Control over the size and morphology of synthesized particles was attained by tuning parameters such as reactant concentration, reaction time, and temperature. The solutions of sodium carbonate $Na_2CO_3$, 0.1 M, and calcium chloride $CaCl_2$ (0.1) solutions were stirred together and mixed at various speeds ranging from 300 to 30,000 rpm with a temperature range of 25 to 37 °C followed by continuous stirring for 20 min. The obtained solution was then taken in falcon tubes for centrifugation. After centrifugation, the pellet was washed a few times with ethanol and dried to be stored for further use (Fig. 2).

### Characterization of calcium carbonate nanoparticles

The characterization of synthesized nanoparticle has been done using several techniques like scanning electron microscopy which was conducted to study the morphology of synthesized calcium carbonate nanoparticles. XRD/max 2500 V diffractometer was used to measure the X-ray diffraction patterns and an FTIR-7600 instrument to get the Fourier transform infrared spectra. Dynamic light scattering (DLS) was used to determine the particle size distribution of calcium carbonate nanoparticles.

### Drug loading to synthesized nanoparticle

To conjugate nanoparticle with EGCG, two dilution of EGCG was prepared with in distilled water at 2 mg/ml and 4 mg/ml concentration. 10 mg of $CaCO_3$ was dispersed in 2 ml of ethanol. The 1 ml of the EGCG solution was added to 2 ml of ethanol–$CaCO_3$ NP mixture. The obtained mixture was ultrasonicated for 20 min for the conjugation of EGCG

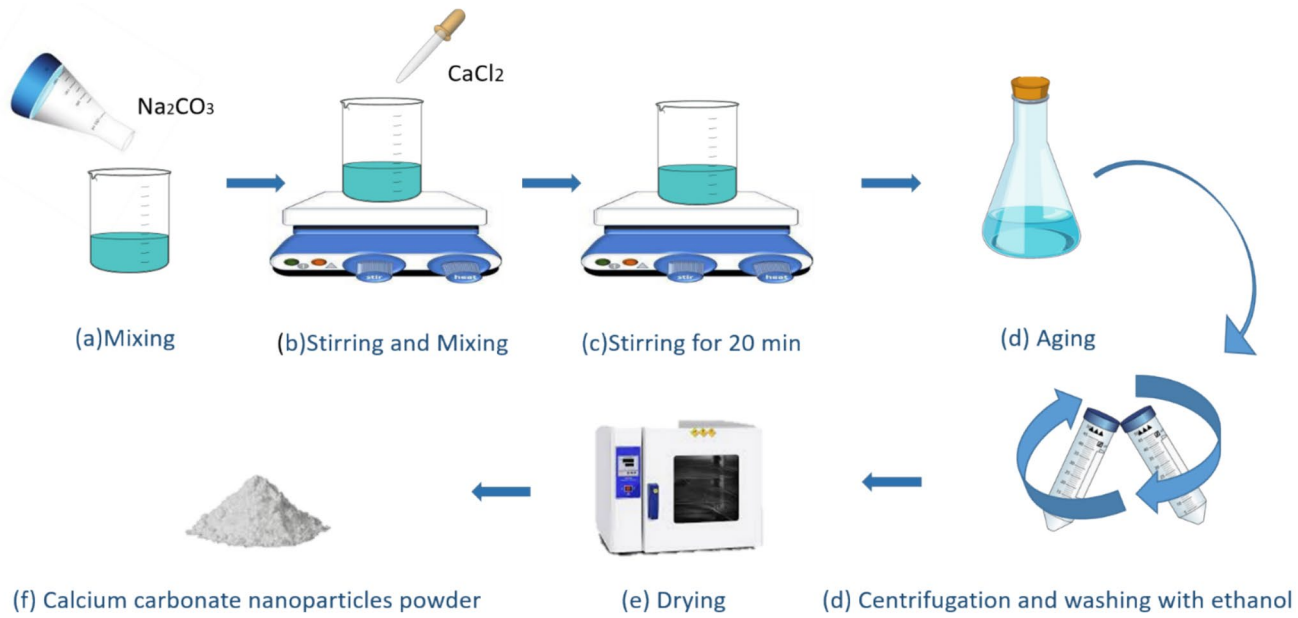



**Fig. 2** Synthesis process of calcium carbonate nanoparticles through the chemical precipitation method. Controlled chemical reaction between calcium chloride ($CaCl_2$) and sodium carbonate ($Na_2CO_3$), under well-defined environmental conditions, should yield calcium carbonate ($CaCO_3$) nanoparticles. Thereafter, the said nanoparticles shall be collected, washed, and dried for reuse

with the $CaCO_3$ nanoparticles. After sonication, dark incubation was carried out using an orbital shaker up to 24 h. Then, the samples were ultra-centrifugated for 10 min at 2000 rpm. The supernatant from each samples was taken to measure the percentage encapsulation using the UV–Vis spectrophotometer at a wavelength of 480 nm. The loading efficiency and loading capacity were determined from the readings in the UV–Vis spectrophotometer at 480 nm.

## Drug release study

In the present work, the release behaviors of EGCG from calcium carbonate nanoparticles were studied at pH 4.8 and 7.4 in PBS buffers. Drug release kinetics was studied using a standard curve for calculating the concentration of EGCG. Centrifuge pellet contained the EGCG-loaded calcium carbonate nanoparticles which was placed in dialysis tubing along with 3 ml of buffer solution. First, the test samples suspension of 3 mL EGCG-loaded sample (4 mg/mL and 6 mg/mL) was added to the dialysis bag (34 kDa). This dialysis bag was then immersed in a PBS buffer of 100 mL and shaken at 70 rpm in a water bath at 37 °C. At a preset time interval, 3 mL of the PBS buffers was withdrawn from the medium, which was replaced by an equal volume of fresh PBS. The released EGCG was determined in triplicate using a UV–Vis spectrophotometer at 424 nm. The experiment was conducted at two different pHs, 4.8, and 7.4 to find out any effect of pH on the drug release property. Figure 3.

The process involves EGCG solution and dispersing $CaCO_3$, followed by mixing and ultra-sonication. After incubation on an orbital shaker, the mixture is ultra-centrifuged, and the supernatant is analyzed for loading efficiency. The EGCG-loaded $CaCO_3$ is then used for UV–Vis analysis, and drug release is compared at pH 4.8 and 7.4.

## Cell cytotoxicity analysis

The human colon cell line, derived from the colon of a 55-year-old male and procured from NCCS Pune, was cultured on a 96-well plate using complete Roswell Park Memorial Institute (RPMI-1640) medium at a density of 1 × 104 cells per well. Cytotoxicity of calcium carbonate nanoparticles (CCN), EGCG-loaded CCN and EGCG alone were tested on colorectal cancer cell line (COLO-320 DM). The cells were subjected to different concentrations of calcium carbonate nanoparticles (1–150 µg/L), EGCG-loaded calcium carbonate nanoparticles (1–150 µg/L), and EGCG (1–150 µg/L). The cells were cultured at 37 °C. Then, 50 µL of MTT solution (with a concentration of 5 mg/mL in PBS) was introduced to the plates. The ELISA plate reader manufactured by Bio-rad was employed to measure the absorbance at 570 and 630 nm (with 630 nm serving as the reference wavelength). The cells were incubated in 96-well plates at 37 ℃, 5% $CO_2$ for 24 h.

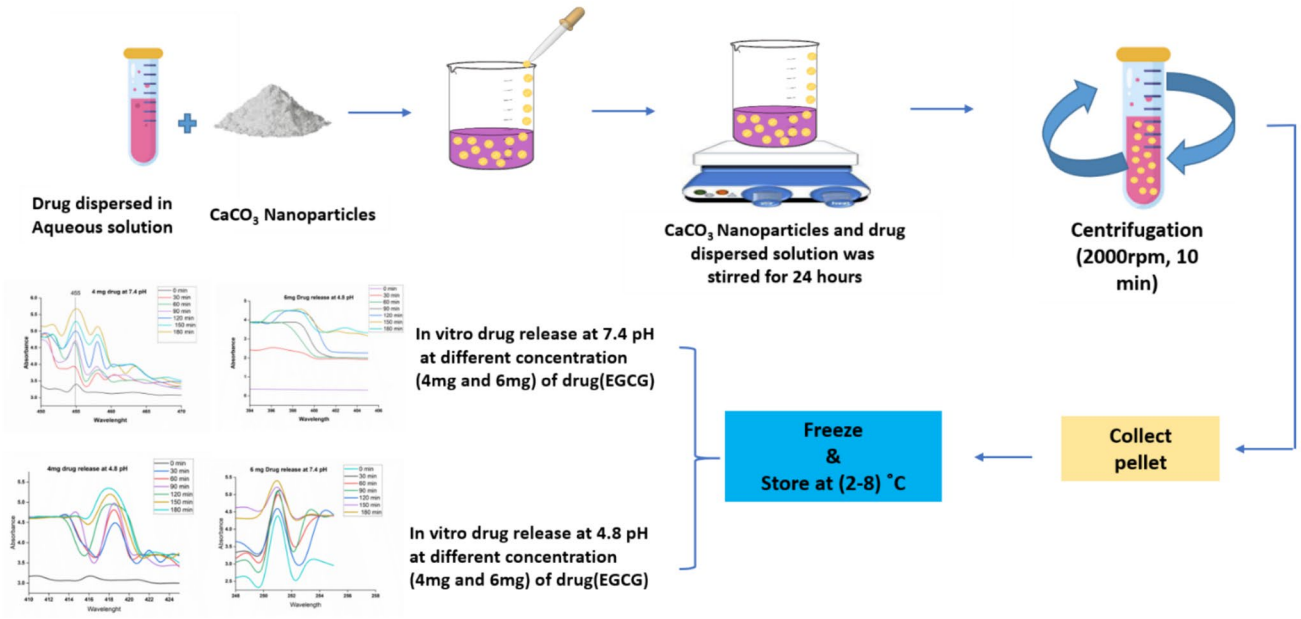



**Fig. 3** Schematic representation of EGCG drug loading and release in $CaCO_3$ nanoparticles

## Cell apoptosis study

The flow cytometry test was employed to quantify apoptosis in COLO-320 DM. The cells were seeded in six-well plates at a density of 3 × 105 and 4 × 105 cells per well, respectively. The plates were then incubated at 37 °C in a 5% $CO_2$ environment for 24 h. Subsequently, the cells were subjected to varying concentrations of calcium carbonate nanoparticles (100 µg/L), EGCG-loaded calcium carbonate (100 µg/L), and EGCG (100 µg/L) for a duration of 48 h. The cells were harvested by trypsinization. Subsequently, the cells were rinsed twice with PBS. The cells were treated with a binding buffer (100 mL), Annexin V-fluorescein isothiocyanate (FITC) (2.5 µL), and propidium iodide (PI) (2.5 µL) as directed by the kit instruction manual from Krishgen Biosystem, Mumbai. Subsequently, the cells were placed in a dark environment and kept at room temperature for a duration of 20 min. The assay findings were detected using flow cytometry (BD FACSCalibur$^{TM}$).

# Results and discussion

## Scanning electronic microscope analysis

The SEM image represented that the synthesized particles used to form a number of quasi-spherical particles with a rough surface texture. Figure 4 shows the existence of particle agglomeration as well. This was more obvious in the samples synthesized by lower molarity concentrations of reactants. A low vacuum mode of operation was used with the imaging process, since the $CaCO_3$ nanopowder is non-conductive. As observed from the SEM images themselves, these $CaCO_3$ nanoparticles are primarily made of flower-shaped crystal formations. Many aggregates are also identifiable, making it appear that the NPs might have an inherent tendency to agglomerate (Fig. 4). The likely reason for their very probable agglomeration is the very high surface energy expected from the small particle size, an allegation that is later verified from the results of the XRD phase study. The morphology investigation revealed that the nanoparticles are circular and flower-like in shape, with diameters from 80 to 300 nm.

## XRD analysis

The crystalline structure of the synthesized samples was characterized by X-ray diffraction analysis. Obtained XRD patterns demonstrated that the formed samples have a single-phase structure and calcite crystal symmetry, which substantiated the formation of pure calcium carbonate without the signature of any detectable secondary phases. The crystal size of the particles was also estimated from the obtained XRD data, which again confirmed the nanoscale dimensions of the synthesized material. The intensity and sharpness of the XRD peaks reflect the degree of crystallinity of the $CaCO_3$ nanoparticles. The predominant XRD peak signifying calcite was typically observed at approximately $2\theta = 26.5$, 27.7, and 37.4 degrees. Additional notable peaks were observed at $2\theta$ values of 45.9, 48.6, and 51.4 degrees,

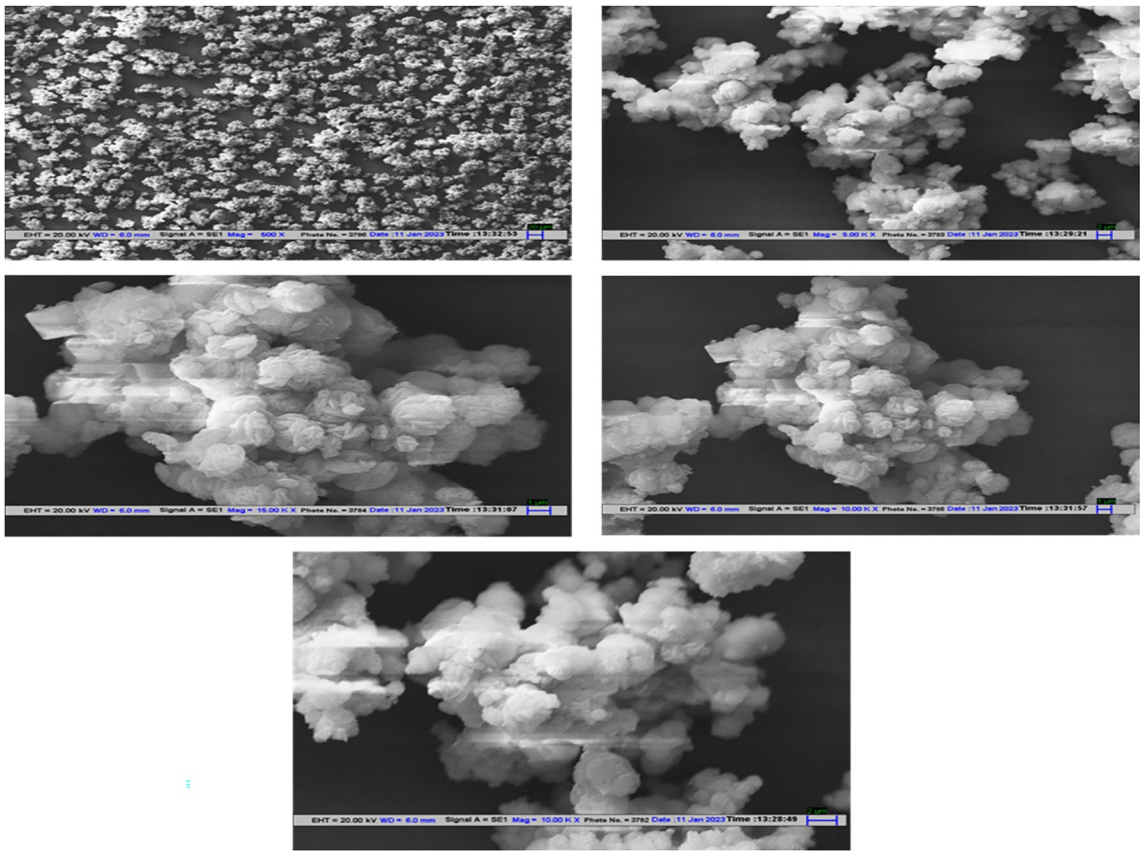


**Fig. 4** SEM images of synthesized calcium carbonate nanoparticles, confirming particle sizes in range of 100–200 nanometer

thereby confirming the formation of calcite. Calcite is the most stable polymorph of $CaCO_3$

## FTIR analysis

FTIR spectroscopy is important in identifying various phases of organic and inorganic compounds and, in particular, calcium carbonate phases due to differences in their carbonate ions, $CO_3^{2-}$. Carbonate ions and similar molecules exert four normal modes of vibration peaks: the bands at 712, 848, and 872/cm were assigned to the calcite vibration mode of a calcium carbonate nanoparticle.

## DLS analysis

To further elucidate on the nanoparticles' size distribution, dynamic light scattering measurements were performed. From DLS analyses, information related to average particles' size in synthesized samples, as well as their size distribution, emerges. The evidence established that variations of synthesis parameters alter the size of the particles, clearly expounding on the phenomenon involved in the synthesis process affecting the final characteristics of the particles. The particle size distributions of drug-loaded and blank $CaCO_3$ nanoparticles are shown in Fig. 5. The $CaCO_3$ nanoparticles without EGCG and those with it loaded had average diameters of 114.64 nm and 289.09 nm, respectively. With polydispersity index (PDI) values of 0.465 for the drug-loaded $CaCO_3$ nanoparticles and 0.196 for the blank CaCO3 nanoparticles, both formulations showed comparatively monodisperse particle size distributions.

## In-vitro drug release kinetics

The drug release properties of EGCG-loaded CCNPs were performed at a temperature of 37 °C with a gentle and consistent stirring motion. At fixed time intervals (0, 30, 60, 90, 120, 150, and 180 min), 1000 µl of each solution was transferred into a cuvette for spectrophotometry absorbance measurements at every time intervals (Fig. 6A, B). The encapsulation efficiency of EGCG inside CCNPs was 66.2%. This might be due to a large surface area of the CCNPs and/or protective effects of calcium carbonate, leading to the awry of premature drug release within biological system. The drug release profile showed a controlled release of EGCG from CCNPs at both pH 7.4 and 4.8 PBS (Fig. 6A). A burst drug release was followed by a much slower and steady release until the equilibrium stage at both pH levels. Results indicate that CCNPs loaded

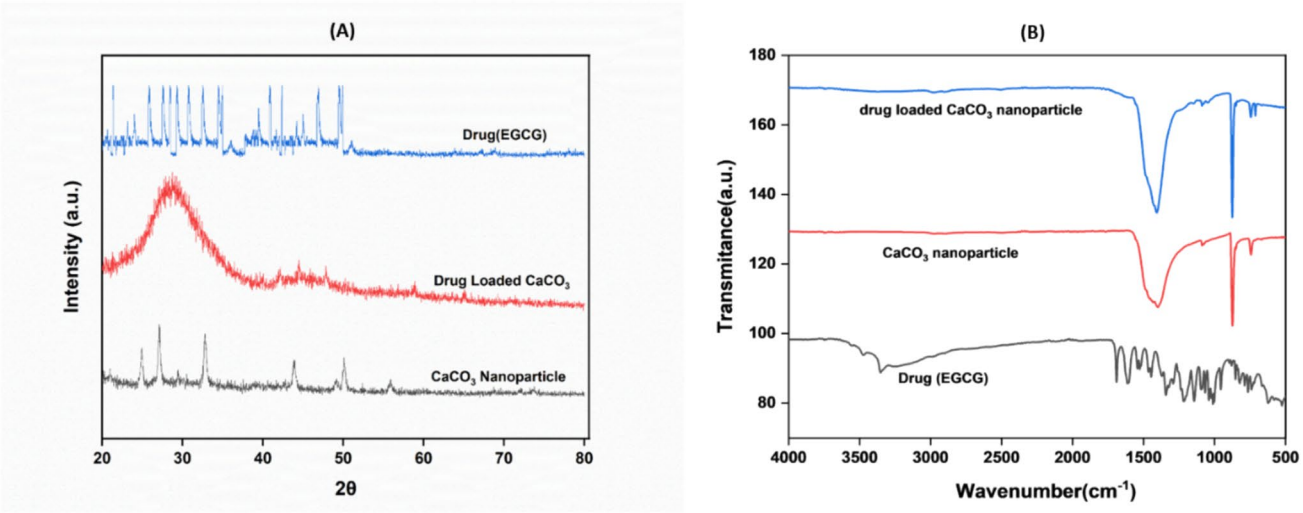



**Fig. 5** **A** X-ray diffraction and **B** FTIR spectra of $CaCO_3$ nanoparticles, drug (EGCG)-loaded $CaCO_3$ nanoparticles and drug alone

with EGCG are suitable drug carriers for targeting pH-specific cancer cells, with a leveled reduction of early drug release in nontarget body regions.

Further studies of EGCG release in the presence of acidic condition at pH 4.8 (mimicking the tumor microenvironment) also showed much faster release at pH 4.8 compared to pH 7.4. Specifically, at pH 4.8, around 50% of EGCG was monitored to be released within the first hour, while it was nearly 20% at pH 7.4, as demonstrated in Fig. 6A, B. The acidic medium assists in the solubilization of the calcium carbonate nanoparticles, thus liberating calcium ions ($Ca^{2+}$) along with carbon dioxide ($CO_2$) gas, as represented in the reaction:

$$CaCO_3(s) + 2H^+(aq) \rightarrow Ca_2^+(aq) + CO_2(g) + H_2O(l)$$

This rapid dissolution of the calcium carbonate matrix leads to an immediate release of the drug encapsulated inside the matrix. All these results taken together demonstrated that CCNPs exhibited a good response in the release profile of EGCG in the presence of ROS, indicative of the potential of the nanoparticles to be used as a carrier for the site-specific delivery of drugs in the cancer treatment process.

## Cell cytotoxicity analysis

The cytotoxicity of calcium carbonate nanoparticles (CCNPs), EGCG-loaded CCNPs, and free EGCG was evaluated in the COLO-320 DM cell line. EGCG-loaded CCNPs show a greater cytotoxic effect on the COLO-320 DM cell line compared to CCNPs and free EGCG. These results could suggest that EGCG-loaded CCNPs had effective interaction with the outer layer of the cancer cells via folate receptors. Controlled release of EGCG specifically at the site of cancer will facilitate robust interaction between EGCG with the DNA base pair of the cancer cells, leading to apoptotic cell death.

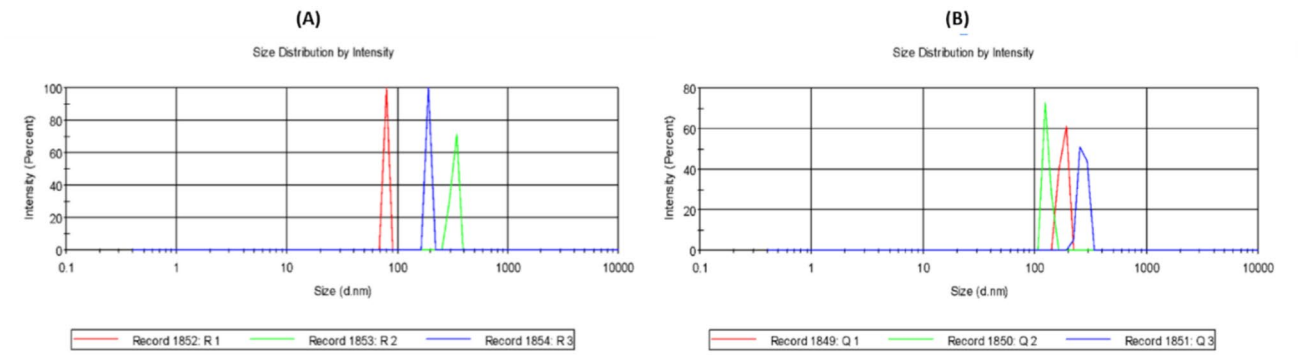



**Fig. 6** The graph illustrates the size distribution of **A** $CaCO_3$ nanoparticles and **B** EGCG-loaded $CaCO_3$ nanoparticles

### Apoptosis analysis

The incidence of early and late apoptosis in the COLO-320 DM cell line treated with EGCG-loaded CCNPs exhibited a considerable increase compared to other treatment methods (CCNPs alone and EGCG alone). The percentage of cell fractions in early (Q3) and late (Q2) apoptosis increased in the EGCG-loaded CCNPs treated group, indicating the cytotoxic effects of the developed nanoparticles against the COLO-320 DM cell line. There were no significant changes seen between the control group and the other groups, which confirms that the NPs had minimal side effects on normal cells. The results align with the MTT findings, suggesting that the EGCG-loaded calcium carbonate NPs exhibit greater cytotoxic effects on the malignant cell line compared to the other samples (Figs. 7, 8).

## Discussion

Surgery, radiotherapy, and chemotherapy are some of the conventional modalities, but they have considerable side effects and long-term disadvantages. In that direction, the dose-dependent and insolubility-related toxicities of drugs also become a major challenge in cancer treatment. Therefore, nanoparticle-mediated DDSs become very feasible ways to ensure targeted and effective delivery for better therapy outcomes. Many types of NPs have been studied for the possibility of being used as nanocarriers. The present study describes the successful synthesis, detailed characterization, and biological evaluation of calcium carbonate nanoparticles as carriers for the anticancer agent epigallocatechin gallate. Chemical precipitation produced the nanoparticles with the desired properties in terms of structure and morphology. SEM results showed the formation of quasi-spherical nanoparticles with a rough surface and aggregation behavior that could be due to high surface energy. The crystalline nature of powder was confirmed by the XRD pattern, performing calcite crystal symmetry and hence indicating the purity of calcium carbonate. FTIR spectra indicated characteristic peaks of carbonate ion thereby justifying the preparation of calcium carbonate. According to DLS analysis, size distribution and average particle size were noted, whereby blank $CaCO_3$ nanoparticles were about 114.64 nm and that of the drug-loaded nanoparticle was 289.09 nm; hence, EGCG was successfully loaded. Drug release patterns of EGCG from the prepared calcium carbonate nanoparticles were carried

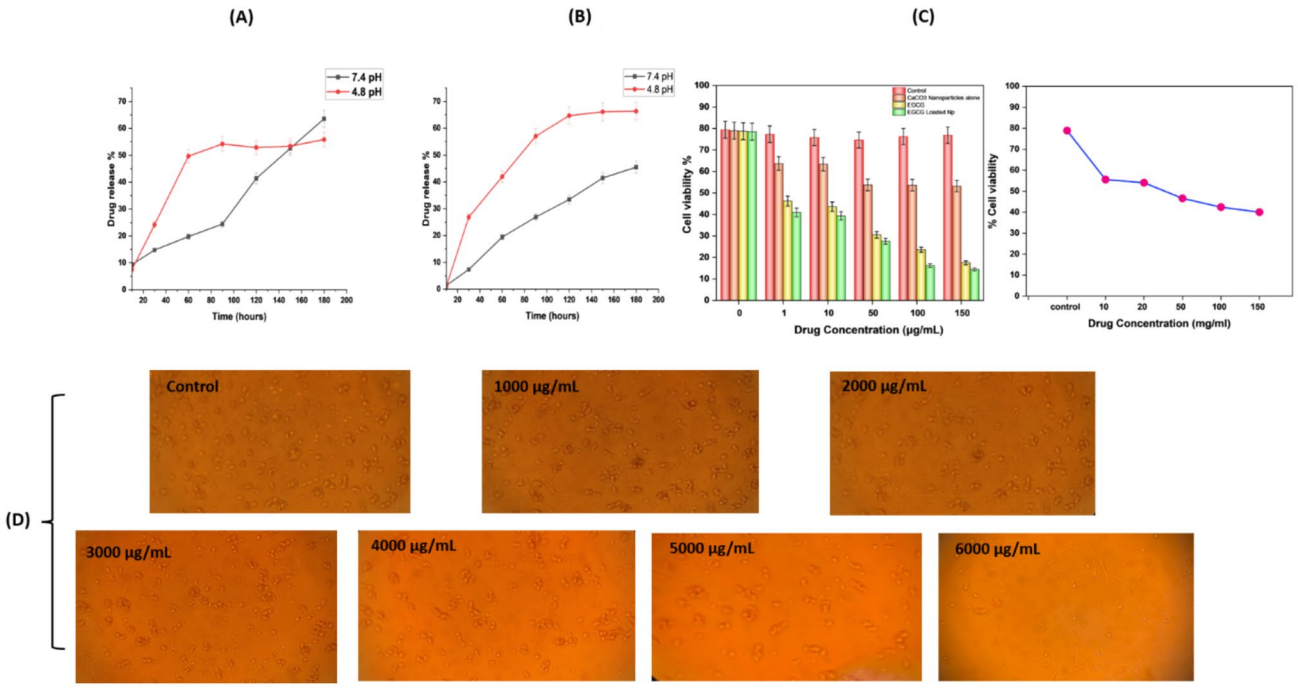



**Fig. 7** The drug release profiles of EGCG-loaded CCNPs were evaluated at pH 7.4 and pH 4.8, with concentrations of 4 mg/ml (**A**) and 6 mg/ml (**B**). Initial burst followed by extended release of drug until equilibrium favors pH-specific targeting in cancer. (**C**) The MTT assay results demonstrated a significant cytotoxic effect of the EGCG-loaded CCNPs compared to CCNPs and free EGCG. This validates the successful targeting of cancer cells and induction of apoptotic cell death. (**D**) More dose dependency towards the apoptotic and morphological changes in EGCG-loaded NNPs treated COLO-320 DM cell against the varying concentrations with a minimum side effect in control groups

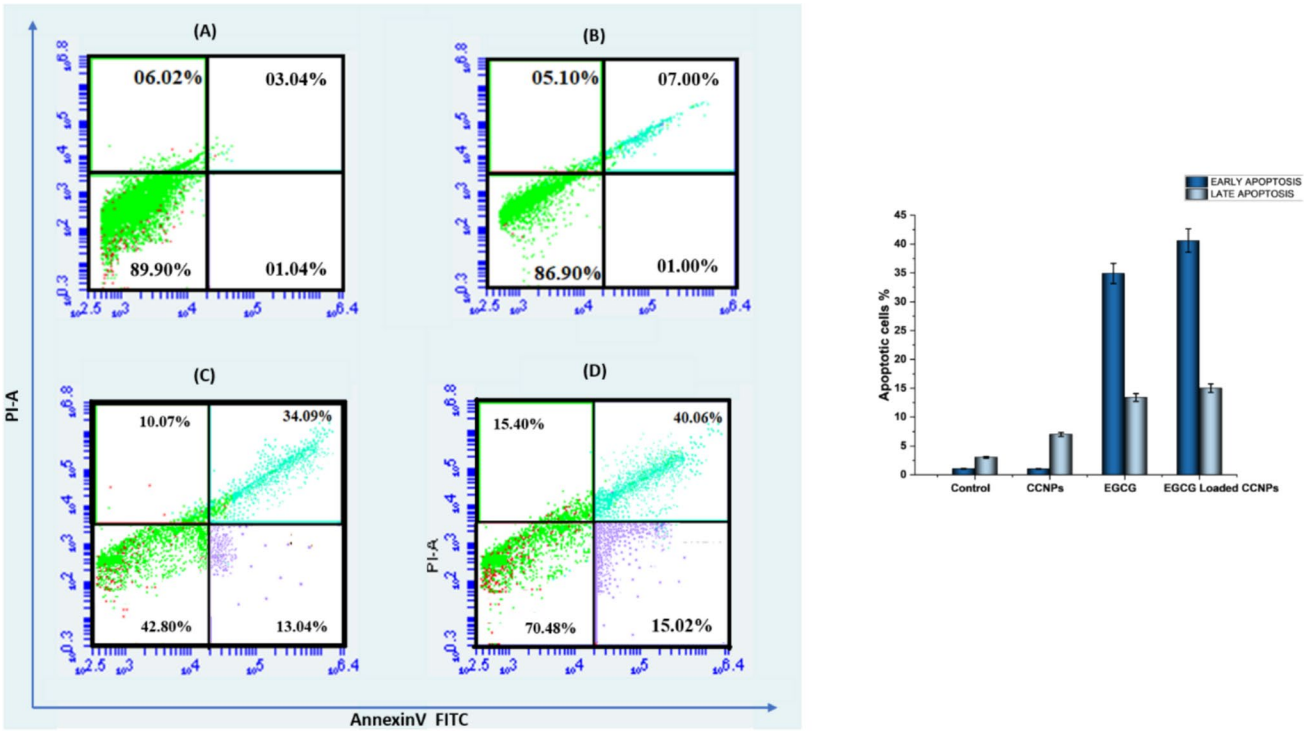



**Fig. 8** Flow cytometry analysis was performed to assess cancer cell death using Annexin V-FITC/PI staining in COLO320DM cells. The quantification of the percentage of cell death was determined. The infusion of EGCG-loaded calcium carbonate nanoparticles (CCNPs) (**D**) results in a notable increase in both early (Q3) and late (Q2) apoptosis when compared to the control COLO-320 DM cells (**A**), CCNPs (**B**) alone and free EGCG (**C**). The control flask was treated with DMSO control for 24 h

out in PBS buffers at pH 4.8 and 7.4. Release kinetics indicated that EGCG-loaded CCNPs provided controlled and sustained drug release. The release rate was pH-dependent, with faster release at the lower pH of 4.8 simulating the acidic tumor microenvironment. This is typical pH-responsive release behavior, which is advantageous for targeted cancer therapy by ensuring preferential drug release at the tumor site and, therefore, reducing system toxicity. The cytotoxic effects on the human colorectal cancer cell line COLO-320 DM were evaluated for CCNPs, EGCG-loaded CCNPs, and free EGCG. Results demonstrated that EGCG-loaded CCNPs exhibited greater cytotoxicity compared to CCNPs alone and free EGCG. Flow cytometry analysis revealed a significant increase in both early and late apoptosis in cells treated with EGCG-loaded CCNPs compared to the other treatments, underscoring their potent apoptotic-inducing effects on colorectal cancer cells, which is in agreement with the cytotoxicity results.

## Conclusion

Our research demonstrates the successful synthesis of EGCG-loaded calcium carbonate nanoparticles as an effective drug delivery system. This approach holds significant promise for development of calcium carbonate nanoparticles as a carrier for EGCG. It showed that such nanoparticles had proper structural properties and controlled release with improved anticancer efficacy. The utilization of calcium carbonate nanoparticles not only reduced the unfavorable effects in cell lines that were reliant on the dosage, but also preserved the anticancer efficacy of EGCG in COLO-320 DM cells. Consequently, it demonstrated that due to the pH-dependent drug release activity of the drug-loaded nanoparticles, EGCG-loaded CCNPs can be employed as a safe, efficient method of targeted cancer therapy. The results demonstrate that the cytotoxicity effects of calcium carbonate nanoparticles, EGCG-loaded calcium carbonate nanoparticles, and EGCG were influenced by the dosage and duration of exposure. As a result, such nanoparticle-mediated drug delivery systems would significantly enhance the potential to improve efficacy besides enhancing safety in cancer treatment. Future studies should aim at clearly resolving into

mechanisms of selective toxicity, and in vivo testing for evaluating the pharmacokinetics, biodistribution, and therapeutic efficacy of EGCG-loaded CCNPs using animal models is of prime importance. Furthermore, exploring the potential of this nanocarrier system for other therapeutic agents could broaden its application within nanomedicine.

**Acknowledgment** The authors express their gratitude to Delhi Technological University for its support in completing this study.

**Author contributions** Ms. Ritu conducted the data collection, data analysis and prepared the manuscript, while Dr. Asmita Das and Dr. Prakash Chandra made significant contributions to manuscript revisions and editing. All the authors collectively reviewed the manuscript, engaged in discussions about the results, and participated in editing the article.

**Funding** Funding and grants for the current research were provided by the University Grants Commission, India to Ms.Ritu.

**Data availability** Our study does not involve any clinical data or clinical trial registration. Therefore, no registration details or trial registration numbers are applicable.

## Declarations

**Conflict of interest** The authors have no competing interests to disclose.

**Ethical approval** The research conducted at Delhi Technological University's Department of Biotechnology in India was an observational study. In accordance with the ethical guidelines of Delhi Technological University, ethical approval was deemed unnecessary for this study, as it did not involve human or animal subjects.

## References

Afzal O, Altamimi ASA, Nadeem MS et al (2022) Nanoparticles in Drug Delivery: From History to Therapeutic Applications. Nanomaterials. https://doi.org/10.3390/NANO12244494

Anand U, Dey A, Chandel AKS et al (2023) Cancer chemotherapy and beyond: Current status, drug candidates, associated risks and progress in targeted therapeutics. Genes Dis 10:1367

Ashique S, Bhowmick M, Pal R, Khatoon H, Kumar P, Sharma H, Garg A, Kumar S, Das U (2024) Multi drug resistance in Colorectal Cancer- approaches to overcome, advancements and future success. Adv Cancer Biol Metastas 10:100114

Baba AI, Catoi C (2007) Chapter 3. Tumor cell morphology. In: Bucharas, RO (ed.), Comparative Oncology, The Publishing House of the Romanian Academy, Bucharest. https://www.ncbi.nlm.nih.gov/books/NBK9553/

Bai S, Lan Y, Fu S, Cheng H, Lu Z, Liu G (2022) Connecting Calcium-Based Nanomaterials and Cancer: From Diagnosis to Therapy. Nanomicro Lett. https://doi.org/10.1007/S40820-022-00894-6

Blanco-Calvo M, Concha Á, Figueroa A, Garrido F, Valladares-Ayerbes M (2015) Colorectal Cancer Classification and Cell Heterogeneity: A Systems Oncology Approach. Int J Mole Sci 16:13610–13632

Brown JS, Amend SR, Austin RH, Gatenby RA, Hammarlund EU, Pienta KJ (2023) Updating the Definition of Cancer. Mole Cancer Res 21:1142

Burrell RA, McGranahan N, Bartek J, Swanton C (2013) The causes and consequences of genetic heterogeneity in cancer evolution. Nature 501:338–345

Cancer Classification | SEER Training. https://training.seer.cancer.gov/disease/categories/classification.html. Accessed 23 Jun 2024

Chehelgerdi M, Chehelgerdi M, Allela OQB et al (2023) Progressing nanotechnology to improve targeted cancer treatment: overcoming hurdles in its clinical implementation. Mol Cancer. https://doi.org/10.1186/S12943-023-01865-0

Chen BH, Hsieh CH, Tsai SY, Wang CY, Wang CC (2020) Anticancer effects of epigallocatechin-3-gallate nanoemulsion on lung cancer cells through the activation of AMP-activated protein kinase signaling pathway. Sci Reports 10:1–11

Debela DT, Muzazu SGY, Heraro KD, Ndalama MT, Mesele BW, Haile DC, Kitui SK, Manyazewal T (2021) New approaches and procedures for cancer treatment: Current perspectives. SAGE Open Med. https://doi.org/10.1177/20503121211034366

Douaiher J, Ravipati A, Grams B, Chowdhury S, Alatise O, Are C (2017) Colorectal cancer—global burden, trends, and geographical variations. J Surg Oncol 115:619–630

Du GJ, Zhang Z, Wen XD, Yu C, Calway T, Yuan CS, Wang CZ (2012) Epigallocatechin Gallate (EGCG) Is the Most Effective Cancer Chemopreventive Polyphenol in Green Tea. Nutrients 4:1679

Dukes CE, Bussey HJ (1958) The spread of rectal cancer and its effect on prognosis. Br J Cancer 12:309–320

American Joint Committee on Cancer, American Cancer Society, SB Edge (2010) AJCC cancer staging handbook : from the AJCC cancer staging manual. Springer

Fujita KI, Kubota Y, Ishida H, Sasaki Y (2015) Irinotecan, a key chemotherapeutic drug for metastatic colorectal cancer. World J Gastroenterol 21:12234

Grumberg V, Chouaïd C, Gaudin A-F, Le Tourneau C, Marabelle A, Bongiovanni-Delarozière I, Cotté F-E, Borget I (2023) Organizational Impact of Immunotherapies in Advanced Cancers in France. JCO Glob Oncol. https://doi.org/10.1200/GO.23.00026/ASSET/IMAGES/LARGE/GO.23.00026TA2.JPEG

Kashyap BK, Singh VV, Solanki MK, Kumar A, Ruokolainen J, Kesari KK (2023) Smart Nanomaterials in Cancer Theranostics: Challenges and Opportunities. ACS Omega 8:14290–14320

Knowlton CA, Mackay MK, Speer TW et al (2024) Colon Cancer. Encycl. Radiat Oncol 1:77–77

Liang T, Feng Z, Zhang X, Li T, Yang T, Yu L (2023) Research progress of calcium carbonate nanomaterials in cancer therapy: challenge and opportunity. Front Bioeng Biotechnol. https://doi.org/10.3389/FBIOE.2023.1266888

Ma DD, Yang WX (2016) Engineered nanoparticles induce cell apoptosis: potential for cancer therapy. Oncotarget 7:40882

Maurya A, Singh AK, Mishra G, Kumari K, Rai A, Sharma B, Kulkarni GT, Awasthi R (2019) Strategic use of nanotechnology in drug targeting and its consequences on human health: A focused review. Interv Med Appl Sci 11:38–54

Min K, Kwon TK (2014) Anticancer effects and molecular mechanisms of epigallocatechin-3-gallate. Integr Med Res 3:16

Patra JK, Das G, Fraceto LF et al (2018) Nano based drug delivery systems recent developments and future prospects. J Nanobiotechnol 16:1–33

Patra JK, Das G, Fraceto LF et al (2018) Nano based drug delivery systems: Recent developments and future prospects. J Nanobiotechnology. https://doi.org/10.1186/S12951-018-0392-8

Puppa G, Sonzogni A, Colombari R, Pelosi G (2010) TNM Staging System of Colorectal Carcinoma: A Critical Appraisal of Challenging Issues. Arch Pathol Lab Med 134:837–852

Ritu CP, Das A (2023) Cholinergic side Effect-Free anticancer drugs: paving the way for safer and more effective cancer treatment. J Biomol Struct Dyn. https://doi.org/10.1080/07391102.2023.2274968

Sawicki T, Ruszkowska M, Danielewicz A, Niedźwiedzka E, Arłukowicz T, Przybyłowicz KE (2021) A Review of Colorectal Cancer in Terms of Epidemiology, Risk Factors, Development Symptoms and Diagnosis. Cancers. https://doi.org/10.3390/CANCERS13092025

Siegel RL, Miller KD, Wagle NS, Jemal A (2023) Cancer statistics, 2023. CA Cancer J Clin 73:17–48

Testa U, Pelosi E, Castelli G (2018) Colorectal Cancer: Genetic Abnormalities, Tumor Progression, Tumor Heterogeneity, Clonal Evolution and Tumor-Initiating Cells. Med Sci. https://doi.org/10.3390/MEDSCI6020031

Xiao Y, Li Z, Bianco A, Ma B (2023) Recent Advances in Calcium-Based Anticancer Nanomaterials Exploiting Calcium Overload to Trigger Cell Apoptosis. Adv Funct Mater. https://doi.org/10.1002/ADFM.202209291

**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.

# Swarm Intelligence in Robotics: Optimizing PSO Parameters for Target Search in Single and Multiple Scenarios

Vikas Sharma
*Department of Information Technology, Delhi Technological University, India*
vikassharma12387@gmail.com

Rahul Gupta*
*Department of Information Technology, Delhi Technological University, India*
rahul_gupta@dtu.ac.in
*Corresponding author

Kapil Sharma
*Department of Information Technology, Delhi Technological University, India*
*kapil@ieee.org*

2025 11th International Conference on Mechatronics and Robotics Engineering (ICMRE) | 979-8-3315-0929-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICMRE64970.2025.10976322

***Abstract*—Target-tracing applications have attracted much interest in unmanned vehicles and mobile robots, especially where human access is limited or dangerous. Several computational techniques such as genetic algorithms, evolutionary computations, and neural networks have been explored to improve the entire system's control mechanism and efficiency. This paper addresses PSO as a high-performance method for enabling collaborative robotic search operations. It has many critical parameters influencing its performance, which are finely optimized using another PSO process to make the best use of this optimization technique. Experimentations set forth show that this indeed does much improve the efficiency and accuracy of searches carried out by robots. These validation experiments work on single-target and multi-target cases to present vivid effectiveness. The results emphasize the ability of PSO to contribute toward autonomous search and rescue operations and all other applications.**

***Keywords—Particle Swarm Optimization (PSO), Collective Robotics, Target Search Algorithms, Optimization Parameters, Swarm Intelligence***

## I. Introduction

Mobile networks and robot systems are becoming increasingly inevitable in the control of unmanned physical systems in many applications, especially in dangerous or inaccessible environments to humans. They have been applied in applications such as search and rescue, environmental monitoring, industrial automation, and defense. Distributed architecture has then been adopted in these systems to decentralize, thus reducing computational complexity and offering more robust and scalable solutions. Key design considerations include prolonged operations while efficiently managing energy, effective communication with minimal interference, and high-speed computation to support real-time decision-making [1]. New methods have been developed from the advancement in technology, including wireless communication, vision-based applications, image processing, and neural networks to enhance the capabilities of a system. Many drawbacks such as communication interference and low transmission ranges are found among these challenges that reduce coordination efficiency among the agents in the network. Stochastic algorithms such as probabilistic techniques for finding optimal solutions are becoming more commonly used at the level of robotic agents. By focusing on local optimization, these algorithms ensure that only the most critical data is transmitted over communication channels, thereby reducing bottlenecks and improving overall performance [2], [3].

Among the many algorithms that could have been selected to apply, one that gained much attraction was PSO due to its performance over such complex optimization problems. PSO is motivated by collective behavior seen in the flocks of birds or schools of fish. In contrast to other evolutionary algorithms, PSO preserves the time persistence of the individual particles, or agents, which enables those particles to interact with each other and, hence, affect other particles' search patterns. This special feature enables PSO to exploit the social dynamics of swarms, resulting in faster and more efficient convergence to optimal solutions [4]. The PSO algorithm highly relies on the fine-tuning of its parameters, that is, inertia weight, cognitive acceleration constants, and social acceleration constants. These parameters drive the balance between exploration (that of searching for new solutions) and exploitation (that of refining known solutions). However, there is no one-size-fits-all set of such parameters as application requirements vary from one scenario to another. Recent advancements in PSO research have included adaptive and dynamic parameters on selection strategies and hybrid algorithms that combine PSO with other optimization techniques such as genetic algorithms or reinforcement learning [5].

This paper proposes an innovative dual-layer PSO approach to face the challenges mentioned above. A secondary PSO fitted to specific applications is first optimized by the first layer of parameters; the second layer uses the optimized PSO to accomplish collective robotic search tasks. This method ensures a dynamic adaptation of the parameters for maximum efficiency. The suggested framework is tested in single-target and multi-target search environments to demonstrate efficiency improvements in search and accuracy. Below is a step-by-step guideline for this paper: Section II is a succinct description of PSO, its principles, recent developments, and applications. Section III details the method of optimizing PSO parameters with a secondary PSO that gives innovations in adaptive parameter tuning and hybrid approaches. Section IV details how the optimized PSO can be implemented using current communication and computation techniques in a collective robotic search framework. Section V details experimental


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:48:57 UTC from IEEE Xplore. Restrictions apply.

results that compare the performance of the proposed approach to traditional methods. The results illustrate the scalability, reliability, and robustness of the system; therefore, its potential for real-world applications in disaster response, autonomous exploration, and smart infrastructure management is highlighted.

## II. Particle Swarm Optimization (PSO)

Kennedy and Eberhart originally introduced Particle Swarm Optimization Figure 1, which has since been developed into very advanced forms. Particle swarm optimization is a bio-inspired optimization algorithm that mimics the swarm intelligence found in nature, such as the collective behavior of bird flocks, fish schools, etc., to solve complex optimization problems [6], [7] PSO is perceived largely in terms of its simplicity, computational efficiency, and versatility across different domains. In contrast, whereas GAs rely on operators such as crossover and mutation, PSO essentially takes steps to dynamically adjust the velocities of particles within a population towards an optimal solution, thereby avoiding additional genetic operators and subsequently reducing parameter complexity [8], [9].

### A. Core Principles of PSO

PSO operates by initializing a swarm of particles, each representing a potential solution to the optimization problem [10]. The particles explore the search space iteratively, guided by their personal best positions ($pbest$) and the global best position ($gbest$) Found by the entire swarm. Their movements are influenced by random weights and acceleration constants, ensuring a balance between exploration (searching new regions) and exploitation (refining known good solutions). The velocity and position update equations are as follows:

- **Velocity Update:** $V(k+1) = w \cdot V(k) + c1 \cdot rand1 \cdot (pbest - P(k)) + c2 \cdot rand2 \cdot (gbest - P(k))$
- **Position Update:** $P(k+1) = P(k) + V(k+1)$

Here, w is the inertia weight that controls the trade-off between exploration and exploitation, and c1, and c2 are cognitive and social components that determine the attraction of the particle toward $pbest$ and $gbest$, respectively.

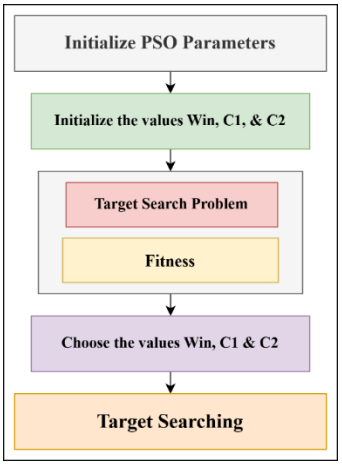



Fig. 1. Flowchart for determining optimal PSO

### B. Recent Developments and Enhancements

Recent advancements in PSO have focused on addressing its inherent challenges, such as premature convergence, stagnation in local minima, and sensitivity to parameter settings. Some of the notable enhancements include [11]:

***Dynamic and Adaptive PSO Variants:***

Researchers have developed adaptive PSO algorithms where parameters like inertia weight ($w$) and acceleration constants ($c1, c2$) Dynamically change during the search process. For instance:

- Linearly decreasing inertia weight encourages exploration in the early stages and exploitation in later stages.
- Adaptive learning strategies use swarm feedback to adjust parameters in real-time, improving convergence.

### C. Hybrid Algorithms

PSO has been combined with other optimization techniques, namely Genetic Algorithms, Differential Evolution, and Reinforcement Learning, to produce hybrid approaches. Hybrid methods take advantage of the strengths of multiple algorithms in order to tackle high-dimensional problems or multi-objective problems more effectively.

### D. Multi-Objective Optimization

Multi-objective PSO algorithms extend the traditional single-objective framework to optimize multiple conflicting objectives simultaneously. Concepts related to Pareto dominance are incorporated into MOPSO to locate a set of optimal trade-off solutions.

### E. Topology Variants

Swarms have been designed using different topologies, such as ring and star, and hierarchical structures that enhance information sharing between particles. While global neighborhoods, as in the star topology, improve the convergence rate, local neighborhoods, as in the ring topology, prevent premature convergence caused by greedy exploration.

### F. PSO in Dynamic and Noisy Environments

Dynamic versions of PSO address optimization problems where the objective function or the constraints of the optimization problem vary with time. In these algorithms, memory-based, diversity enhancement, or prediction models are used to address change in the environment. PSO algorithms against noisy scenarios employ robust fitness evaluation methods to ensure convergent solutions.

### G. Quantum-inspired PSO

Quantum principles have been integrated into PSO to enhance exploration abilities. Inspired by quantum principles, particles are represented as probability distributions by quantum-inspired PSO, so they can explore the search space much more effectively, especially in high-dimensional problems.

**Contributions of This Paper**


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:48:57 UTC from IEEE Xplore. Restrictions apply.

This paper builds upon these advancements by presenting a novel dual-layer PSO framework. The first layer optimizes the parameters of the secondary PSO, ensuring that it adapts to the specific requirements of the target application. This secondary PSO is then applied to collective robotic search scenarios, addressing single-target and multi-target search problems. Key innovations include:

- Adaptive parameter tuning strategies that dynamically adjust based on swarm behavior.
- Hybrid integration of PSO with complementary techniques for enhanced robustness.
- A focus on scalability and real-time performance in dynamic environments.

Particle Swarm Optimization continues to evolve, driven by advancements in adaptive strategies, hybridization, and real-world applications. This paper demonstrates how the dual-layer PSO framework effectively leverages these enhancements to deliver superior performance in complex optimization tasks. The proposed approach holds significant potential for further expansion into emerging fields such as autonomous systems, smart infrastructure, and next-generation artificial intelligence applications.

## III. Optimal PSO

This paper investigates the use of Particle Swarm Optimization (PSO) for parameter optimization in target-search applications at the ground level. The problem is divided into two parts for simplicity: first, the single-target case is analyzed, followed by the more complex multiple-target scenario. The performance of the PSO algorithm is influenced significantly by its key parameters, including the inertia weight ( $w$ ) and acceleration constants ($c1\ and\ c2$).

For efficient performance, the dynamic range of $w$ is set between 0.2 and 1.2, while the commonly adopted values for $c1\ and\ c2$ are initialized at 2. These parameters play a critical role in controlling the velocity of particles as they explore the search space. Initially, when the swarm begins moving, the particles are randomly oriented, requiring a higher velocity for broad exploration. As particles near the target, it becomes crucial to reduce their velocity to prevent overshooting. A higher $w$ at this stage retains momentum, while a lower $w$ enables finer adjustments and improves accuracy. Dynamically adjusting these parameters throughout the search can enhance performance but adds complexity to the optimization process.

This paper introduces a dual-hierarchy PSO framework. This framework optimizes the parameter values ($w, c1, and\ c2$) By applying an outer PSO layer to adjust the parameters for an inner PSO layer. The process involves:

### *A. Fitness Function and Iterative Convergence*

The fitness function for the outer swarm is defined as the number of iterations the inner swarm takes to achieve an error within a specified limit. This design ensures that minimizing the number of iterations in the inner swarm corresponds to reduced computation time and increased efficiency. The optimized parameters ( $w = 0.8, c1 = 2, c2 = 2$ ) The outer swarm consistently improves the performance of the target-search problem by achieving faster convergence.

This approach was tested on both single-target and multiple-target search problems to validate its effectiveness and consistency. Results show that the dual-hierarchy method not only improves the accuracy of target localization but also reduces computational overhead. Additionally, the hierarchical approach offers several advantages:

- **Scalability:** The method can adapt to different problem complexities, such as varying numbers of targets.
- **Dynamic Adaptability:** Parameters are fine-tuned in real-time based on the needs of the specific application.
- **Computational Efficiency:** By minimizing the number of iterations, the method reduces overall computation time.

The proposed dual-hierarchy PSO framework demonstrates significant potential in addressing complex optimization challenges in target-search problems. By effectively fine-tuning parameters through an outer PSO layer, it ensures improved performance and computational efficiency. This innovative approach is highly applicable to advanced robotics, autonomous navigation, and dynamic systems, where adaptive optimization plays a critical role.

## IV. Target Search Problem

The increase in the use of collective mobile robots for target search applications is because of the efficiency and cost-effectiveness advantages that they offer. One main reason for the increasing interest is that these robots can operate under minimal human intervention. Swarm intelligence techniques, such as Particle Swarm Optimization (PSO), enable low-cost, expendable robots to finish tasks that would otherwise be impossible or extremely expensive for humans to do [12]. By leveraging swarm-based algorithms, multiple robots can work collaboratively to explore and search areas that would otherwise be difficult or dangerous for human teams. This paper explores the use of the optimized PSO algorithm for both single-target and multiple-target search applications [13].

### *A. Single-Target Search*

The setup is found within a defined search space for the single-target search scenario. The main assumption made here is that there exists one target, and its position is known. The objective of the robots is to converge together at the location of the target. Each robot is guided towards the target by the use of the PSO algorithm. The particles in the PSO algorithm are regarded as moving robots, while each robot's motion depends on both the history of the same robot and on social exchanges with other robots in the swarm. These interactions are aimed to be guided by the PSO algorithm so that the swarm achieves the target efficiently.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:48:57 UTC from IEEE Xplore. Restrictions apply.

In this case, the fitness function is Euclidean distance from every robot to the target, as defined by Equation. Euclidean distance is the standard method of calculating the straight-line distance between two points in the search space, ensuring that the calculated motion of each robot moves optimally toward the target.

Fitness= $\sqrt{(T_x - P_x)^2 + (T_y - P_y)^2}$

Where:

- $Tx\ and\ Ty$ are the coordinates of the target,
- $Px\ and\ Py$ are the coordinates of a robot.

In this scenario, the goal of the swarm is for all the robots to work together and reach the target location. Each robot's movement is adjusted using the PSO algorithm, which factors in its own best position ($pbest$) and the best position achieved by the swarm as a whole ($gbest$).

### B. Multiple-Target Search

For multi-target searches, guiding robots toward different targets simultaneously involves more complexity. In this case, it not only calls for navigating the swarm through a bigger area but also efficient collaboration to locate and reach multiple target points. The capabilities of the PSO algorithm are extended to support the accommodation of multiple targets by having the swarm of robots dynamically adjust their positions according to the location of each target.

Both for one and multiple targets, the use of PSO ensures optimal efficiency of the operation of the swarm, reducing both the time and energy required by the robots to accomplish their task. Swarm intelligence is quite more effective in such applications because it offers the ability for distributed problem-solving where each robot contributes towards accomplishing the overall goal, without needing a central control unit. This approach of using PSO for guiding mobile robots in target search tasks offers several benefits:

1. **Scalability:** Multiple robots can be used in larger search spaces, allowing for efficient exploration of areas that are too vast or inaccessible for humans.
2. **Cost-effectiveness:** Swarm intelligence techniques, like PSO, can be applied to low-cost robots, making this approach more affordable compared to human intervention or traditional robotic systems.
3. **Flexibility:** Robots can be deployed in various environments, such as disaster areas, hazardous zones, or large outdoor spaces, where human involvement would be risky or impractical.

It means that this PSO can be applied to the collective search problem whether there is one goal or multiple goals, thus proving the very applicability of this approach toward real-world situations where automation and efficiency are required. Therefore, the solution for the mobile robots that utilize swarm intelligence through a lead by PSO proves a reliable method for target search tasks. Using the PSO algorithm, each robot within the swarm will be able to utilize local knowledge and collective experience to efficiently scour the search space. This is both to improve swarm-level performance but also to achieve highly complex search missions with minimal human intervention.

## V. Results

### A. Single Target

TABLE I: PARAMETER VALUES WITHOUT PSO

| C1 | C2 | Win=0.8 |
|---|---|---|
| 2 | 2 | 371.1 |
| 0.5 | 0.5 | 185.4 |
| 0.5 | 2 | 197 |
| 2 | 0.5 | 127.4 |

TABLE II: PARAMETER VALUES WITH PSO

| Parameter Value | No. of Iterations |
|---|---|
| Win=0.3520<br>C1=0.5022<br>C2=1.4599 | 113 |
| Win=0.4500<br>C1=1.1914<br>C2=1.5284 | 127 |
| Win=0.4500<br>C1=1.1914<br>C2=1.3747 | 129 |

### B. Multiple Target

TABLE III: PARAMETER VALUES WITHOUT PSO

| C1 | C2 | Win=0.6 | | | |
|---|---|---|---|---|---|
| | | G1 | G2 | G3 | G4 |
| 2 | 2 | 144 | 144 | 140 | 144 |
| 0.5 | 0.5 | 139 | 140 | 140 | 142 |
| 0.5 | 2 | 189 | 182 | 167 | 219 |
| 2 | 0.5 | 179 | 237 | 169 | 278 |

TABLE IV: PARAMETER VALUES WITH PSO

| Parameter Value | G1 | G2 | G3 | G4 |
|---|---|---|---|---|
| Win=0.5500<br>C1=1.3881<br>C2=2.3259 | 142 | 148 | 144 | 142 |
| Win=0.5500<br>C1=1.3881<br>C2=2.377 | 129 | 129 | 122 | 128 |
| Win=0.5500<br>C1=0.5500<br>C2=2.0898 | 137 | 137 | 141 | 132 |

The results demonstrate the significant improvements achieved using optimized Particle Swarm Optimization (PSO) parameters in both single-target and multiple-target search scenarios. For single-target searches, unoptimized configurations resulted in high iteration counts, ranging from 127.4 to 371.1, depending on parameter values such as $C1, C2$, and $Win$. In contrast, optimized PSO parameters ($Win = 0.3520, C1 = 0.5022, C2 = 1.4599$) reduced iterations to just 113, showcasing enhanced efficiency. Similarly, in multiple-target scenarios, unoptimized parameters produced variable performance, with iteration counts ranging from 140 to 219 across four targets. However, optimized parameters ($Win =$


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:48:57 UTC from IEEE Xplore. Restrictions apply.

$0.5500, C1 = 1.3881, C2 = 2.3259$) delivered consistent results, reducing iteration counts to a range of 129 to 148. These findings emphasize the effectiveness of the dual-hierarchy PSO framework in balancing exploration and exploitation, significantly improving search accuracy and computational efficiency. Additionally, the higher value of the social acceleration constant ($C2$) compared to the cognitive constant ($C1$) highlights the importance of swarm interactions over individual contributions in achieving superior convergence.

## VI. Conclusion

Based on the study, we have shown how Particle Swarm Optimization works in improving the capabilities of collective mobile robots in target search applications. Optimizing the inertia weight and acceleration coefficients of the PSO algorithm has led to a significant improvement in the efficiency of the algorithm. With optimal values for these parameters, we ensured that the PSO algorithm could perform at its best for single-target and multiple-target search conditions. Such an approach to pre-optimization not only streamlines the algorithm's execution but also saves time and computational resources by eliminating manual adjustments. By applying the method of offline parameter optimization, the proposed procedure can be used in real-time environments, thus readily integrating it into practical applications. Moreover, our results indicate that the social component of PSO is also significant because a higher value was assigned for c2, which corresponds to the social acceleration constant, than for c1, the cognitive acceleration constant. This fact indicates that interaction among particles such as robots plays a more crucial role than individual experience in navigating towards the target.

Future work will emphasize extending the applicability of this methodology for a wider variety of real-world applications such as investigating optimal swarm configurations and the dynamic modification of swarm size based on the intricacy of the underlying search space. With more experimentation and enhancement, the PSO-based approach has the potential to revolutionize autonomous robotic systems in several fields, including environmental monitoring, search and rescue missions, and autonomous vehicle navigation, through more efficient and reliable solutions.

## References


[1] Y. Cai and S. X. Yang, "A potential field-based PSO approach for cooperative target searching of multi-robots," in *Proceeding of the 11th World Congress on Intelligent Control and Automation*, IEEE, 2014, pp. 1029–1034.

[2] Y. Du, "A novel approach for swarm robotic target searches based on the DPSO algorithm," *IEEE Access*, vol. 8, pp. 226484–226505, 2020.

[3] J. Yang, X. Wang, and P. Bauer, "Extended PSO based collaborative searching for robotic swarms with practical constraints," *IEEE Access*, vol. 7, pp. 76328–76341, 2019.

[4] Y. Cai and S. X. Yang, "An improved PSO-based approach with dynamic parameter tuning for cooperative target searching of multi-robots," in *2014 World Automation Congress (WAC)*, IEEE, 2014, pp. 616–621.

[5] Y. Cai and S. X. Yang, "An improved PSO-based approach with dynamic parameter tuning for cooperative multi-robot target searching in complex unknown environments," *Int. J. Control*, vol. 86, no. 10, pp. 1720–1732, 2013.

[6] J. Tang, G. Liu, and Q. Pan, "A review on representative swarm intelligence algorithms for solving optimization problems: Applications and trends," *IEEE/CAA J. Autom. Sin.*, vol. 8, no. 10, pp. 1627–1643, 2021.

[7] J. Li and Y. Tan, "A two-stage imitation learning framework for the multi-target search problem in swarm robotics," *Neurocomputing*, vol. 334, pp. 249–264, 2019.

[8] M. Bakhshipour, M. J. Ghadi, and F. Namdari, "Swarm robotics search & rescue: A novel artificial intelligence-inspired optimization approach," *Appl. Soft Comput.*, vol. 57, pp. 708–726, 2017.

[9] Z. H. Ismail and M. G. M. Hamami, "Systematic literature review of swarm robotics strategies applied to target search problem with environment constraints," *Appl. Sci.*, vol. 11, no. 5, p. 2383, 2021.

[10] J. Pugh and A. Martinoli, "Inspiring and modeling multi-robot search with particle swarm optimization," in *2007 IEEE swarm intelligence symposium*, IEEE, 2007, pp. 332–339.

[11] M. Dadgar, S. Jafari, and A. Hamzeh, "A PSO-based multi-robot cooperation method for target searching in unknown environments," *Neurocomputing*, vol. 177, pp. 62–74, 2016.

[12] M. G. M. Hamami and Z. H. Ismail, "A systematic review on particle swarm optimization towards target search in the swarm robotics domain," *Arch. Comput. Methods Eng.*, pp. 1–20, 2022.

[13] K. Derr and M. Manic, "Multi-robot, multi-target particle swarm optimization search in noisy wireless environments," in *2009 2nd conference on human system interactions*, IEEE, 2009, pp. 81–86.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 06,2025 at 07:48:57 UTC from IEEE Xplore. Restrictions apply.

Water Air Soil Pollut (2025) 236:399
https://doi.org/10.1007/s11270-025-08052-5

# Technological Optimization, Health Risk Reduction, and Economic Analysis of Electrokinetic Remediation of Cadmium-contaminated Soil

**Sonam Taneja · Öznur Karaca · Anil Kumar Haritash**

Received: 28 January 2025 / Accepted: 19 April 2025
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2025

**Abstract** The electrokinetic remediation (EKR) technique was employed for cadmium (Cd)-contaminated soil to study the cross-effect of voltage gradient and electrolyte concentration on removal efficiency. Response Surface Methodology over Central Composite Design (CCD) was used to design the experiments. In addition, the effect of electrode material on EKR was investigated using two different electrodes, graphite, and stainless steel. The distribution and migration behavior of Cd in soil was explored through chemical speciation and sequential extraction procedure. The findings revealed that the voltage gradient and EDTA concentration were positively correlated with the removal efficiency at the optimized conditions of 2.5 V/cm and 0.15 M EDTA to achieve 27.8% removal in 10 h. The voltage gradient had a more profound effect on the removal efficiency as compared to EDTA concentration. The EKR treatment effectively reduced the bioavailability of Cd by removing water-soluble and exchangeable fractions and rendering the transformation of Cd to reducible and residual fractions. The Human Health Risk Assessment study was performed which revealed that EKR was successful in reducing the non-carcinogenic adverse effects by 29% and total carcinogenic risks from one in a thousand to one in a hundred thousand in adults, while children still exposed to high potential non-carcinogenic effects. The economic evaluation of all experiments concluded that the stainless steel electrode was more suited for Cd removal as compared to the graphite electrode with better removal (30%) and low specific cost (5.8 US$/g Cd). The study shows that EKR is successful in reducing the overall toxicity of Cd in soil and humans at a relatively low cost and less treatment time.

**Keywords** Cadmium · Electrokinetic remediation · RSM · Health Risk Assessment · Sequential extraction

S. Taneja (✉) · A. K. Haritash
Department of Environmental Engineering, Delhi Technological University, New Delhi 110042, India
e-mail: sonamtaneja_2k20phden04@dtu.ac.in

A. K. Haritash
e-mail: akharitash@dce.ac.in

Ö. Karaca
Department of Geological Engineering, Çanakkale Onsekiz Mart University, Çanakkale, Turkey
e-mail: oznurkaraca@comu.edu.tr

## 1 Introduction

Heavy metal contamination in soil poses intricate public health concerns due to its chemical stability and environmental persistence. Globally, about 20% of agricultural land is estimated to be contaminated with heavy metals (FAO, 2021). The introduction of elevated levels of metals into the soil can potentially

Published online: 29 April 2025

contaminate the aquifers and subsequently, transport the pollutants to larger distances (Taneja et al., 2023a). Cadmium (Cd) metal, in particular, is a non-essential, toxic heavy metal, which is ranked 7th among the top 20 pollutants on the list of Priority Substances according to the Agency for Toxic Substances and Disease Registry (ATSDR), (FAO, 2021). The background level concentration of Cd is 0.3–3 mg/kg in soil from the geogenic origin (WHO, 1993). However, anthropogenic activities mainly, mining, slag disposal, and excessive use of fertilizers can elevate the concentration in soil up to 200 mg/kg (Jawjit et al., 2024; Kubier et al., 2019).

Cd is highly soluble and is readily available for plant uptake, due to which it exhibits severe phytotoxicity and bioaccumulation (Genchi et al., 2020). Cd toxicity in humans occurs mainly through chronic exposure to Cd-contaminated food and water sources with concentrations above the recommended value of 3 µg/L (WHO, 1993). Long-term exposure to Cd can increase the maximum tolerable daily intake of 0.025 mg/kg body weight, leading to serious health implications such as cancer, kidney damage, bone fragility, and neurological disorders (Genchi et al., 2020). Furthermore, the persistence of Cd in the environment adversely affects soil quality (Jawad & Jahanger, 2024), microbial activity, plant growth, and the overall health of the ecosystem (Bolan et al., 2014), necessitating the removal of Cd from the soil. Conventional methods to remediate Cd-contaminated soils, such as soil washing, immobilization, and phytoremediation, have shown varying degrees of success. However, these methods are often limited by high costs, long remediation periods, or secondary environmental impacts (Khalid et al., 2017; Saeed et al., 2024).

Electrokinetic Remediation (EKR) has proven to be a promising technology for the removal of heavy metals from soil, with minimum disturbance to the soil environment (Taneja et al., 2024). The technique employs direct electric current to the contaminated soil using the electrodes installed across the soil column. The current initiates three key transport processes, electromigration, electro-osmosis, and electrolysis to cause mobilization and subsequently transfer of metals from soil into the electrode wells (Acar et al., 1995). The implications of EKR on polluted soils are multi-fold. It not only reduces the metal toxicity but plays a critical role in metal recovery which can then be reintroduced into the production processes, thus contributing to the circular economy (Asadollahfardi et al., 2021). Moreover, the application of electric current across the soil has proven to improve its physical properties such as stabilization (Pandey et al., 2023).

The success of EKR in removing metals depends mainly on the soil type, composition, and operating conditions, which influence the metal-soil interactions, such as, precipitation/dissolution, sorption/desorption, and soil pH (Yeung et al., 1997). Numerous studies have been conducted to overcome the formation of pH gradients and soil clogging that occurs due to electrolysis reactions and is the major hindrance in metal removal. Techniques like periodic voltage application (Cai et al., 2022) and polarity reversal (Wang et al., 2024) have resulted in higher desorption and less accumulation of metal ions near the cathode, which allows sufficient migration of metals. However, the reversal frequency, if not optimized, may lead to the re-adsorption of metals near the electrodes, reducing net removal. The approaching anode (AA) technique has also been tested to remove Cd from soil that resulted in 25.9% removal as compared to 17.9% from the conventional EKR approach (Zhang et al., 2016). This technique reduce the accumulation zone and improve migration with low energy consumption but, the physical movement of anode in soil may disturb the soil structure and increase the operational complexity in field conditions.

Electrolyte amendments with organic acids like lactic acid (Cai et al., 2022) and citric acid (Song et al., 2018) and chelating agents like EDTA (Ge et al., 2022) have been studied to improve the desorption and mobilization of metals by lowering the pH or by forming soluble metal-complexes, respectively. Citric acid, being a biodegradable organic acid, has displayed excellent mobilization and removal of Cd while lowering the pH of soil (Lai et al., 2023). EDTA, on the other hand, forms stable soluble complexes over neutral to alkaline pH soil and thus can be employed without the potential risk of soil acidification. The use of surfactants like Tween 80 (Taneja et al., 2023b) and Sodium Dodecyl Sulphate (SDS) (Peng & Tian, 2010) as electrolytes have been shown to remove metals from soil by reducing the interfacial tension between the metal and soil particles, however, it increases the risk of secondary pollution and environmental toxicity. The use of microorganisms as electrolytes has resulted in significant Cd removal

as compared to EDTA and citric acid by behaving as a biosurfactant with low pH and high electrical conductivity (Tsui et al., 2022). This bio-electrokinetic approach addresses the sustainability concerns of EKR, however, further investigation is needed to understand the microbial strains and optimize the operating conditions for effective removal.

The majority of the studies are carried out considering individual parameters at a time, which requires performing numerous experiments, thus making the OFAT (one factor at a time) approach less suited for EKR studies. Response Surface Methodology (RSM) is a useful tool for the design of experiments and to assess the relation between the operating variables for optimization. RSM is considered to be the most effective parametric approach to design minimum experiments and statistical optimization (Wang et al., 2024). Few studies have employed RSM to study the effect of individual factors as well as their interactions. Li et al., (2020) employed Box-Behnken Design (BBD) to investigate the interactive effects of voltage gradient, soil moisture, and citric acid concentration. Whereas, Wang et al., (2024) used BBD to conclude that 1.6 V/cm voltage gradient, 0.2 M citric acid, and 0.55 times/day polarity reversal frequency were the optimum conditions to achieve 65.3% Cd removal in 16 days.

The present study employs RSM using the Central Composite Design (CCD) which is more suited for two-factorial design. The objective of this study is to evaluate the efficiency of EKR to remove Cd ions from the soil while analyzing the cross-impacts of voltage gradient and electrolyte composition. While many studies have been conducted on Cd removal using EKR, limited attention has been given to the economic feasibility and toxicity reduction associated with EKR. The present study aims to address this gap by performing a comprehensive human health risk assessment methodology and economic evaluation to assess the potential reduction in public health risks by mitigating soil Cd toxicity in a cost-effective manner.

## 2 Material and Methods

### 2.1 Chemicals and Reagents Used

All chemicals and reagents used in this study were of analytical grade. Cd metal granules (99%) purity were purchased from SRL Chemicals, India, and ethylenediaminetetraacetic acid (EDTA) (98%) was purchased from Central Drug House (CDH) Pvt. Ltd. India. Type I ultrapure water was used to prepare all stock and standard solutions.

### 2.2 Soil Collection and Preparation

The soil sample was collected from the campus of Delhi Technological University at a depth of 0–20 cm from the surface. The surface layer containing litter, stones, and other impurities was removed, and the soil was air-dried, ground, and sieved through a 2 mm mesh size. A Cd stock solution was prepared using Cd granules to introduce Cd into the soil by thorough mixing to achieve the desired concentration of 250 mg/kg of Cd. The spiked soil was then mixed regularly with repeated wetting and drying cycles for one week to ensure uniform distribution. The physicochemical properties of soil before spiking were performed using standard methods as presented in Table 1.

### 2.3 Electrokinetic Remediation Setup

The experimental setup used in the study is shown in Fig. 1. The assembly constituted a soil cell, electrolyte chambers, electrodes, and power supply. The electrolytes used in this study were distilled water and EDTA (0.1 M and 0.2 M), and electrodes were made of graphite and stainless steel. For each experiment, 1.5 kg of spiked soil was pre-treated with the desired enhancing solution to attain a moisture content of 25%. The soil was left to equilibrate for 24 h to improve electro-osmotic flow and then added to the soil cell. Before initiating the experiment, the electrode wells were filled with 100 ml of the selected electrolyte, the electrodes were connected to the power supply through wires and clamps suspended in the wells. Each experiment was run in two cycles for 5 h each. After the first cycle of 5 h, the electrolytes were collected and the experiment was suspended for 14 h. Before initiating the second cycle, the electrode wells were refilled with fresh electrolytes and the experiment was then rerun for 5 h. After the completion of the second cycle, the electrolytes were collected, and the soil was extruded in four equal sections from S1 through S4

in the direction from the anode towards the cathode. The soil sections and electrolytes were then analyzed.

### 2.4 Experimental Design and Statistical Analysis

Response Surface Methodology (RSM) was used to design the experiments for this study using Central Composite Design (CCD). Since the model takes into consideration the continuous factors, only two parameters that showed continuous variations were considered i.e. voltage gradient ($X_1$) and electrolyte composition ($X_2$). Minimum and maximum levels for the parameters were ranged as 0.5–2.5 V/cm for $X_1$ and 0.0–0.2 M EDTA for $X_2$. Considering the factorial structure of CCD, the total number of runs was obtained using the formula: $N = 2^K + 2K + n_c$, where K is the number of factors ($K = 2$), and $n_c$ is the replicate number of the central point ($n_c = 1$). The Design Expert software was used to design 09 experimental runs ($N = 2^2 + 2(2) + 1 = 9$) for the response removal efficiency (Y) to derive the relation with the independent variables ($X_1$ and $X_2$). To test the adequacy of the model, one-way Analysis of Variance (ANOVA) was performed with predictability at 95% confidence level. In addition, the effect of electrode material on the efficiency of EKR for Cd removal was also investigated by using stainless steel as the electrode at 2.5 V/cm gradient using two different electrolyte concentrations, 0.0 M (R10) and 0.1 M (R11). Since it is a discrete variable, it was not considered for the RSM model. The experiments designed for this study and the corresponding response are presented in Table 2.

### 2.5 Analytical Methods

The pH and electrical conductivity (EC) of electrolytes were measured using a bench-top multiparameter (Systronics, LMMP 30 model). The Cd concentration was determined using an Atomic Absorption Spectrophotometer (AAS) (Analytik-Jena novAA 350) after the electrolytes were centrifuged at 5000 rpm. For soil samples, a 1:10 soil–water suspension was prepared using a shaker which was used for the analysis of pH and EC. To determine the concentration of Cd in soil samples, 5 g of oven-dried soil samples were thermally oxidized in a muffle furnace at 550ºC. After cooling, the sample was transferred into conical flasks and mixed with 50 ml of 5% nitric acid over an orbital shaker for 2 h. The recovered solution was centrifuged at 5000 rpm for 10 min and then analyzed over AAS. The removal efficiency of Cd was calculated according to the Eq. 1,

$$RX\ 100 \tag{1}$$

where, $C_0$ is the initial concentration of Cd in soil (mg) and C is the residual concentration of Cd in the soil after EKR treatment (mg). Energy expenditure is an important factor in assessing the economic viability of EKR technique. In this study, energy consumption is calculated using Eq. 2.

**Table 1** Characterization of soil before EKR treatment

| Soil characteristics | Values | Method |
|---|---|---|
| Soil Texture | Sandy Loam | ASTM D6913 |
| pH | 8.1 | ASTM D4972 - 19 |
| Electrical conductivity (mS/cm) | 0.17 | ASTM D4972 - 19 |
| Total Dissolved Solids (ppm) | 90 | ASTM D4972 - 19 |
| Organic Matter (%) | 1.6 | ASTM D2974 |
| K (mg/kg) | 488 | APHA 3111 |
| Na (mg/kg) | 5230 | APHA 3111 |
| Ca (mg/kg) | 3215 | APHA 3111 |
| Plastic Limit (%) | 12.72 | ASTM D4318 |
| Liquid limit (%) | 29.4 | ASTM D4318 |
| Cd (mg/kg) | BDL (< 0.002) | APHA 3500-Cd |

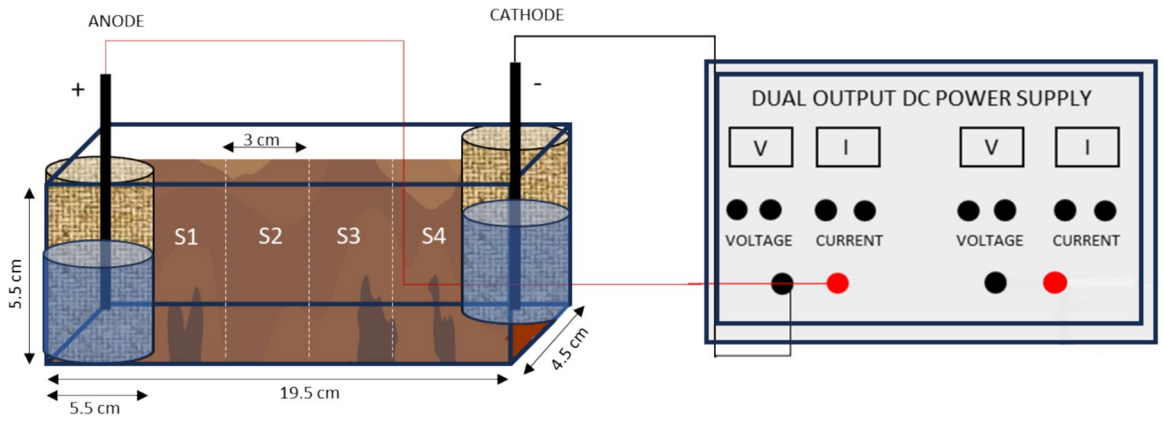



**Fig. 1** Schematic representation of EKR setup constituting a dual output power supply, soil cell, electrode wells, electrodes and wires

$$E = \frac{1}{Vs}\int VI(t)dt \tag{2}$$

where, V is the average voltage, I is the average current measured during EKR in time t and Vs is the volume of soil. The total cost of all experiments was also calculated which includes the cost of electricity, electrode, and chemicals used. A sequential extraction procedure was performed before and after EKR soil samples of the experiment with maximum removal as described by Reddy et al., (2001).

To assess human toxicity, a Human Health Risk Assessment (HHRA) was carried out using the simplified methodology developed by USEPA, 1989. Non-carcinogenic risk was estimated by calculating the Average Daily Dose (ADD, mg/kg/day) of Cd through oral and inhalation routes of exposure pertaining to the days of exposure (Eq. 3 and 4). The ADD was then calculated against the R*f*D to obtain the Hazard Quotient (HQ) (Eq. 5). The sum of HQ from the two pathways is the Hazard Index (HI) (Eq. 6), which determines the non-cancer risks associated with a contaminant, with HI < 1 having no adverse effects, and HI > 1 having potential adverse non-carcinogenic effects. Carcinogenic risks, on the other hand, were estimated through the product of ADD with the slope factor (SF) (Eq. 7 and 8). Total Carcinogenic Risk (TCR) is estimated from the sum of risks from the two pathways (Eq. 9). The normal value of TCR ranges from $10^{-6}$ to $10^{-4}$, above which, the associated contaminant poses adverse carcinogenic risk (Adimalla et al., 2020).

**Table 2** Experimental design using RSM and the results obtained

| Run order | Experimental Parameter | | Response |
|---|---|---|---|
| | A: Voltage gradient (V/cm) | B: EDTA concentration (mol/l) | Removal Efficiency (%) |
| 1 | 0.5 | 0 | 4.5 |
| 2 | 0.5 | 0.1 | 6.4 |
| 3 | 0.5 | 0.2 | 5.9 |
| 4 | 1.5 | 0 | 5.6 |
| 5 | 1.5 | 0.1 | 8.4 |
| 6 | 1.5 | 0.2 | 7.8 |
| 7 | 2.5 | 0 | 19.7 |
| 8 | 2.5 | 0.1 | 28.8 |
| 9 | 2.5 | 0.2 | 26.7 |
| 10 | 2.5 (SS) | 0 | 22.7 |
| 11 | 2.5 (SS) | 0.1 | 30.2 |

$$ADDing = \frac{CX\ IngR\ X\ EF\ X\ ED\ X\ EF}{BW\ X\ AT} X\ 10^{-6} \tag{3}$$

$$ADDinh = \frac{CX\ Ingh\ X\ ED\ X\ EF}{BW\ X\ AT\ X\ PEF} \tag{4}$$

$$HQ = \frac{ADDi}{RfDi} \quad (5)$$

$$HI = \sum HQ \quad (6)$$

$$CRing = ADDing\ X\ SFing \quad (7)$$

$$CRinh = ADDinh\ X\ SFinh \quad (8)$$

$$TCR = \sum (CRing + CRinh) \quad (9)$$

The values and physical significance of the selected variables for this study are presented in Table 3, where C is the concentration of Cd in soil.

## 3 Results and Discussion

### 3.1 RSM Model Development

The general quadratic RSM equation (Eq. 10) was used to fit the removal efficiency and the best model was obtained:

$$Removal\ Efficiency\ (\%) = 9.16 + 9.73A + 1.77B + 1.40AB + 8.07A^2 - 2.83B^2 \quad (10)$$

The Analysis of Variance (ANOVA) and regression coefficient are shown in Table 4. The F-value > 4 indicates that the response is independent of noise. For this model, the F-value was 49.03 and the p-value was 0.004 ($P < 0.05$ for 95% confidence level), implying that the model is significant. The difference between $R^2$ predicted (0.8607) and $R^2$ adjusted (0.967) was less than 0.2 which indicates that the model is credible with good predictability. The reasonably close values between predicted and experimental results validate the accuracy of the model (Fig. 2).

### 3.2 Response Plots and Optimization of Removal Efficiency

The 3D surface plot of the removal efficiency of EKR for Cd from soil is shown in Fig. 3a. The results show that the best removal was achieved in R8 (28.8%), followed by R9 (26.74%) and R7 (19.7%). At a lower voltage gradient (0.5–1.5 V/cm), the removal efficiencies of Cd ranged from 4.45% to 5.58%. A significant rise in removal efficiency was observed with the increase in voltage gradient from 1.5 to 2.5 V/cm at a constant electrolyte concentration. When the EDTA concentration increased from 0 M to 0.1 M, a slight increase in removal was observed. At the maximum voltage, increasing the EDTA concentration beyond 0.1 M did not increase the removal, rather showed a decline. The contour plot (Fig. 3(b)) illustrates that the removal efficiency of Cd was positively affected by the voltage gradient and the addition of EDTA as electrolyte. However, the effect of voltage gradient was significantly greater than that of EDTA concentration.

The experimental results suggest that a higher voltage gradient and EDTA amendments can effectively achieve higher removal of Cd. However, the increase in voltage also increases the power consumption. Based on the experimental results and model predictions, an optimization study was done using the desirability function of Design Expert Software to achieve efficient removal at low energy consumption and chemical costs (Fig. 4). The obtained optimal conditions were 2.5 V/cm voltage gradient and 0.15 mol/L EDTA concentration to achieve the corresponding removal efficiency of 27.8%.

**Table 3** Selected values for the variables used in HHRA (Adimalla et al., 2020; Wcisło et al., 2022)

| Variables | Physical significance and units | Values | |
|---|---|---|---|
| | | Adults | Children |
| *IngR* | Ingestion rate of soil (mg/day) | 100 | 200 |
| *InhR* | Inhalation rate of soil ($m^3$/day) | 12.3 | 7.6 |
| *EF* | Exposure frequency (days/year) | 365 | 365 |
| *ED* | Exposure duration (years) | 30 | 6 |
| *BW* | Average body weight (kg) | 70 | 20 |
| AT | Average exposure time (non-carcinogenic, days) | 8760 | 2190 |
| R*f*D*ing* | Reference Dose (Ingestion) | 4.00E- 02 | |
| R*f*D*inh* | Reference Dose (Inhalation) | 1.00E- 03 | |
| SF*ing* | Slope Factor (Ingestion) | 0.38 | |
| SF*inh* | Slope factor (Inhalation) | 6.3 | |

### 3.3 Variations in pH and Electrical Conductivity

The distribution of pH in electrolytes and soil is shown in Fig. 5(a). The changes in pH of the electrolyte were consistent throughout with a low pH of 0.8–4 at the anode and a high pH of 8.5–12 at the cathode, due to the formation of $H^+$ and $OH^-$ ions at the anode and cathode, respectively. In the absence of electrolyte amendments (R1, R4, R7, R10), the pH of the anolyte lowered with the increase in the voltage gradient from 0.5 to 2.5 V/cm, while the catholyte pH remained constant at 10. However, upon the addition of EDTA (R2, R3, R5, R6, R8, R9, R11), the catholyte pH rose to 12. Soil, on the other hand, was able to neutralize the acidic front from the anode and did not show significant variations in the pH from the initial value (7.9), as was also observed in the previous experiments (Taneja et al., 2023a).

The variations in the electrical conductivity (EC) of electrolytes and soil are shown in Fig. 5(b). EC is a measurement of the migration of ions that conduct current across the soil column. The results showed higher EC values in electrolytes for experiments at voltage gradient 1.5 V/cm (R4, R5, R6). The results confirmed that this voltage gradient was able to overcome the resistance offered by soil to cause the movement of ions towards the oppositely charged electrodes. The EC in soil was significantly less than the electrolytes as most of the ions were transported to the electrode wells from the soil. Higher EC values in soil were observed for experiments R3, R6, and R9 due to the addition of EDTA at 0.2 M concentration which increased the ionic strength. It was observed that EC decreased from the anode towards the cathode with maximum values at S1 section for all the runs, except R4, in which a dip at S1 was observed. This could be due to the effective transfer of ions from the soil into the anode well. However, in experiments R5 and R6, EC was higher at the S4 soil section, which could be due to the accumulation of ions as hydroxides causing the focussing zone near the cathode, as also reported in previous studies (Li et al., 2012; Wu et al., 2021).

### 3.4 Electro-osmotic Volume (EOV) and Moisture Content

The soil pore fluid plays a crucial role in facilitating contaminant removal. During EKR, the applied voltage causes the movement of soil pore fluid through the process of electro-osmosis, which can transport the contaminants toward the electrode wells (Cameselle & Pena, 2016). To investigate the influence of voltage and EDTA on electro-osmosis, the electro-osmotic volume (EOV) was measured as the cumulative catholyte collected in the cathode well after the 2-day experiment (Fig. 6). The results revealed that as the voltage gradient increased from 0.5 to 2.5 V/cm, the EOV increased from 158 to 218 mL. This could be attributed to the low current generation and subsequently high electrical resistance by soil at a low voltage gradient, resulting in less EOV (Cameselle et al., 2021). At the same voltage gradient, as EDTA concentration increased to 0.1 M, EOV increased, however, at 0.2 M EDTA concentration, EOV decreased. The results suggest that an increase in the ionic strength decreases the EOV which could be due to the compression of double

**Table 4** Regression analysis and ANOVA for the response removal efficiency

| Source | Sum of Squares | df | Mean Square | F-value | *p*-value | |
|---|---|---|---|---|---|---|
| Model | 741.19 | 5 | 148.24 | 49.03 | 0.0045 | **Significant** |
| A | 568.43 | 1 | 568.43 | 187.99 | 0.0008 | |
| B | 18.73 | 1 | 18.73 | 6.19 | 0.0886 | |
| AB | 7.84 | 1 | 7.84 | 2.59 | 0.2057 | |
| $A^2$ | 130.14 | 1 | 130.14 | 43.04 | 0.0072 | |
| $B^2$ | 16.06 | 1 | 16.06 | 5.31 | 0.1046 | |
| Residual | 9.07 | 3 | 3.02 | | | |
| Cor Total | 750.26 | 8 | **$R^2$** | 0.98 | | |
| Std Dev | 1.74 | | **$R^2$ adjusted** | 0.96 | | |
| Mean | 12.64 | | **$R^2$ predicted** | 0.86 | | |
| C.V % | 13.75 | | **Adeq precision** | 16.19 | | |

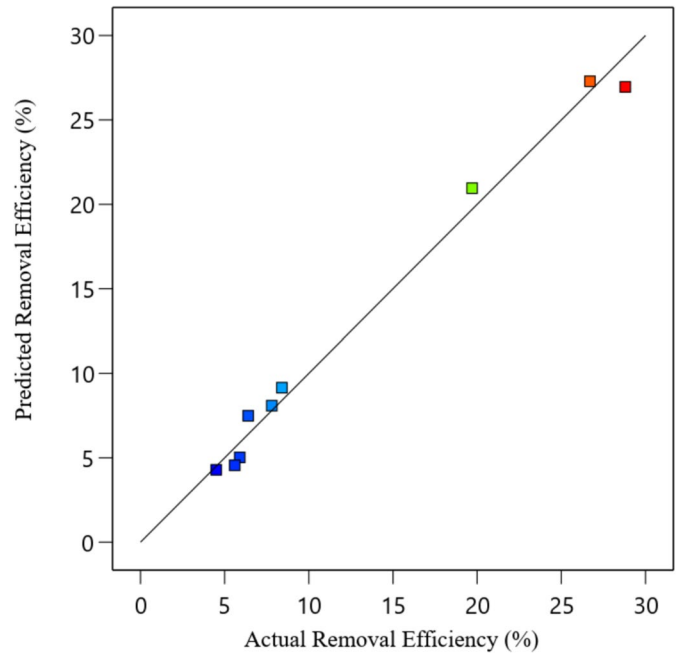



**Fig. 2** Plot of predicted values from the model against the values obtained from experimental results for Removal Efficiency (%)

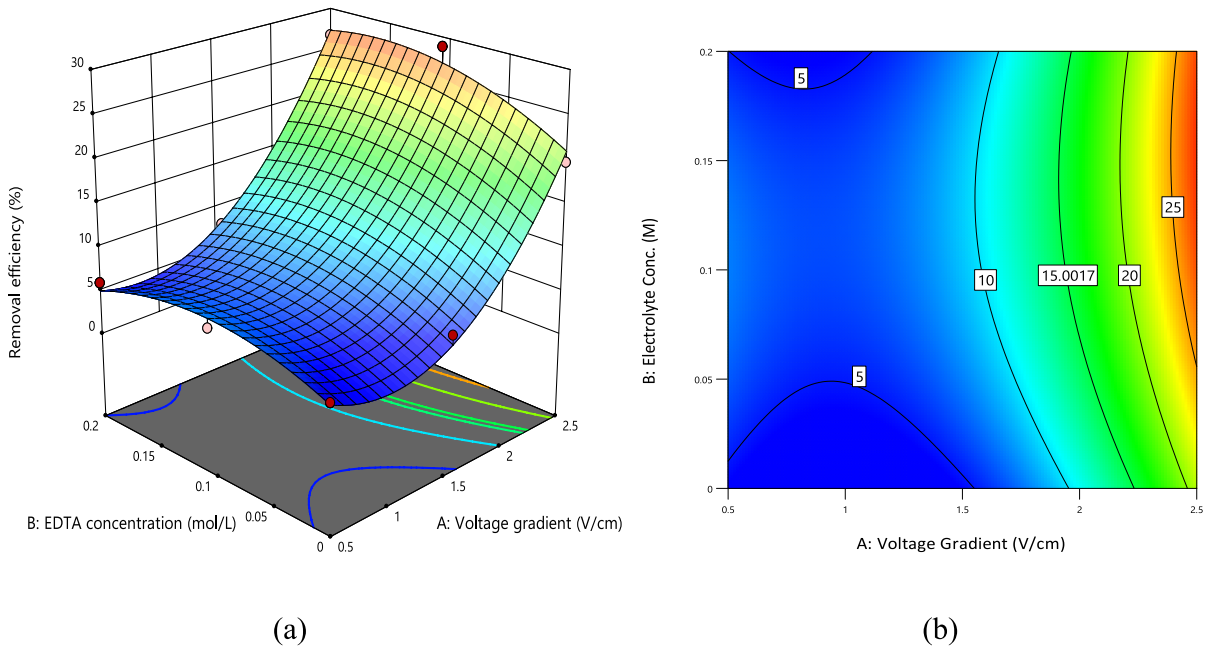



**Fig. 3** (**a**) 3-D Surface response plot and (**b**) Contour plot for Removal Efficiency (%)

diffuse layer thickness, which was also observed by Ng et al., (2016).

The soil moisture content (%) is another important factor measured to understand the movement of soil pore fluid and water loss during EKR. The variations in moisture content in soil sections are shown in Fig. 6. The experiments with no EDTA amendments showed a general trend of increasing moisture from anode to cathode (S1 to S4). As the voltage increased, the moisture content decreased significantly. The

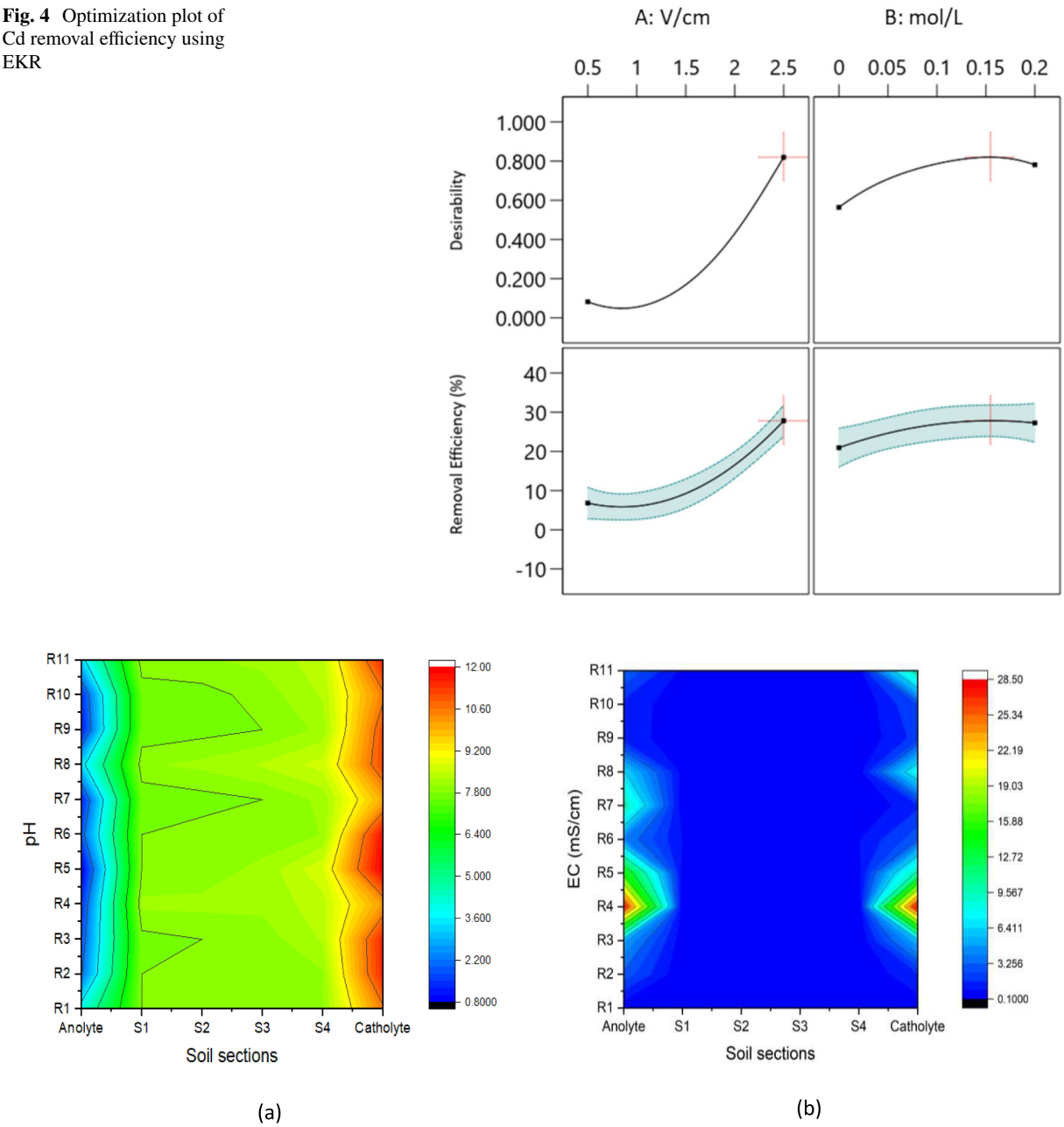



**Fig. 4** Optimization plot of Cd removal efficiency using EKR

**Fig. 5** Contour plots of (**a**) pH and (**b**) electrical conductivity (mS/cm) in electrolytes and soil sections after EKR

possible reasons could be (a) the loss of water due to the heat generated at higher voltage gradient (b) the increased electro-osmosis due to which the pore fluid accumulated in the electrode wells. Upon the addition of 0.1 M EDTA, the moisture content was found to be higher in soil sections S1 and S4, that is the soil sections close to the wells. As the EDTA concentration increased to 0.2 M, the trend shifted to cathode towards anode. This could be due to the drag force of Cd-EDTA complexes on the soil pore fluid towards the anode. Overall, the water loss was observed in all experiments, specifically at higher voltages. The loss

of water induces soil compaction that may lead to trapping or accumulation of metal in the soil, preventing its removal (Li et al., 2020; Reddy et al., 2002).

### 3.5 Distribution of Residual Cd in Soil

Figure 7 exhibits the normalized residual concentration of Cd in soil sections for experiments R1 to R11 and their corresponding removal efficiency. It was found that experiments R1 to R6 had a removal rate below 20%, suggesting that the voltage gradient of up to 1.5 V/cm was not sufficient to overcome the electrical resistance to cause the migration of Cd. At the constant voltage, amendments with EDTA could only improve the removal rate from 5.5 to 8.4%. From R7 onwards, the removal efficiency increased substantially to attain a peak at 28.8% (R8), which indicated that increasing the voltage gradient beyond 1.5 V/cm along with EDTA (0.1 M) could improve electromigration and electro-osmosis. However, the removal rate decreased to 26.7% (R9) upon further increasing the concentration of EDTA to 0.2 M at higher voltage. The use of stainless steel (SS) electrode (R11) as compared to graphite electrode (R8) under similar operating conditions, yield better removal (30%), reflecting the predominant nature of SS electrode for Cd removal, possibly due to increased surface area and less electro-deposition than graphite electrode.

The normalized residual concentration of Cd in soil post-EKR varied significantly across the experiments and the sampling zones. The initial experiments at a low voltage gradient (R1 to R4) exhibited higher values, indicating limited removal. In contrast, as the voltage gradient increased, a decrease in the residual concentration was observed. The variability in residual concentration across the soil sections displayed heterogeneous patterns at low voltage experiments. As the voltage gradient increased, the Cd accumulation shifted to the S2 and S3 sections. Conversely, S1 sections generally achieved better removal, potentially due to proximity to the anolyte well, where the Cd ions migrated. The antagonistic movement of $Cd^{2+}$ ions and $[Cd\text{-}EDTA]^{-}$ complexes in opposite directions resulted in the accumulation of Cd in the middle of soil sections, It could also be the reason for the overall low removal of Cd from the soil.

### 3.6 Species Distribution of Cd and EDTA in Soil

To understand the migration behaviour of Cd ions in the soil column, Visual Minteq software was employed. Figure 8(a) and (b) exhibit the chemical speciation of Cd in soil when the electrolytes used were distilled water and EDTA, respectively. When

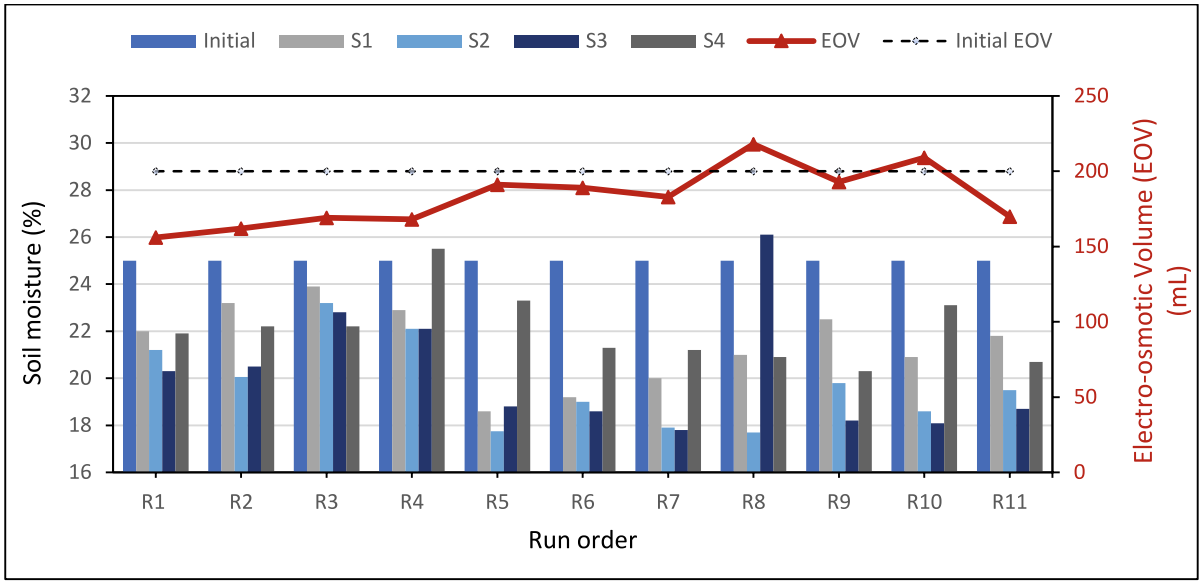



**Fig. 6** Moisture content (%) in different soil sections of all run orders and their corresponding electro-osmotic volume (EOV) after electrokinetic treatment

distilled water was used as the electrolyte (Fig. 8a), $Cd^{2+}$ ions dominated the system as a highly mobile ionic species over a wide range of pH (2–8) that tend to migrate towards the cathode. As the pH increased from 8 to 10, $Cd^{2+}$ began to precipitate as cadmium hydroxide species, mainly $Cd(OH)^{+}$, $Cd(OH)_2$, and $Cd(OH)_3$ with stability constants ($K_s$) 3.9, 7.7, and 10.3, respectively (Giannis et al., 2008), which clogged the soil pores near the cathode region (Eq. 11 and 12). Since the pH of the soil in this study was in the range of 7–8, relatively higher precipitation and surface complexation reactions led to the increased adsorption of $Cd^{2+}$ on soil and subsequently limited removal.

$$Cd^{2+} + H_2O \leftrightarrow CdOH^{+} + H^{+} \quad (11)$$

$$Cd^{2+} + 2H_2O \leftrightarrow (CdOH)_2 + 2H^{+} \quad (12)$$

EDTA is a tetracarboxylic acid and a strong chelating agent that can form soluble metal complexes over a wide range of pH (Cameselle et al., 2021). At low pH (below 2), Cd was predominantly present as $Cd^{2+}$ ions, while EDTA existed in its protonated forms, mainly $H_3EDTA^{-}$ ($K_s$ = 2.68) and $H_4EDTA$ ($K_s$ = 2.0), which do not have the ability to chelate. At higher pH (above 5), EDTA existed as deprotonated forms, $HEDTA^{3-}$ ($K_s$ = 10.17), and $H_2EDTA^{2-}$ ($K_s$ = 6.11), which poses the ability to chelate Cd and form negatively charged soluble complexes, [Cd-EDTA]$^{2-}$ ($K_s$ = 18.2) and [Cd-HEDTA]$^{-}$ ($K_s$ = 21.5), as given in Eq. 13 and 14. The higher thermodynamic stability of Cd-EDTA complexes as compared to Cd-hydroxide complexes led to reduced hydroxide formation of Cd, significantly improving the Cd sequestration in the aqueous phase even at higher pH, as also indicated in Fig. 8b.

$$Cd^{2+} + EDTA^{4-} \leftrightarrow Cd(EDTA)^{2-} \quad (13)$$

$$Cd^{2+} + HEDTA^{3-} \leftrightarrow Cd(HEDTA)^{-} \quad (14)$$

The mobility of Cd in soil is also influenced by the soil texture. Sandy loam soil used in this study was characterized by a higher fine sand fraction and a relatively low clay content. The higher sand

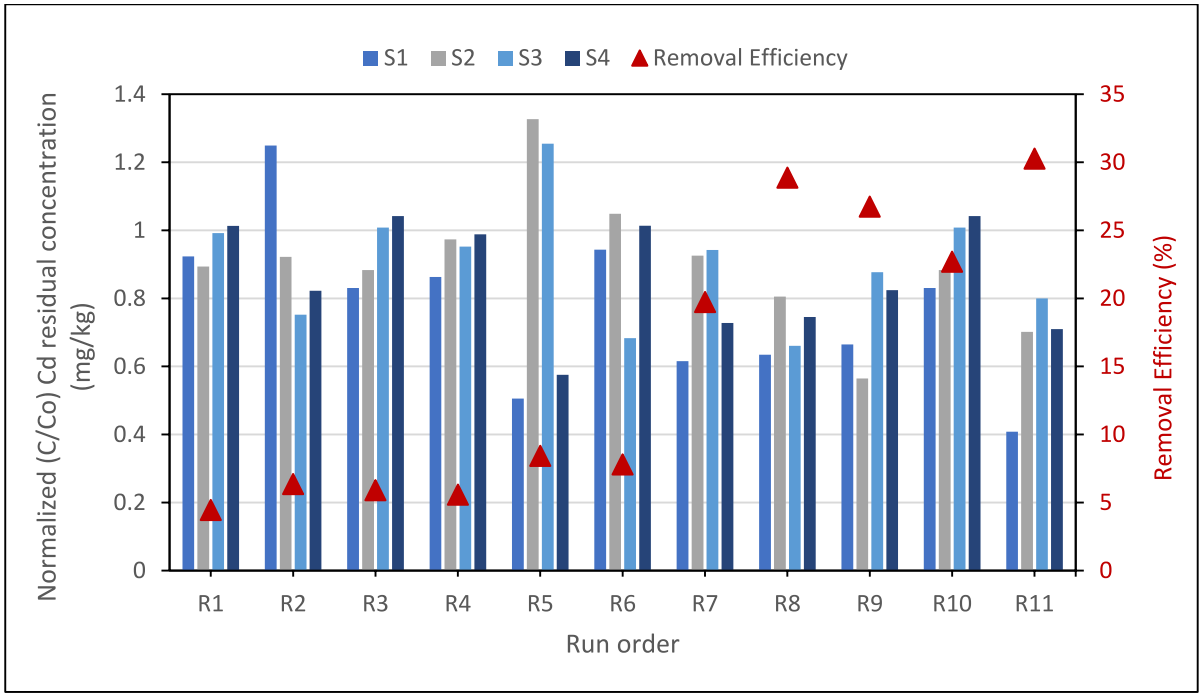
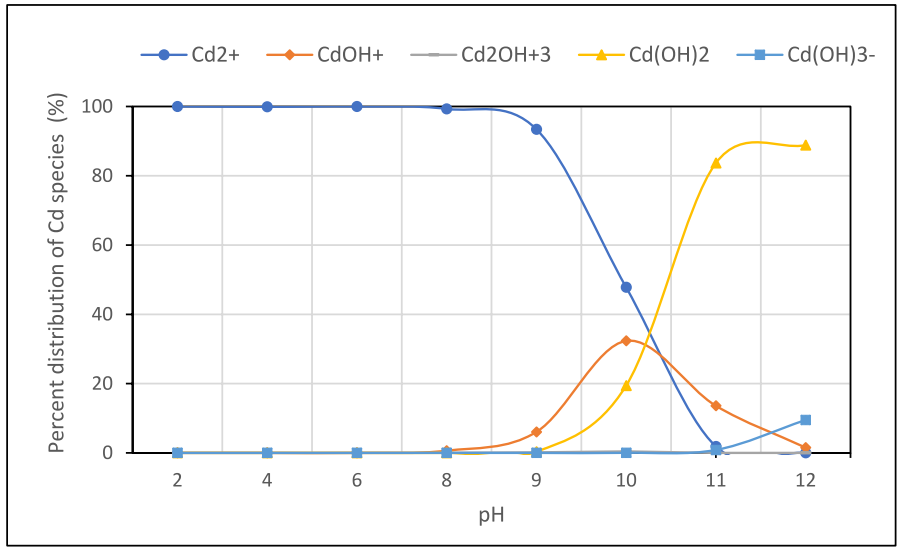
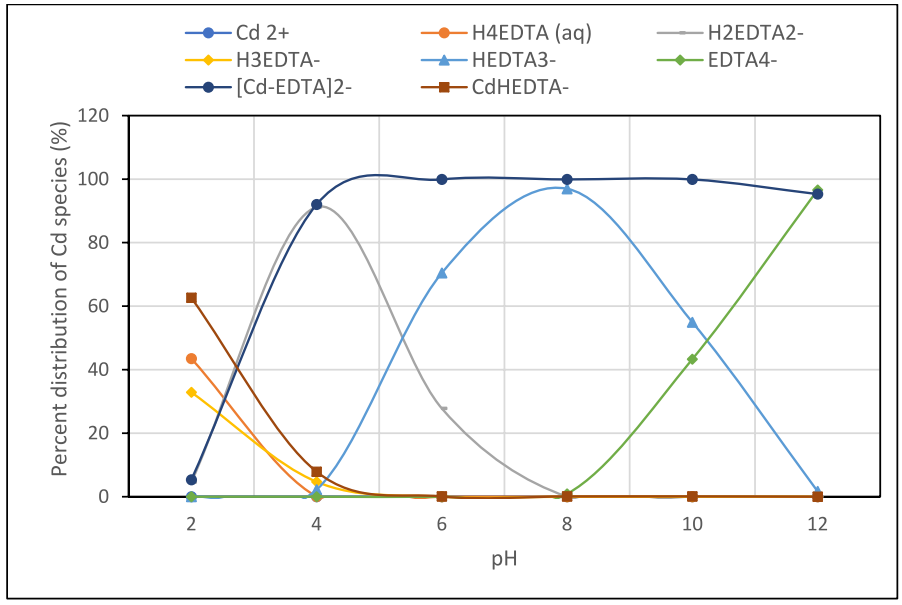



**Fig. 7** Normalized (C/Co) residual concentration (mg/kg) of Cd in soil after EKR treatment and removal efficiency of all experiments




(a)




(b)

**Fig. 8** Chemical speciation of Cd in soil under (**a**) distilled water as electrolyte and (**b**) EDTA as electrolyte using Visual Minteq Software

fraction contributes towards high permeability and fewer adsorption sites for Cd, thereby facilitating its mobility (Lai et al., 2023). However, the pH of soil had a more pronounced effect on Cd mobility. A high pH of 8.1 in the soil facilitated the formation of Cd hydroxide precipitates, which reduced the mobility and its subsequent migration. The alkaline conditions also enhanced the negative surface charge of soil particles, leading to strong adsorption. Nevertheless, when EDTA was used as the electrolyte, it was able to form soluble stable complexes with Cd even at high pH, preventing Cd precipitation. Even though the use of EDTA did solubilize Cd, a significant portion of metal remained within the soil system.

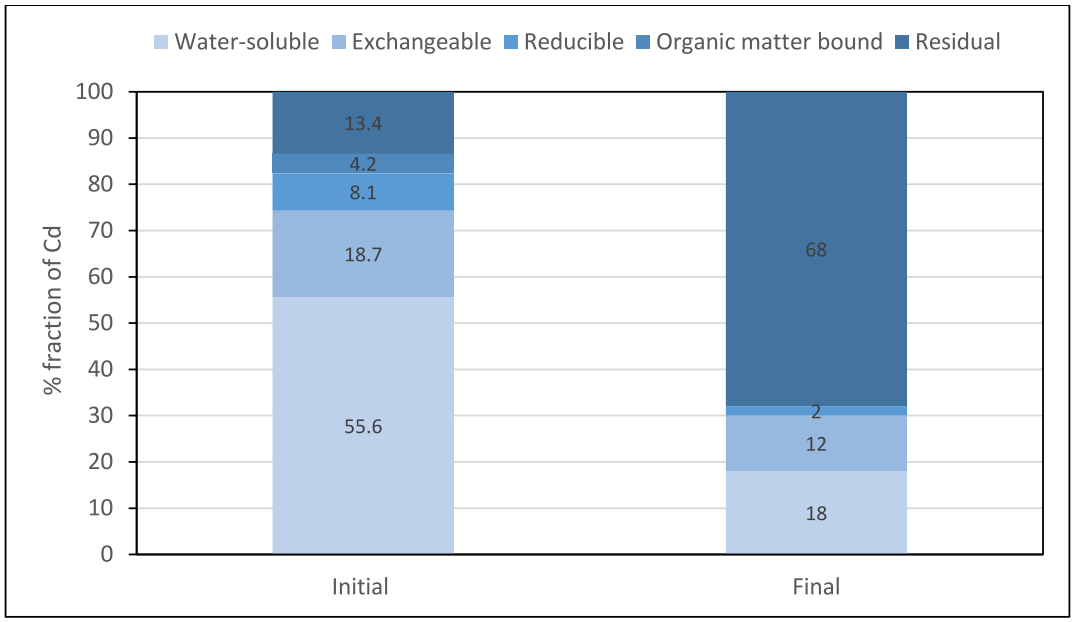



**Fig. 9** Fractionation of Cd in different forms before and after EKR using sequential extraction procedure

### 3.7 Toxicity Reduction and Health Risk Assessment

Cd is one of the most mobile elements among other heavy metals such as lead, nickel, copper, etc. However, in alkaline conditions, Cd binds with oxides of Fe or Mn, carbonates, or organic matter present in the soil, reducing mobility (Luan et al., 2022). To investigate the interaction of Cd with soil, a sequential extraction procedure was performed before and after EKR as shown in Fig. 9. It was observed that in the initial soil sample, the majority of Cd appeared to be in either water-soluble or exchangeable forms, while reducible and organic matter-bound fractions were low. Upon treatment with EKR, the water-soluble and exchangeable fractions significantly reduced, while reducible and residual fractions increased as Cd was immobilized into less bioavailable forms. The effective removal of bioavailable Cd and reduction of residual Cd into less bioavailable fractions confirmed that EKR was successful in reducing the overall toxicity of Cd in soil, which would further reduce the risk of groundwater infiltration of Cd in the field conditions.

The Human health risk assessment methodology was adopted to investigate the Cd-associated health risks to individuals, as shown in Tables 5 and 6. The non-carcinogenic adverse health effects were determined using the Hazard Index (HI) (Table 5). The primary route of exposure for Cd was found to be oral, followed by inhalation. In adults, the HI was less than 1, indicating no potential adverse effects. In children, the HI was reduced from 2 to 1.4, indicating a 29% reduction in risks post-EKR treatment, though still above the threshold value, posing potential adverse non-carcinogenic effects to children.

**Table 5** Average Daily Dose, Hazard quotient, and Hazard Index to assess non-carcinogenic adverse health effects

| Experiment | Cd (mg/kg) | $ADD_{ing}$ | $ADD_{inh}$ | $HQ_{ing}$ | $HQ_{inh}$ | HI |
|---|---|---|---|---|---|---|
| Adult | | | | | | |
| Untreated | 304.8 | 5.44E- 04 | 4.92E- 05 | 5.44E- 01 | 4.92E- 02 | 5.94E- 01 |
| Treated | 216.7 | 3.87E- 04 | 3.50E- 05 | 3.87E- 01 | 3.50E- 02 | 4.22E- 01 |
| Children | | | | | | |
| Untreated | 304.8 | 2.03E- 03 | 5.70E- 05 | 2.03E + 00 | 5.70E- 02 | 2.09E + 00 |
| Treated | 216.7 | 1.45E- 03 | 4.05E- 05 | 1.44E + 00 | 4.05E- 02 | 1.49E + 00 |

**Table 6** Total Carcinogenic Risk (TCR) in adults and children before and after EKR

| Experiment | Cd (mg/kg) | $CR_{ing}$ | $CR_{inh}$ | TCR |
|---|---|---|---|---|
| Adult | | | | |
| Untreated | 304.8 | 2.07E- 04 | 3.10E- 04 | 5.17E- 04 |
| Treated | 216.7 | 1.47E- 04 | 2.20E- 04 | 3.68E- 04 |
| Children | | | | |
| Untreated | 304.8 | 7.72E- 04 | 3.59E- 04 | 1.13E- 03 |
| Treated | 216.7 | 5.49E- 04 | 2.55E- 04 | 8.04E- 04 |

This could be attributed towards low body weight, weak immunity, and more susceptibility to exposure to contaminants in children as compared to adults. The total carcinogenic risk (TCR) for Cd is proportional to exposure concentration. For both adults and children, TCR was above the acceptable range of $10^{-6}$. However, the TCR reduced by 28% from one in a thousand ($10^{-3}$) to one in a hundred thousand ($10^{-4}$) post-treatment, though still above the acceptable limit (Table 6). The results reflect the potency of EKR to reduce available forms of Cd by removing from the soil system and immobilizing the residual Cd into less bioavailable forms, subsequently leading to reduced overall health risks.

### 3.8 Economic Evaluation

The economic evaluation of EKR for Cd removal is shown in Table 7. The energy consumption had a significant positive correlation with the voltage gradient with a maximum of 133.3 kWh/$m^3$ in R11, which subsequently increased the electricity cost. The total cost, however, also considered the cost of electrodes and chemicals used. Among the high voltage gradient experiments, R10 displayed low cost, while maximum cost was observed in R3. The specific cost, that is, the cost of experiment for the removal of one gram of Cd was calculated to determine the most cost-effective run. Although Cd removal was maximum in R11 and R8, R10 was found to be the most cost-effective (5.0 US$) due to the reduced use of chemicals and relatively cost-effective electrodes. It was found that experimental runs with SS electrodes performed better, both technologically and economically without compromising the structure of the electrode.

To compare the economic viability of EKR with conventional remediation methods such as soil washing and phytoremediation, several factors need to be considered. Soil washing has excellent removal efficiency. However, the additional cost of excavation, transportation, chemicals etc. (100–200 US$/ton soil) (Mulligan et al., 2001) makes it less cost-efficient. Phytoremediation, on the other hand, is an excellent eco-friendly method with low cost (12–15 US$/ton) (da S Trentin et al., 2019). Nonetheless, the long treatment duration and site selectivity with less contamination limit its large-scale application. EKR, being an *in-situ* technique, can minimize the transportation and chemical costs (150–180 US$/ton) (Akansha et al., 2024). Additionally, integrating EKR with phytoremediation

**Table 7** Energy consumption (kWh/$m^3$) and cost estimation (US$) of EKR for the removal of Cd from soil

| Run | Voltage (V) | Current (A) | Time (h) | Energy Consumption (kWh/$m^3$) | Cost (₹) | | | Total Cost | | Cost/gm Cd removal | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Electricity | Electrodes | Chemicals | (₹) | (US$) | (₹/g) | (US$/g) |
| R1 | 10 | 0.01 | 10 | 0.7 | 0.003 | 88.9 | 0 | 88.9 | 1.1 | 8156.2 | 97.9 |
| R2 | 10 | 0.02 | 10 | 1.3 | 0.006 | 88.9 | 28.6 | 117.5 | 1.4 | 5657.5 | 67.9 |
| R3 | 10 | 0.06 | 10 | 4 | 0.018 | 88.9 | 57.2 | 146.1 | 1.8 | 10,446.7 | 125.4 |
| R4 | 30 | 0.04 | 10 | 8 | 0.036 | 88.9 | 0 | 88.9 | 1.1 | 5200.9 | 62.4 |
| R5 | 30 | 0.12 | 10 | 24 | 0.108 | 88.9 | 28.6 | 117.6 | 1.4 | 4535.4 | 54.4 |
| R6 | 30 | 0.08 | 10 | 16 | 0.072 | 88.9 | 57.2 | 146.2 | 1.8 | 5984.5 | 71.8 |
| R7 | 50 | 0.1 | 10 | 33.3 | 0.15 | 88.9 | 0 | 89.1 | 1.1 | 1446.5 | 17.4 |
| R8 | 50 | 0.38 | 10 | 126.7 | 0.57 | 88.9 | 28.6 | 118.1 | 1.4 | 1341.4 | 16.1 |
| R9 | 50 | 0.2 | 10 | 66.7 | 0.3 | 88.9 | 57.2 | 146.4 | 1.8 | 1913.7 | 23.0 |
| R10 | 50 | 0.07 | 10 | 23.3 | 0.105 | 25 | 0 | 25.1 | 0.3 | 413.9 | 5.0 |
| R11 | 50 | 0.4 | 10 | 133.3 | 0.6 | 25 | 28.6 | 54.2 | 0.7 | 484.5 | 5.8 |

**Table 8** Summary of published literature on Cd removal by EKR and results from the present study

| S.No | Initial Cd (mg/kg) | Time (hours) | Electrolyte | Power supply | Amendments | Removal Efficiency (%) | References |
|---|---|---|---|---|---|---|---|
| 1 | 294 | 130 | 0.1 M EDTA | 1 V/cm | EDTA | 15.1 | Gu et al., 2018 |
| 2 | - | 192 | 0.5 M lactic acid | 1 V/cm | Power breaking | 38 | Cai et al., 2022 |
| 3 | 80 | 96 | CA + $CaCl_2$ | 2 V/cm | addition of $CaCl_2$ | 22.5 | Lai et al., 2023 |
| 4 | 0.86 | 384 | 0.3 M CA | 2 V/cm | Polarity reversal | 59 | Wang et al., 2024 |
| 5 | 20 | 240 | 0.1 M EDTA | 2.6 $A/m^2$ | EDTA | 40.4 | Xu et al., 2021 |
| 6 | 250 | 10 | 0.1 M EDTA | 2.5 V/cm | EDTA + periodic voltage | 30 | *Present study* |

may offer a promising approach for the remediation of highly contaminated sites. EKR can significantly bring down the contamination levels in less duration, subsequently allowing phytoremediation to stabilize and restore the remediated site at low costs.

### 3.9 Comparison with Previously Published Literature

Table 8 summarizes the previously published studies on Cd removal using EKR and the results from the present study. The table highlights that the removal efficiency is not only influenced by the voltage and electrolyte used, but also dependent on factors like initial metal concentration, and the duration of treatment. The present study employed EDTA amendment with a high voltage gradient of 2.5 V/cm to achieve a moderate removal efficiency of 30% in 10 h. While the efficiency is slightly less compared to other studies, the results showed an interesting insight into reducing the treatment duration significantly by enhancing ion mobility with periodic voltage application. Thus, the practical applicability of EKR shows great potential for rapid remediation.

## 4 Conclusion

The study investigated the effectiveness of EKR to remove Cd from soil under the influence of varying voltage gradients and EDTA concentrations using RSM methodology. The experimental results, predicted values from the model and the regression analysis suggested that the quadratic model developed was reliable. The influence of voltage gradient was significantly dominant over electrolyte concentration. The optimized conditions were found to be 2.5 V/cm voltage gradient and 0.15 M EDTA concentration to achieve 27.8% Cd removal in 10 h of treatment duration. The study concluded that stainless steel electrodes were superior to graphite in removing Cd in a technologically and economically effective manner. The overall Cd removal was relatively less due to less duration and fewer amendments. Nevertheless, upon extrapolating the removal rate of this study with respect to the duration, it is expected that the removal may increase two-fold with 3 cycles of operation. The findings revealed that EKR can yield significant metal removal when the initial concentration is high and can be employed for the treatment of highly contaminated sites as a rapid remediation method, which would bring down the overall toxicity and associated health risks.

**Acknowledgments** The authors are thankful to Delhi Technological University for providing the necessary facilities and support to carry out this research.

**Authors' Contribution** **Sonam Taneja:** Conceptualisation; Investigation; Writing-Original draft.
**Oznur Karaca:** Supervision; Writing-Reviewing and Editing.
**Anil Kumar Haritash:** Methodology; Supervision; Writing-Reviewing and Editing.

**Funding** No funding was received for conducting this study.

**Data Availability** All data used or analysed during this study is included in this article.

**Declarations**

**Ethics Approval and Consent to Participate** Not Applicable.

**Consent to Participate** Not Applicable.

**Consent for Publication** Not Applicable.

**Conflict of interest** The authors have no competing interests to declare that are relevant to the content of this article.

## References


Acar, Y. B., Gale, R. J., Alshawabkeh, A. N., Marks, R. E., Puppala, S., Bricka, M., & Parker, R. (1995). Electrokinetic remediation: Basics and technology status. *Journal of Hazardous Materials, 40*(2), 117–137. https://doi.org/10.1016/0304-3894(94)00066-P

Adimalla, N., Chen, J., & Qian, H. (2020). Spatial characteristics of heavy metal contamination and potential human health risk assessment of urban soils: A case study from an urban region of South India. *Ecotoxicology and Environmental Safety, 194*, 110406.

Akansha, J., Thakur, S., Chaithanya, M. S., Gupta, B. S., Das, S., Das, B., Rajasekar, N., & Priya, K. (2024). Technological and economic analysis of electrokinetic remediation of contaminated soil: A global perspective and its application in Indian scenario. *Heliyon, 10*(2), e24293. https://doi.org/10.1016/j.heliyon.2024.e24293

Asadollahfardi, G., Sarmadi, M. S., Rezaee, M., Khodadadi-Darban, A., Yazdani, M., & Paz-Garcia, J. M. (2021). Comparison of different extracting agents for the recovery of Pb and Zn through electrokinetic remediation of mine tailings. *Journal of Environmental Management, 279*, 111728. https://doi.org/10.1016/j.jenvman.2020.111728

Bolan, N., Kunhikrishnan, A., Thangarajan, R., Kumpiene, J., Park, J., Makino, T., Kirkham, M. B., & Scheckel, K. (2014). Remediation of heavy metal(loid)s contaminated soils – To mobilize or to immobilize? *Journal of Hazardous Materials, 266*, 141–166. https://doi.org/10.1016/j.jhazmat.2013.12.018

Cai, Z., Sun, Y., Deng, Y., Zheng, X., Sun, S., Sinkkonen, A., & Romantschuk, M. (2022). Enhanced Electrokinetic Remediation of Cadmium (Cd)-Contaminated Soil with Interval Power Breaking. *International Journal of Environmental Research, 16*(3), 31. https://doi.org/10.1007/s41742-022-00409-6

Cameselle, C., & Pena, A. (2016). Enhanced electromigration and electro-osmosis for the remediation of an agricultural soil contaminated with multiple heavy metals. *Process Safety and Environmental Protection, 104*, 209–217. https://doi.org/10.1016/j.psep.2016.09.002

Cameselle, C., Gouveia, S., & Cabo, A. (2021). Enhanced electrokinetic remediation for the removal of heavy metals from contaminated soils. *Applied Sciences, 11*(4), 4. https://doi.org/10.3390/app11041799

da S Trentin, A. W., Reddy, K. R., Kumar, G., Chetri, J. K., & Thomé, A. (2019). Quantitative Assessment of Life Cycle Sustainability (QUALICS): Framework and its application to assess electrokinetic remediation. *Chemosphere, 230*, 92–106. https://doi.org/10.1016/j.chemosphere.2019.04.200

FAO U. (2021). Global assessment of soil pollution: Report. *FAO and UNEP*. https://doi.org/10.4060/cb4894en

Ge, X., Song, X., Xie, J., Huang, W., Yuhui, W., Xu, Z., Wang, Y., Hou, X., & Cao, X. (2022). Comparison of the influence of EDTA, nitrilotriacetic acid, diethylenetriamine pentaacetic acid, chitosan, fulvic acid and pine needle extract on the removal of multiple heavy metal by electrokinetic remediation. *International Journal of Electrochemical Science, 17*(10), 221012. https://doi.org/10.20964/2022.10.08

Genchi, G., Sinicropi, M. S., Lauria, G., Carocci, A., & Catalano, A. (2020). The Effects of Cadmium Toxicity. *International Journal of Environmental Research and Public Health, 17*(11), 11. https://doi.org/10.3390/ijerph17113782

Giannis, A., Gidarakos, E., & Skouta, A. (2008). Transport of cadmium and assessment of phytotoxicity after electrokinetic remediation. *Journal of Environmental Management, 86*(3), 535–544. https://doi.org/10.1016/j.jenvman.2006.12.003

Gu, Y., Yeung, A. T., & Li, H. (2018). Enhanced electrokinetic remediation of cadmium-contaminated natural clay using organophosphonates in comparison with EDTA. *Chinese Journal of Chemical Engineering, 26*(5), 1152–1159. https://doi.org/10.1016/j.cjche.2017.10.012

Jawad, H. M., & Jahanger, Z. K. (2024). The Effect of Oil Contaminated on Collapse Pattern in Gypseous Soil Using Particle Image Velocimetry and Simulation. *Civil Engineering Journal, 10*(7), 7. https://doi.org/10.28991/CEJ-2024-010-07-016

Jawjit, S., Narom, N., & Thongkaow, P. (2024). Evaluating Household Hazardous Waste Generation, Composition, and Health Risks in an Urban Municipality. *Journal of Human, Earth, and Future, 5*(3), 471–482.

Khalid, S., Shahid, M., Niazi, N. K., Murtaza, B., Bibi, I., & Dumat, C. (2017). A comparison of technologies for remediation of heavy metal contaminated soils. *Journal of Geochemical Exploration, 182*, 247–268. https://doi.org/10.1016/j.gexplo.2016.11.021

Kubier, A., Wilkin, R. T., & Pichler, T. (2019). Cadmium in soils and groundwater: A review. *Applied Geochemistry, 108*, 104388. https://doi.org/10.1016/j.apgeochem.2019.104388

Lai, Z., Mei, Z., Hui, Z., ShaoNan, Z., XueQiang, Z., & Ping, L. (2023). Insights into the enhanced electrokinetic remediation of lead and cadmium contaminated soil with the composite electrolyte of citric acid and calcium chloride. *Separation Science and Technology, 58*(7), 1319–1330. https://doi.org/10.1080/01496395.2023.2189058

Li, G., Guo, S., Li, S., Zhang, L., & Wang, S. (2012). Comparison of approaching and fixed anodes for avoiding the 'focusing' effect during electrokinetic remediation of chromium-contaminated soil. *Chemical Engineering Journal, 203*, 231–238. https://doi.org/10.1016/j.cej.2012.07.008

Li, X., Yang, Z., He, X., & Liu, Y. (2020). Optimization analysis and mechanism exploration on the removal of cadmium from contaminated soil by electrokinetic remediation. *Separation and Purification Technology, 250*, 117180.

Luan, Y., Xu, J., Zhou, J., Wang, H., Han, F., Wang, K., & Lv, Y. (2022). Migration and Removal of Labile Cadmium Contaminants in Paddy Soils by Electrokinetic Remediation without Changing Soil pH. *International Journal of Environmental Research and Public Health, 19*(7), 7. https://doi.org/10.3390/ijerph19073812

Mulligan, C. N., Yong, R. N., & Gibbs, B. F. (2001). Remediation technologies for metal-contaminated soils and groundwater: An evaluation. *Engineering Geology, 60*(1), 193–207. https://doi.org/10.1016/S0013-7952(00)00101-0

Ng, Y.-S., Sen Gupta, B., & Hashim, M. A. (2016). Remediation of Pb/Cr co-contaminated soil using electrokinetic

process and approaching electrode technique. *Environmental Science and Pollution Research, 23*(1), 546–555. https://doi.org/10.1007/s11356-015-5290-0

Organization (WHO), W. H. (1993). *Standard maxima for metals in Agricultural soils*. WHO.

Pandey, B. K., Shukla, C., Sillanpää, M., & Shukla, S. K. (2023). A systematic review on application of electrokinetics in stabilization and remediation of problematic soils. *Innovative Infrastructure Solutions, 8*(9), 226. https://doi.org/10.1007/s41062-023-01186-7

Peng, G., & Tian, G. (2010). Using electrode electrolytes to enhance electrokinetic removal of heavy metals from electroplating sludge. *Chemical Engineering Journal, 165*(2), 388–394. https://doi.org/10.1016/j.cej.2010.10.006

Reddy, K. R., Xu, C. Y., & Chinthamreddy, S. (2001). Assessment of electrokinetic removal of heavy metals from soils by sequential extraction analysis. *Journal of Hazardous Materials, 84*(2–3), 279–296.

Reddy, K. R., Saichek, R. E., Maturi, K., & Ala, P. (2002). Effects of soil moisture and heavy metal concentrations on electrokinetic remediation. *Indian Geotechnical Journal, 32*(2), 258–288.

Saeed, K., Al-Khyat, S., Abd Hacheem, Z., & Fartosy, S. H. (2024). Evaluating the efficiency of alkaline activator with silica-rich wastes in stabilizing cadmium-contaminated soil. *Civil Engineering Journal, 10*(7), 2123–2143.

Song, Y., Benamar, A., Mezazigh, S., & Wang, H. (2018). Citric Acid-Enhanced Electroremediation of Toxic Metal-Contaminated Dredged Sediments: Effect of Open/Closed Orifice Condition. *Electric Potential and Surfactant. Pedosphere, 28*(1), 35–43. https://doi.org/10.1016/S1002-0160(18)60003-7

Taneja, S., Karaca, O., & Haritash, A. K. (2023a). Combined effects of high voltage gradient and electrolyte conditioning on electrokinetic remediation for chromium (VI)-contaminated soils. *Rendiconti Lincei. Scienze Fisiche e Naturali, 34*(2), 635–646. https://doi.org/10.1007/s12210-023-01159-z

Taneja, S., Karaca, O., & Haritash, A. K. (2023b). Treatment of Pb-contaminated soil by electrokinetics: Enhancements by varying voltage, chelant, and electrode material. *Journal of Geochemical Exploration, 250*, 107240. https://doi.org/10.1016/j.gexplo.2023.107240

Taneja, S., Karaca, O., & Haritash, A. K. (2024). Electrokinetic remediation: Past experiences and future roadmap for sustainable remediation of metal-contaminated soils. *Journal of Geochemical Exploration, 259*, 107437. https://doi.org/10.1016/j.gexplo.2024.107437

Taneja, S., Yadav, S., Pipil, H., Karaca, O., & Haritash, A. K. (2023). Soil–Water Interactions and Arsenic Enrichment in Groundwater. In *Hydrogeochemistry of Aquatic Ecosystems* (pp. 97–120). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119870562.ch5

Tsui, L., Paul, A., Chen, Y.-T., & Tz-Chi, E. (2022). Potential mechanisms contributing to the high cadmium removal efficiency from contaminated soil by using effective microorganisms as novel electrolyte in electrokinetic remediation applications. *Environmental Research, 215*, 114239. https://doi.org/10.1016/j.envres.2022.114239

Wang, Z., Tang, T., Chen, L., Wu, S., Li, X., & Liu, Y. (2024). Electrokinetic remediation of cadmium-contaminated soil using polarity reversal method: Optimization analysis and mechanism exploration. *Journal of Environmental Management, 370*, 122732. https://doi.org/10.1016/j.jenvman.2024.122732

Wcisło, E., Bronder, J., Rodríguez-Valdés, E., & Gallego, J. L. (2022). Health Risk Assessment of Post-mining Hg-As-Contaminated Soil: Implications for Land Remediation. *Water, Air, & Soil Pollution*, *233*. https://doi.org/10.1007/s11270-022-05712-8

Wu, J., Wei, B., Lv, Z., & Fu, Y. (2021). To improve the performance of focusing phenomenon related to energy consumption and removal efficiency in electrokinetic remediation of Cr-contaminated soil. *Separation and Purification Technology, 272*, 118882. https://doi.org/10.1016/j.seppur.2021.118882

Xu, Y., Lu, Q., Li, J., Wan, L., Chen, S., & Lu, Y. (2021). Effect of humus on the remediation of arsenic-contaminated soil by electrokinetic technology. *Environmental Technology & Innovation, 21*, 101297. https://doi.org/10.1016/j.eti.2020.101297

Yeung, A. T., Hsu, C., & Menon, R. M. (1997). Physicochemical soil-contaminant interactions during electrokinetic extraction. *Journal of Hazardous Materials, 55*(1), 221–237. https://doi.org/10.1016/S0304-3894(97)00017-4

Zhang, P., Jin, C., Sun, Z., Huang, G., & She, Z. (2016). Assessment of Acid Enhancement Schemes for Electrokinetic Remediation of Cd/Pb Contaminated Soil. *Water, Air, & Soil Pollution, 227*(6), 217. https://doi.org/10.1007/s11270-016-2879-y

**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.

OXFORD

*The Computer Journal*, 2025, 1–38
https://doi.org/10.1093/comjnl/bxaf045
**Original Article**

# Towards secure healthcare: SA-GBO-ODBN model utilizing Blockchain and deep learning for data handling and diagnosis

Nikhil Sharma and Prashant Giridhar Shambharkar*

Department of Computer Science and Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur Village, Rohini, New Delhi-110042, India
*Corresponding author. Department of Computer Science and Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur Village, Rohini, New Delhi-110042, India. E-mail: prashant.shambharkar@dtu.ac.in

## Abstract

The integration of Electronic Health Records (EHRs) in healthcare has significantly advanced the field but introduced challenges related to data security and precise diagnoses due to the large data volume. To address these issues, we propose the SA-GBO-ODBN model, combining Blockchain and deep learning (DL) for secure medical data management and diagnostics. This model includes Hyperledger Fabric for tamper-proof storage, optimal key generation, data encryption and decryption, and disease detection functionalities. The Key features of the proposed framework include emergency contact notifications, user data access management, and administrative data modifications. The framework employs SHA-256 and elliptical curve cryptography (ECC) for enhanced data security. ECC uses the Self-Adaptive Gradient-Based Optimizer (SA-GBO) to generate optimal encryption and decryption keys. Hyperledger blockchain technology enables secure medical data sharing, patient visit data storage, and EHR link recording in external databases via multiple channels. After decryption, the Optimized Deep Belief Network–based approach diagnoses epilepsy using real-time EEG datasets. The qualitative and quantitative performance analysis shows the proposed framework's superiority over existing techniques, with accuracy, False Positive Rate (FPR), and FNR of 98.93%, 0.0199, and 0.0034 for the Bonn EEG dataset, and 99.40%, 0.0196, and 0.0034 for the CHB-MIT dataset, respectively.

**Keywords:** Blockchain; optimized deep belief network (ODBN); deep learning (DL); electronic health records (EHRs); gradient-based optimizer (GBO); security

## 1. INTRODUCTION

Recently, the rise in chronic diseases and the ongoing COVID-19 pandemic have led to a marked increase in people's health awareness. Electronic Health Records (EHRs) are common in the real world and can be shared as vital healthcare data resources. The expanding amount of data consumes a considerable large amount of local resources, which are restricted by the capabilities of devices in the nearby area [1]. Hospitals could keep their EHRs in their cloud to share the data between patients, doctors, and suppliers [2]. However, the healthcare system (HS) nodes are consistently linked via an open channel, exposing the entire network to potential eavesdropping, data manipulation, and other security-related concerns. The rapid advancement of hacking techniques and communication protocols has heightened security concerns. Attackers are increasingly sophisticated in their efforts to breach systems, aiming to compromise the integrity, availability, confidentiality, and reliability of data. This evolving threat landscape underscores the urgency of addressing these security vulnerabilities [3, 4]. These attacks can target healthcare network components with malicious software or malware and reduce the functionality of actual Internet of Things (IoT) devices or access the patient's data without their permission [5]. Most currently used medical data-sharing protocols use essential encryption features and fine-grained access control. EHRs are shared among many users [6]. ABE (Attribute-based encryption) links the encrypted message and a user's confidential key to a group of characteristics and authorization rules. Suppose patients' EHRs are encrypted utilizing traditional CP-ABE (Cipher Policy Attribute-based encryption). In that case, anyone with cloud servers can access the precise access policy data [7], which may indicate that the patient has cardiac problems, which would be unacceptable. This may result in devices within the network inefficiently using resources and experiencing disruptions in their typical communication patterns [8].

For the issues mentioned above, combining DL and Blockchain can offer an exceptional solution in the healthcare industry [9, 10]. A chain, a block, and a blockchain are types of open databases (a type of digital information) using which information cannot be changed once it has been saved in a series of immutable blocks [11]. Due to blockchain's decentralized and immutable characteristics, smart contracts (SC) can enhance trust among parties engaged in data transmission by promptly executing and enforcing their terms. Additionally, the consensus mechanisms guarantee the integrity of the blockchain's distributed records. As a result, it is feasible to assume that the patient's medical-related information is secure [12], reliable, and trustworthy while being transmitted over HS [13]. We develop a self-adaptive deep learning model with a secure blockchain model to address the abovementioned issues and attain safer and more effective data access control in the healthcare sector.

**Received:** April 28, 2024. **Revised:** February 7, 2025. **Accepted:** March 31, 2025
© The British Computer Society 2025. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

The exponential growth in EHRs has led to revolutionary opportunities in the healthcare industry. It transformed how data is managed and improved diagnostics capabilities. Despite granting data owners control over their information and sharing permissions, the large quantity of healthcare records presents significant challenges in guaranteeing precision and security. Due to the constant risk of data leakage [14, 15], encryption algorithms must be used to completely control the security of EHRs. When data security is compromised, maintaining the secrecy and precision of user data in EHRs becomes difficult. To overcome the issues allied with the increasing generation of EHRs in the healthcare sector, our research proposed the SA-GBO-ODBN model. This novel method fuses blockchain and DL technology to manage and diagnose secure medical information effectively. It is a flexible and efficient information security framework to protect EHR data. This diagnostic model enhances the proficiency and accuracy of disease diagnosis within the HS. The suggested framework comprises various techniques designed to strengthen medical records' availability and integrity. It ensures strong security measures to safeguard patients' sensitive data by using optimal key generation, Hyperledger-based data management, and advanced encryption algorithms.

### 1.1. Motivation

The motivation for this research is driven by the need to solve two major issues in healthcare: secure sharing of medical data and accurate disease diagnosis. Currently, patient medical records are often spread across various institutions, making it difficult to gather comprehensive datasets for research while maintaining patient privacy and effective diagnosis. This fragmentation slows the collection of large, representative samples essential for accurate medical insights. Additionally, there is a lack of transparency in how data is stored and used, increasing the risk of misuse, including fraud and other malicious activities. Blockchain and deep learning technologies present promising solutions to these problems. Blockchain's decentralized and transparent system and SC allow for secure and auditable data sharing among different entities. At the same time, deep learning models can improve disease diagnosis accuracy by using large and diverse datasets. However, integrating these technologies involves overcoming the complexities of smart contract logic and business processes to make them understandable and effective. This research uses blockchain and deep learning to securely store and share medical data, enhancing disease diagnosis accuracy. By using blockchain for secure data management and deep learning for precise medical analysis, the goal is to design a reliable framework that protects patient privacy, ensures data integrity, and promotes collaborative medical research. This approach addresses the urgent need for privacy-preserving data solutions and aims to improve healthcare outcomes using advanced technologies.

### 1.2. Major contribution

The major contributions of the paper are presented as follows:

- In this paper, a novel SA-GBO-OBDN model has been proposed that is used for effectively managing and diagnosing medical records by strengthening the transparency, security, and traceability of Healthcare data across distributed networks.
- This work presents a systematic blockchain-based framework for securely managing vast healthcare records using Hyperledger technology.
- The SA-GBO technique has been employed to generate optimal keys, enhancing security during the encryption and decryption process.
- The proposed security framework employs SHA-256 and the Elliptic Curve Cryptography (ECC) cryptographic algorithm to strengthen the security of sensitive healthcare records, thereby preventing data breaches and unauthorized access.
- The ODBN (Optimized Deep Belief Network) diagnostic model has been suggested to detect epilepsy disease using two standard real-time datasets, the BONN and CHB-MIT datasets, after the decryption process.
- We propose the integration of an Intrusion Detection System (IDS) to continuously monitor IoT network traffic, enabling real-time detection of security threats and anomalies. By proactively identifying and mitigating potential attacks, the IDS strengthens the overall security framework, minimizing risks and preventing malicious activities.
- The proposed framework's performance has been evaluated and compared to existing techniques to illustrate its effectiveness for diagnosing and managing healthcare data.
- The standard performance parameters have been used for error analysis, which thoroughly comprehends the model's performance across different dimensions and differs from those often used in existing research work.

***Structure:*** The rest of the paper is organized as follows: Section 2 discusses the related work proposed by different researchers using technologies like blockchain mechanisms and DL in the healthcare industry. Section 3 explains the role of blockchain in data security and discusses the proposed architecture framework. Section 4 illustrates the result achieved by the suggested approach and provides a detailed comparison with other existing techniques. Section 5 presents the conclusion of the paper.

## 2. RELATED WORK

This section explores the research conducted by various experts in the healthcare field using emerging technologies like blockchain, ML, and DL. Islam *et al.* [16] aim to enhance the accuracy of activity classification in videos with the assistance of fog or cloud computing-based blockchain framework for activity recognition and monitoring. The framework focuses on a process of multi-class cooperative categorization. Frames depicting different human activities are isolated to identify critical features, which are further analyzed to create an action vocabulary, enhancing precision and efficiency. An activity classification system utilizes a support vector machine (SVM) model based on the estimated framework.

Wang *et al.* [17] suggested an EHR sharing framework designed on blockchain that is safe and protects the privacy of patient records. The data requester can utilize the specified keyword provided by the data provider to structure the EHRs needed on the blockchain network. Upon obtaining consent from the data owner, they can accept the re-encryption ciphertext from the cloud server. The primary approach of the scheme involves utilizing conditional searchable encryption and proxy re-encryption to ensure data security, access control, and privacy preservation. Xu *et al.* [18] developed the HealthChain framework, an intelligent medical care system that employs blockchain mechanisms to safeguard the privacy of large-scale healthcare records. This system helps users securely store data gathered through IoT devices and receive medical recommendations from physicians. On the other hand, doctors can upload their diagnoses by examining the

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

patient's medical reports generated using IoT devices. Blockchain technology ensures that all data and diagnoses are immutable and cannot be tampered which minimizes the risk of medical disputes. This system uses a two-step process that separates data publishing from access control transactions. Additionally, it stores all encrypted data in the Interplanetary File System (IPFS) to reduce costs and maintain anonymity.

Wang *et al.* [19] suggested a Blockchain-enabled framework named GuardHealth that ensures the privacy and security of sensitive records in the healthcare industry. The system employs a decentralized approach to manage confidentiality, authentication, privacy, and data sharing, which has been attained by leveraging consortium SC and Blockchain technology, which secures data sharing and storage while thwarting unauthorized access. An advanced Graph Neural Network technique is also applied to detect malware nodes. Zhou *et al.* [20] developed a human-in-the-loop-assisted method (HITLM) that has been used to protect patient data confidentiality in smart healthcare systems (SHS). One method uses a block design format to hide various health-related readings of the patients generated using smart healthcare IoT devices or hospital equipment. This approach enables private access to medical information produced from healthcare monitoring systems.

Hossain *et al.* [21] suggested a DL-driven pathology detection framework for the SHS. The signals from a patient's EEG data can be employed to identify pathology. An intelligent EEG headset transmits EEG signals to a mobile edge computing server throughout the proposed framework. The server processes the signals before sending them to a cloud server. The system employs a DL technique that performs most of the processing and ascertains the presence of pathology in the patient. A cloud server authentication manager links the framework of clients and stakeholders. Hossein *et al.* [22] demonstrated that balancing transparency and access control is a significant issue in managing healthcare data while maintaining privacy. They proposed a solution to this problem by enabling direct owners to state access control policies for their records. The BC Health framework consists of two distinct chains for storing data, data transactions and access policies. They have used a clustering approach to ensure the BC Health architecture's scalability, low latency, and minimal overhead.

Zaabar *et al.* [23] discussed the use of blockchain mechanisms in order to bolster the security and privacy of EHRs. A new design was suggested to avoid issues related to centralized storage by utilizing decentralized databases. The patient EHRs were stored on a decentralized Orbit DB that leverages IPFS. A blockchain network utilizing Hyperledger Fabric was also implemented to utilize Hyperledger Composer for storing data hashes and managing data access during retrieval. This blockchain-driven framework was designed to advance the security of extant SHS and strengthen the flexibility of healthcare management systems (HMS). Zhang *et al.* [24] proposed a Blockchain-enabled access control framework with hidden policies offered as a critical component of the SHS. Multiple authorities were introduced into the scheme to prevent single-point failure. Online-offline encryption mode minimizes users' online data processing workload by assigning computation tasks to their idle time, while policy concealment preserves users' sensitive data. Decryption outsourcing between smartphones and mobile edge computing servers was supported by making appropriate transactions using blockchain technology and SC.

Lin *et al.* [25] introduced a blockchain-enabled proxy re-encryption technique for the e-HS that ensures data security. The proposed approach ensures data confidentiality by concealing the situation within the re-encryption key, preventing the proxy from gaining knowledge of the condition. The ciphertext-searching procedure is utilized by running the SC on the blockchain mechanisms to guarantee accurate and comprehensive findings. Ahmad *et al.* [26] proposed EHRs that uses fog layers to enhance the scalability and flexibility of IoT-based healthcare services, offering quick response times and low latency. Gupta *et al.* [27] proposed a framework that employs blockchain technology to securely manage and share critical health information among different stakeholders, addressing a significant challenge in the secure exchange of health data. Ahmed *et al.* [28] designed a blockchain and DL-based framework specifically for identifying the virus in radiological images, achieving a classification accuracy of 96%. Alruwaili *et al.* [29] proposed blockchain-based SHS that employed the Jellyfish Search Optimization with a Dual-Pathway Deep Convolutional Neural Network (JSO-DPCNN) approach for feature extraction and a multiplicative long short-term memory (MLSTM) method for disease detection. Rahal *et al.* [30] introduced a method of model ensembling that combines the weights of several local DL models to create a single global model. This global model can diagnose complex medical conditions more accurately than individual local models. Ahmad *et al.* [31] proposed an Explainable AI-based Computer-Aided Epileptic Seizure Detection System (XAI-CAESDs) designed to address the challenges in detecting epileptic seizures from EEG data. Saidi *et al.* [32] proposed a multi-classification Federated Learning (FL) method enhanced by blockchain technology to reinforce data security and privacy. Myrzashova *et al.* [33] proposed a blockchain-based framework that allows multiple hospitals to train and share encrypted federated models, uniting resources from small and large hospitals in distant areas, thus solving the data leakage issue. Table 1 depicts the different blockchain-based suggested framework and their pros and cons. The symbol ✔ shows that the proposed framework offers a particular feature, whereas the ✗ represents that the model does not meet the feature requirements.

The proposed method offers the following security requirements and features, which are explained as follows:

***i) Confidentiality:*** The proposed method employed a data encryption and decryption approach along with blockchain mechanisms to ensure the confidentiality of sensitive medical records. The Hyperledger fabric is used to build access control techniques, which protect confidentiality by permitting only authorized users to view and modify the data.

***ii) Integrity:*** Unifying ECC and SHA-256 algorithms improves data integrity and guarantees that data remains trustworthy and unaltered. Using blockchain technology, the Hyperledger fabric offers tamper-proof storage for medical data.

***iii) Availability:*** The decentralized nature of Blockchain guarantees high availability of medical records by eliminating single points of failure. Hyperledger blockchain technology provides multiple data storage and sharing channels that improve accessibility to medical data while preserving its availability.

***iv) Distributed:*** The distributed techniques improve data availability, resilience, and fault tolerance against system failures and malicious attacks. The proposed model employed distributed ledger technology, which enables the distribution and transmission of medical records among other participants.

***v) Decentralized:*** The proposed model employed Hyperledger fabric's distributed ledger technology, which permits users to store and transfer medical records in a distributed manner. The system's reliability is improved, and it is less vulnerable to single points of failure by spreading data among several network nodes.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 1. Different Blockchain and DL-based suggested framework with pros and cons

| Ref. | Methodology | Dataset | Pros | Cons | C | I | A | Di | De | Tp | Au | Sc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Islam *et al.* [16], 2019 | Blockchain, ML | Hollywood2 Action, UCF50 Sport Action, KTH Dataset | The authors designed a framework for recognizing human activities that incorporates blockchain technology and Fog-Cloud computing. | The authors did not provide an explanation of how blockchain has been implemented in the suggested framework, and no assessment of the performance of blockchain, fog, and Cloud platforms has been carried out. | ✗ | ✔ | ✔ | ✔ | ✔ | ✗ | ✔ | ✔ |
| Wang *et al.* [17], 2019 | Blockchain, EHRs | – | A blockchain technology designed for cloud-based medical care networks. | The discussion of privacy concerns in blockchain transactions has been dealt with. | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✗ |
| Xu *et al.* [18], 2019 | Blockchain, IoT | – | A Hybrid Blockchain framework is implemented to facilitate access control and ensure privacy preservation throughout the storage of data in the IPFS storage system. | The information can still be retrieved by the doctor, even though the patient has the right to remove the permission granted to the doctor by publishing a new Internet of Things key to the user-chain. | ✔ | ✔ | ✗ | ✔ | ✔ | ✔ | ✔ | ✗ |
| Wang *et al.* [19], 2020 | Blockchain | – | The GuardHealth design utilizes a consortium blockchain to facilitate data sharing while ensuring authentication, confidentiality, and efficient data preservation. | The proposed model has limited discussion on the feasibility, scalability, and security analysis. | ✔ | ✗ | ✗ | ✔ | ✔ | ✔ | ✔ | ✗ |
| Zhou *et al.* [20], 2020 | Deep Learning, Cloud Computing, Blockchain | – | A proposed SHS for the recognition of irregularities in electroencephalogram (EEG) signals. The system facilitates the incorporation of MEC, cloud, Software Defined Networks (SDN), Network Slicing (NS), and a DL technologies in the shape of a Convolutional Neural Network (CNN) model. | The author emphasizes the significance of resolving trust, security, and privacy concerns in SHS but does not mention how the proposed scheme mitigates these challenges. | ✔ | ✗ | ✗ | ✗ | ✗ | ✗ | ✔ | ✗ |
| Hossain *et al.* [21], 2020 | Blockchain, AI, IoT | TUH abnormal EEG Corpus v2.0.0 | Utilized the MEC's location awareness feature to alert medical professionals in close proximity to patients. Transmitted compressed data to the cloud storage. | The experimental sample size is comparatively small. | ✗ | ✗ | ✗ | ✔ | ✗ | ✗ | ✔ | ✗ |
| Hossein *et al.* [22], 2021 | Blockchain | – | i) The authors suggested a blockchain enabled framework that empowers data owners to choose their preferred access permissions for their privacy-sensitive healthcare records. The framework employed a dual-chain system, with one chain dedicated to storing data transactions and another chain dedicated to storing access regulations. ii) Utilize clustering strategies in order to advance the scalability of the framework. | The model required more time to retrieve the data. | ✔ | ✔ | ✗ | ✔ | ✔ | ✔ | ✗ | ✔ |

*(continued)*

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 1. Continued.

| Ref. | Methodology | Dataset | Pros | Cons | C | I | A | Di | De | Tp | Au | Sc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zaabar *et al.* [23], 2022 | Blockchain, Hyperledger fabric | – | In order to monitor patient vital signs data, the HealthBlock framework takes advantage of the ideas of permissioned blockchain networks and decentralized storage. The adoption of an IPFS-based database called OrbitDB was also suggested by the authors. | The attack success rate is less than 20% when the training is in its latter stages. According to the results of experiments, a consensus process that is founded on gradient verification is capable of successfully overcoming a segment of poisoning attacks. | ✔ | ✔ | ✗ | ✔ | ✔ | ✔ | ✗ | ✔ |
| Zhang *et al.* [24], 2022 | Blockchain, ABE, | – | The developed strategy mitigates the risk of a single point of failure and lowers the costs linked with online procedures by employing online-offline encryption. The authors furnished a formal security proof in line with the random oracle paradigm. | It can be difficult to guarantee that all of the essential security standards are satisfied because it is not always possible to achieve the desired degree of efficiency. | ✔ | ✗ | ✗ | ✔ | ✔ | ✔ | ✔ | ✔ |
| Lin *et al.* [25], 2022 | Blockchain | – | In order to prevent the proxy from discovering the condition, it is contained within the re-encryption key. | Due to the fact that it is not always possible to attain the appropriate level of efficiency, it might be challenging to provide an assurance that all of the essential security standards are achieved. | ✔ | ✗ | ✗ | ✔ | ✔ | ✔ | ✔ | ✔ |
| Ahmad *et al.* [26], 2022 | IoT, Blockchain | – | The integration of fog computing enhances the scalability and flexibility of the system, allowing it to efficiently handle increasing amounts of health data. | The system's reliance on both fog and cloud computing infrastructure may necessitate substantial initial investment and ongoing maintenance. | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✗ | ✔ |
| Gupta *et al.* [27], 2023 | Blockchain, Deep Learning | Monkeypox skin lesion dataset | The use of blockchain technology addresses the challenge of secure health information sharing, which is a significant advancement in healthcare monitoring systems. | While the model shows promise for monkeypox detection, its effectiveness for other diseases remains theoretical and will require additional research and validation. | ✔ | ✔ | ✗ | ✔ | ✔ | ✔ | ✗ | ✗ |
| Ahmed *et al.* [28], 2023 | Blockchain, Deep Learning | Chest X-rays Dataset | The integration of blockchain technology ensures secure, transparent, and efficient management of healthcare data, addressing significant concerns about data privacy and integrity. | The scalability of the proposed system for widespread use in varied healthcare environments remains untested and might present challenges in real-world applications. | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✗ | ✔ |
| Alruwaili *et al.* [29], 2023 | Blockchain, Deep Learning | Monkeypox dataset | The use of Ethereum-based public BC ensures the privacy and security of healthcare images, mitigating risks associated with data breaches and unauthorized access. | The scalability of the proposed system in handling large-scale healthcare data across diverse medical facilities remains to be thoroughly tested. | ✔ | ✔ | ✗ | ✔ | ✔ | ✔ | ✗ | ✔ |
| Almalawi *et al.* [34], 2023 | IoT, Artificial Intelligence, Blockchain | IoT healthcare Security (Kaggle) | The hybrid encryption model (Lionized Remora optimization-based serpent) provides enhanced protection against unauthorized users, cyber-attacks, and data breaches. | The author does not mention potential weaknesses or areas where the encryption approach may be susceptible to attacks. | ✔ | ✗ | ✗ | ✗ | ✗ | ✔ | ✗ | ✗ |
| Kumar *et al.* [35], 2023 | Blockchain, Deep Learning | ToN-IoT, CICIDS-2017 | The proposed model provides a comprehensive solution to tackle security issues related to unprotected data. | The author does not address potential weaknesses that may arise during implementation. | ✗ | ✔ | ✗ | ✔ | ✔ | ✔ | ✗ | ✔ |

*(continued)*



Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 1. Continued.

| Ref. | Methodology | Dataset | Pros | Cons | C | I | A | Di | De | Tp | Au | Sc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rahal *et al.* [30], 2024 | Blockchain, Deep Learning | Breast cancer Wisconsin, Lung cancer prediction, Diabetes UCI dataset | The model ensembling approach is scalable and can be adapted to diagnose various medical conditions, making it a versatile solution for healthcare diagnostics. | Ensuring interoperability between different healthcare sectors and organizations for sharing model weights can be challenging, potentially hindering the seamless integration of the proposed system. | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✗ | ✔ |
| Ahmad *et al.* [31], 2024 | XAI | UCI, CHB-MIT, and Bonn EEG datasets | The multi-domain feature extraction (frequency, time, and non-linear) captures the complex nature of EEG signals, enhancing the robustness of the detection system. | The effectiveness of the system relies heavily on the quality and diversity of the input EEG datasets, which may vary across different healthcare environments. | ✔ | ✔ | ✗ | ✗ | ✔ | ✔ | ✗ | ✗ |
| Saidi *et al.* [32], 2024 | Federated Learning, Blockchain | Chest X-ray and CT Scans, GitHub | FL enables multiple stakeholders to collaborate on model development, leading to more comprehensive and accurate diagnostic models. | The computational resources needed for training FL models and maintaining blockchain networks can be substantial, potentially limiting accessibility for smaller healthcare facilities. | ✔ | ✔ | ✗ | ✔ | ✔ | ✔ | ✗ | ✔ |
| Myrzashova *et al.* [33], 2024 | Federated Learning, Blockchain | NIH Chest Ray dataset | By utilizing blockchain-based FL, the framework ensures that patients' medical data remains private and secure, addressing significant privacy concerns in healthcare. | The computational resources needed for training and maintaining the blockchain network can be substantial, potentially limiting its accessibility for smaller healthcare facilities. | ✔ | ✔ | ✗ | ✔ | ✔ | ✔ | ✗ | ✔ |
| **Proposed Method, 2024** | **Blockchain, Deep Learning** | **Dataset 1: BONN University EEG, Dataset 2: CHB-MIT Scalp EEG Dataset** | **The suggested model addresses the critical challenges in EHRs and offers a comprehensive, secure, and competent solution for managing healthcare data and diagnosis.** | — | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |

C- Confidentiality, I- Integrity, A- Availability, Di- Distributed, De- Decentralized, Tp-Tamper-Proof, Au- Authentication, Sc- Scalability

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

***vi) Tamper-Proof:*** Hyperledger fabric guarantees tamper-proof storage of medical records by employing the immutable nature of Blockchain mechanisms. Any effort to alter the information that has been stored would require the permission of all the participants in the network, which would make it extremely difficult to manipulate.

***vii) Authentication:*** Blockchain-based authentication mechanisms verify users' identities and strengthen that only authorized individuals can access medical records. ECC key generation and verification approaches improve authentication by offering secure cryptographic methods.

***viii) Scalability:*** The proposed method integrates Blockchain and deep learning approaches to address scalability challenges in EHRs. It employs a Hyperledger fabric for secure data management and optimal key generation using the SA-GBO technique. The DL-based ODBN model has been utilized for epilepsy disease detection, which shows scalability in handling large datasets.

### 2.1. Problem statement

Traditional HSs face challenges and issues, such as interoperability, integrity, authentication, scalability, and security in providing e-healthcare services. SHS has developed an attribute-based access control system with blockchain-empowered hidden policies to address these issues. This system utilizes a multi-authority framework to mitigate the possibility of a singular point of failure, guaranteeing robustness and reliability. A hybrid online-offline encryption technique reduces the computational load on users, particularly during online interactions, by assigning duties to users' inactive periods. A policy-hiding technique safeguards user privacy by concealing the access control policies. The solution utilizes blockchain and SC to ease secure and equitable payments for outsourced decryption activities between mobile edge computing servers and users. To increase the efficacy and proficiency of healthcare services, Wang *et al.* [17] addressed the need for a secure and privacy-preserving system for exchanging EHRs between healthcare providers and patients. This research offers an EHR sharing protocol that allows users to look for pertinent EHRs on the Consortium blockchain and retrieve the re-encryption ciphertext from a cloud server using data supplied by the data provider.

However, the system proposed in the paper involves a complex combination of cloud technology and blockchain, which may require more work to implement and maintain for some healthcare organizations. Xu *et al.* [18] addressed the crucial challenge of designing a system that ensures privacy while handling large amounts of health data. An adequate system must solve essential issues such as data security, ownership, and sharing among players in the healthcare ecosystem. Ensuring the immutability of data and diagnoses is crucial for reducing the likelihood of medical disputes. This suggests that data and diagnoses should be securely stored and safeguarded from unwanted modifications or refusals. Existing health chain schemes prioritize data privacy and security within specific businesses or networks but typically need more capacity to work efficiently with other HSs, limiting their overall usefulness. Wang *et al.* [19] demonstrated the urgent requirement for a robust and secure data management system in SHS. It is essential to have a system to protect sensitive healthcare records against unauthorized access, data loss, and tampering. The detection and prevention of abnormal activity in healthcare data is crucial to avoid situations such as inappropriate diagnosis and treatment and injury to patients. The study offers a comprehensive description of the Guard Health system, although it requires an in-depth evaluation of its performance, scalability, and practicality in real-world scenarios. In order to demonstrate the efficacy and feasibility of the proposed solution, further research and validation are required. Zhou *et al.* [20] employed a HITLM scheme that used blockchain technology and ML algorithms to protect sensitive healthcare data while allowing authorized parties to access and use the data for analysis and decision-making. The proposed solution involves a multi-layered architecture that includes data encryption, access control, and user feedback mechanisms to ensure the privacy and security of healthcare data. However, the authors do not offer a comprehensive evaluation of the performance and scalability of the suggested approach, conceivably preventing its applicability to more comprehensive healthcare environments.

## 3. PROPOSED METHODOLOGY

This section discusses the model and techniques employed in designing the proposed framework which are explained as follows:

### 3.1. Overview: role of Blockchain for security

The Ethereum blockchain functions similarly to other blockchain networks by generating a secure and decentralized ledger using interconnected blocks. Each block within the Ethereum blockchain consists of a set of data and a header that plays a significant role in managing the integrity and durability of the network [36]. The Ethereum block header comprises essential components, including the preceding block's hash, an independent identification value for the current block, a timestamp, a nonce and a cryptographic signature. These components are crucial for preserving the security and integrity of the Ethereum blockchain, as shown in Fig. 1. The independent identification value associated with each block is known as the block hash. This hash is generated using a cryptographic algorithm that combines the data within the block and the header information. It provides a unique identification for the block, which makes it easier for the block to be included in the blockchain.

### 3.2. Proposed model framework

SA_GBO-OBDN model is an advanced hybrid framework proposed to secure patients' EHRs and provide early disease prediction by employing the OBDN approach in the system. It offers different operational phases, such as data encryption and decryption, optimal critical generation process, secure data transactions using blockchain, and early disease diagnosis.

Figure 2 represents the working architecture of the proposed framework in the smart HS. The operation was conducted in two phases, which are discussed step by step as follows in Table 2.

In the first phase, the patient's personal health-related information (PHI) gets encrypted and decrypted using the Self-Adaptive Gradient-Based Optimizer- Elliptical Curve Cryptography (SA-GBO-ECC) technique, where SA-GBO generates the optimal key with small key sizes in a short period. The SA-GBO-based optimal key generation technique generates a secure unique key for encrypting EHRs. This combination of the model strengthens the framework by improving confidentiality and data integrity. It is a robust security mechanism that minimizes the security concerns related to healthcare data management. This model precisely protects the private patient's information by using the ECC technique to develop a secure framework against attackers, unauthorized access, and data breaches. In the second phase, after the decryption process, a DL-based ODBN model has

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

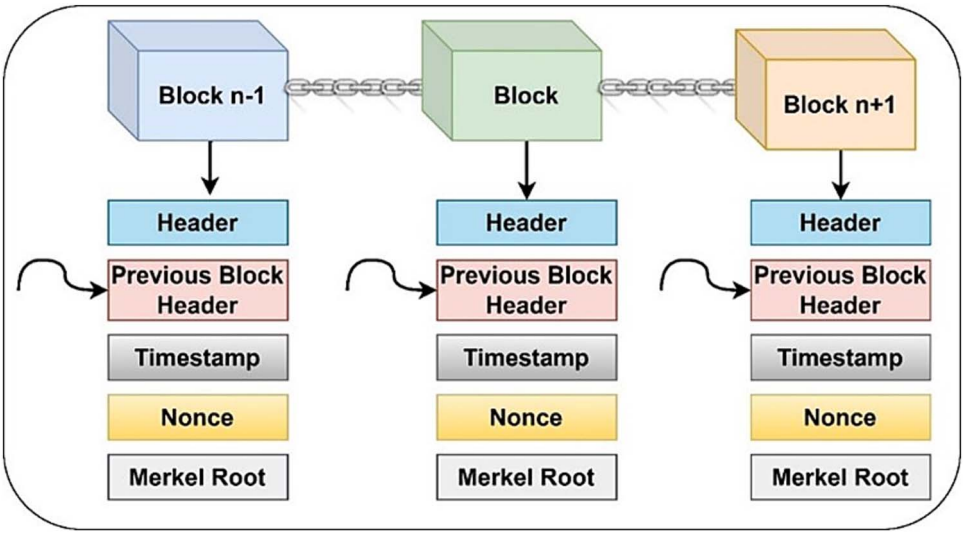



Figure 1. Blockchain architecture.

Table 2. Step-by-step working of the proposed framework in two phases

| |
|---|
| ***Phase 1-*** Secure Data Transmission: In this phase, the patient's healthcare data generated using smart IoT devices is collected and securely shared between patients, doctors, administrators, or stakeholders by following further steps. |
| *Step 1–Data Encryption:* The patient's PHI is encrypted utilizing the ECC algorithm to guarantee data security while transferring the information. This protects data from unauthorized access. SHA-256 (Secure Hash algorithm-256 bit) is also employed to generate a fixed-size unique hash value for input data, which ensures data integrity, security, efficiency, standardization, and tamper resistance to the system. |
| *Step 2–Optimal Key Generation and Selection:* The SA-GBO generates an optimal key for smoothly performing the encryption and decryption process. It generates a small key, which takes less storage space and speeds up the procedure compared to large keys generated by other algorithms. While the ECC technique has its own internal key generation process, SA-GBO can be employed to find an optimal key within the capabilities of the ECC algorithm. |
| *Step 3–Data storage to the cloud:* Once the information is encrypted, it is transferred to a cloud-based storage system for future retrieval. This ensures the implementation of acceptable encryption and access control procedures. |
| *Step 4–Blockchain Transmission:* To enhance data security, the encrypted data is transferred further using blockchain networks. This reduces the risk of unauthorized access and data breaches. |
| *Step 5 (Decryption):* The data can only be decrypted by an authorized user, such as doctors or hospital officials, whenever required using a secure optimal key. |
| ***Phase 2-*** Epilepsy detection using Deep Learning model: In this phase, after the decryption process, the deep learning model was utilized to detect Epileptic Seizures disease in the data. |
| *Step 6–Disease detection:* The ODBN model has been applied to the decrypted data to detect early epilepsy seizure disease, which helps minimize the risk of injuries, accidents, and other complications associated with it. |

been applied to detect epilepsy disease using two standard EEG data, i.e. BONN and CHB-MIT datasets.

Patients have complete control over who can ingress their personal health records (PHRs) and grant or revoke access to physicians or medical organizations as needed. The successful decryption and retrieval of a patient's comprehensive health data requires a highly effective encryption key. This is necessary for an authorized user to be able to access a patient's PHR for the diagnosis purpose. To ensure high security and compact key size, the ODBN driven diagnostic model is used to find the Epilepsy disease using EEG data. This approach enables secure management of patient health records and protects sensitive information from unauthorized access.

### 3.3. Data storage system using Blockchain

The fusion of ML for disease diagnostics and blockchain mechanisms for secure data storage and transmission is becoming very popular in the smart healthcare industry. This integrated approach addresses the critical issues of maintaining patient data confidentiality and integrity while supporting the efficient development of disease diagnosis models using ML approaches. Cryptocurrency is a shared ledger technology with several advantages, including decentralization, availability, privacy, integrity, and applicability. To begin with decentralization, it ensures that data stored on the blockchain is copied across multiple computers and eliminates the jeopardy of a single point of failure. In addition to availability, it allows access to the data, even in computer failure situations, while integrity defends against unauthorized modifications. The applicability enables tracing all data stored on the Ethereum blockchain, while privacy allows members to remain unidentified.

The Ethereum blockchain combines interlinked blocks, each containing data and a header. The blocks are allied using a cryptographic hash function, which develops a secure and tamper-resistant chain of transactions. The header of each Ethereum block consists of various components. Every blockchain block includes a cryptographic signature known as a hash, which is a unique identifier linking it to the previous block. This deep connection system raises the fundamental characteristics of the blockchain by maintaining chronological order and guaranteeing

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

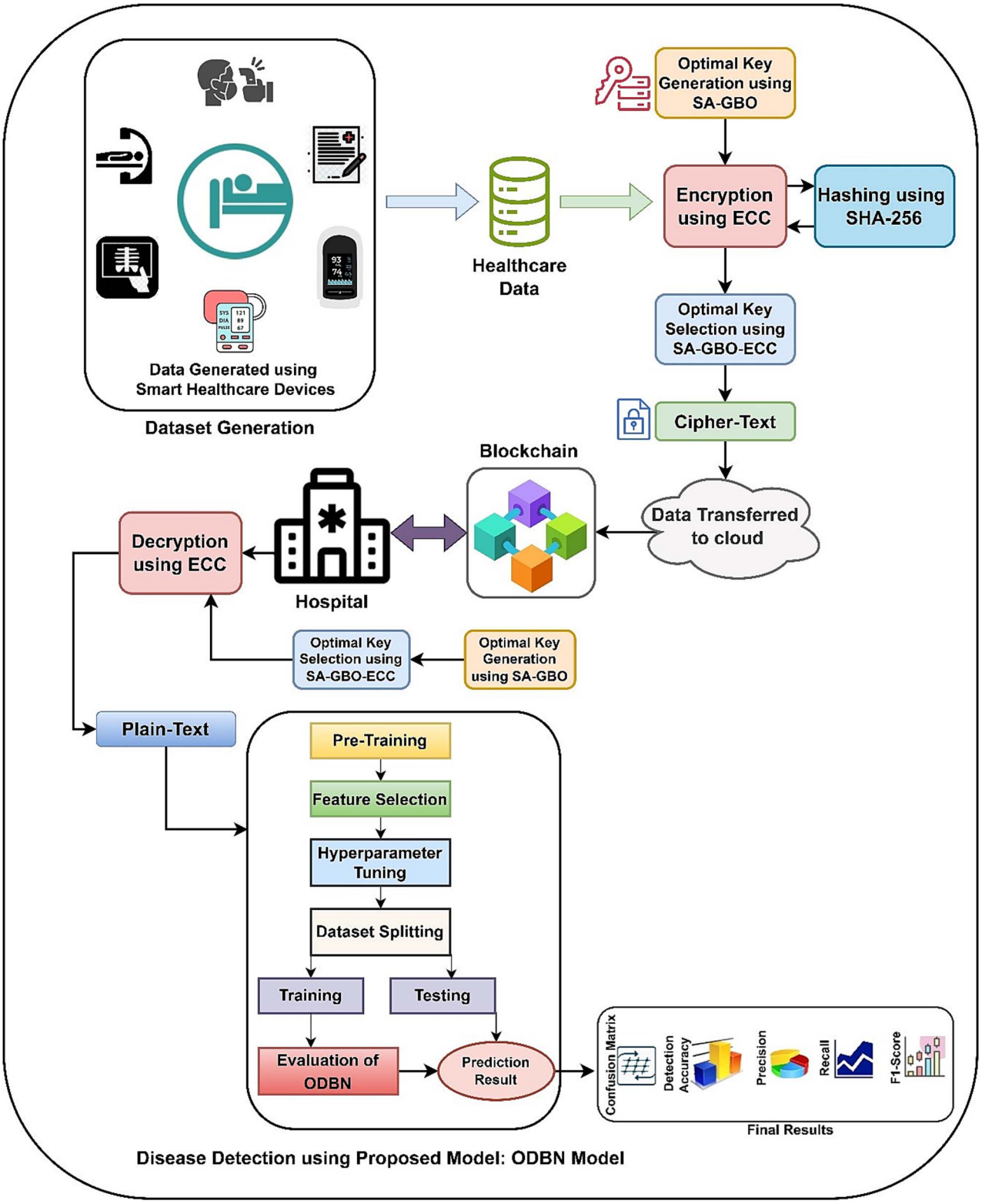



Figure 2. Proposed secure EHRs framework for data security and disease detection.

data integrity. Any effort to modify a previous block would cause a hash to change. This discrepancy would alert the entire network to the existence of malicious activity.

Moreover, each block includes a unique identity in its header, enhancing its uniqueness. This identifier can reference and locate specific blocks within the Ethereum blockchain. It provides a direction to validate the authenticity of the block. Another significant component in the header is the signature. This cryptographic signature ensures the authenticity and integrity of the block. It validates that an authorized user created the block, which has been unaltered since its beginning. The signature improves the security of the blockchain by protecting it from redundant modifications. It utilizes a consensus mechanism called Practical Byzantine Fault Tolerance (PBFT), which can

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

incorporate privacy features like data secrecy and confidential transactions. These features will guarantee the robust security and confidentiality of patient data while enabling secure data transmission and management. This consensus algorithm can handle a significant number of transactions per second, which is essential for effectively managing the large PHR of patients. Ethereum, like other blockchains, utilizes a square-chaining architecture. Each block contains a cryptographic reference to the block that appeared before it, creating a linear and interconnected series. This method supports the blockchain's fundamental consistency and unchangeability [37]. Modifying a block would change its reference, disrupting the chain and causing immediate detection throughout the network.

Due to privacy concerns, a component in a channel is restricted to accessing the blockchain data of its authorized peers exclusively. This access control system depends on both network connectivity and cryptographic methods. The system utilizes ECC for secure key exchange and SHA-256 hashing for data integrity checking. ECC is a strong encryption technique based on the mathematical characteristics of elliptic curves. It produces two distinct keys: a private key, which is kept secret, and a public key, which is shared publicly. This allows for secure communication, ensuring that only the entity with the appropriate private key can decrypt communications that have been encrypted with the shared public key. When a transaction is made on the blockchain, the sender uses their private key to sign the transaction, which can be verified using their public key.

On the contrary, SHA-256 is a commonly employed hashing algorithm that receives variable-length input and generates a fixed-length output, referred to as a hash. This hash is unique to the input and can be utilized to verify data integrity within the blockchain. The system usually produces a unique "fingerprint" using SHA-256 before including transaction data. Despite its weaknesses, SHA-256 is a fundamental measure to detect unauthorized modifications. Furthermore, ECC further enhances security. This encryption technique creates a private key for decryption, which is kept confidential, and a public key for encryption, which is shared explicitly. Transactions are encrypted using the public key and can only be decrypted by the holder of the associated private key. The combination of hashing and ECC techniques helps prevent attacker data alteration. It ensures the legitimacy and permanence of transactions.

This system provides scalability using dedicated channels for individual transactions. This method enables adaptable data storage among different commit nodes and supports concurrent processing of multiple transactions. As a result, the system can handle more node requests, resulting in enhanced overall scalability. Authorized peers are vital in gathering client transactions and using SC to evaluate their compliance with predetermined rules. This guarantees faithfulness to regulations and protects the system's integrity. Peers in a network have two key roles: collecting user interactions and structuring them to provide a reliable record of the blockchain's activities. This promotes trust and transparency in the system. On every currently operational chain, it is the responsibility of the requesting peers to broadcast transactions to the committed peers. In this way, the integrity of the data and the consensus inside the network are maintained. All the participants are guaranteed to get the same transaction information. This blockchain framework offers an effective alternative for storing individual EHRs. This system utilizes a distinct architecture in which each medical institution manages its local blockchain, designed exclusively for storing an individual's EHRs. The proposed framework guarantees data privacy and management inside each institution. Medical organizations can utilize the extensive features of the Ethereum blockchain without developing the complete blockchain infrastructure from the beginning. They need the essential connectivity and access protocols to connect with the current Ethereum network, allowing them to utilize its features, such as consensus mechanisms and SC.

## 3.4. ECC cryptographic algorithm-based data encryption process

ECC is a form of public-key cryptography based on the algebraic structure of elliptic curves over finite fields. It provides equivalent security to traditional cryptographic systems with significantly smaller key sizes, making it suitable for use in resource-constrained environments like the IoT. The security of ECC is built on the Elliptic Curve Discrete Logarithm Problem (ECDLP), which asserts that given a point $A = a.G$ on the curve, it is computationally infeasible to derive the private scalar $a$ from the public point $A$.

In this paper, we used Curve25519, which is an elliptic curve used extensively for high-performance, secure cryptographic applications, particularly in environments with constrained resources, such as the IoT. It is optimized for speed, security, and resistance to common cryptographic attacks, which makes it suitable for key exchange protocols and digital signatures. This approach is based on a Montgomery curve defined over a prime field $\mathbb{F}_p$, with $p = 2^{255} - 19$ which is represented using Eq. (1):

$$y^2 = x^3 + 486662x^2 + x \tag{1}$$

Where, $p = 2^{255} - 19$ is a large prime that defines the field $\mathbb{F}_p$. The constant coefficient 486,662 was chosen for performance and security, as it facilitates efficient modular arithmetic and avoids small-group vulnerabilities.

Curve25519 operates with a base point $G$ with x-coordinate 9. The security of Curve25519 is based on the ECDLP, which asserts that given $A = a.G$, it is computationally infeasible to derive the private scalar from $a$ the public point $A$.

### 3.4.1. *Operational phases in Curve25519 key exchange*

Curve25519 is commonly used in the Diffie-Hellman key exchange protocol, which enables two parties to securely establish a shared secret over an insecure channel. We outline the primary operational phases, such as setup, key generation, encryption, and decryption.

**(i) Setup Phase.**

In Curve25519, the setup phase is standardized, and no additional configuration by the user is required. The following parameters are fixed: Prime Field, i.e. $\mathbb{F}_P$ *with* $p = 2^{255} - 19$ with curve equation (Eq. 1), and Base-point $G$ with the x-coordinate set to 9. The use of a fixed base point $G$facilitates efficient scalar multiplication, while the field choice ensures modular arithmetic can be implemented optimally for computational speed and security.

**(ii) Key Generation Phase.**

The Key generation in Curve25519 involves generating a private key and computing a corresponding public key.

***(a) Private Key Generation:*** A private scalar $a$ is randomly chosen in the range $1 < a < p - 1$. To ensure efficient scalar multiplication and security against specific attacks, Curve25519 requires a process known as key clamping:

- The private key's three least significant bits are set to 0.
- The second-most significant bit is set to 1.
- The most significant bit is cleared.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

This improves resistance to small subgroup attacks and certain side-channel attacks by effectively restricting the scalar to a subset of possible values.

***(b) Public Key Generation:*** The public key $A$ is computed by performing scalar multiplication of the private key $a$ with the base point $G$ on the curve as shown in Eq. (2):

$$A = a.G \tag{2}$$

The Scalar multiplication on Montgomery curves is optimized, which allows the public key $A$ to be computed efficient and securely. The generated public key $A$ can be safely shared, as it does not expose information about the private scalar $a$.

**(iii) Key Exchange and Encryption Phase.**

In the Curve25519, encryption is typically achieved through an Elliptic Curve Diffie-Hellman (ECDH) key exchange protocol. In this process, two communicating parties, Martha and Harris, derive a shared secret that can be used as an encryption key.

- *Martha's Key Pair:* Martha generates her private key $a$and computes her public key $A = a \cdot G$.
- *Harris's Key Pair:* Similarly, Harris generates his private key $b$ and computes his public key $B = b \cdot G$.
- *Public Key Exchange:* Martha sends $A$ to Harris, and Harris sends $B$ to Martha over a public channel.
- *Shared Secret Computation:* Upon receiving Harris's public key $B$, Martha computes the shared secret $S$ as shown in Eq. (3):

$$S = a.B = a.(b.G) \tag{3}$$

Likewise, Harris, upon receiving Martha's public key $A$, computes the shared secret $S$ as shown in Eq. (4):

$$S = b.A = A = b.(a.G) \tag{4}$$

By properties of elliptic curve scalar multiplication, both Martha and Harris get the same shared secret $S = ab \cdot G$.

*Key Derivation:* The shared secret $S$, hashed using a cryptographic hash function using SHA-256 to derive a symmetric key $K = Hash(S)$. This symmetric key $K$ is used to encrypt and decrypt messages between Martha and Harris using a symmetric encryption algorithm.

**(iv) Decryption Phase.**

As Curve25519 primarily facilitates key exchange, decryption occurs at the symmetric level. With the derived shared key $K$, Martha and Harris securely exchange encrypted messages. The symmetric encryption key $K$ derived from the shared secret $S$ is used in a symmetric cipher for decrypting messages. Any cipher text transmitted between Martha and Harris is decrypted using shared key as shown in *Eq. (3) and (4)*, which ensures secure communication based on the Diffie-Hellman-derived symmetric encryption.

Curve25519 is a high-performance, secure elliptic curve that supports modern cryptographic protocols, especially in resource-constrained environments. The setup, key generation, key exchange, and decryption phases facilitate secure communications with a small footprint, which makes Curve25519 suitable for IoT applications and high-throughput cryptographic environments. Its efficiency, combined with rigorous security properties, has made it an essential tool in the cryptographic domain.

## 3.5. Gradient-based optimizer

The GBO method as depicted in Fig. 3 integrates population-based tactics with gradient-based methods to address complex optimization problems. The strategy uses Newton's method to guide the agents' search paths as they explore the problem space. The main two elements of the algorithm are the local escaping operator and the gradient search rule.

***A. Initialization stage***

The GBO algorithm creates an initial population of vectors at the beginning of the optimization process, like many other metaheuristic algorithms. Each agent in the population acts as a "vector" in a J-dimensional search space, and these vectors are generated independently using a uniform random distribution. There are $N$ vector agents in the population. Eq. (5) represents the initialization procedure:

$$\begin{gathered} J_{in} = J_{\min} + \mathbf{rand}\,(0,1) \times (J_{\max} - J_{\min}) \\ Here, i = 1, 2, \ldots, \mathbf{N} \\ n = 1, 2, \ldots, \mathbf{J} \end{gathered} \tag{5}$$

where a random number is rand (0, 1) that is defined in the [0, 1] and $J_{\min}$, and $J_{\max}$ are the limits of the decision variable $J$.

***B. Gradient search rule stage***

To balance exploitation and exploration, a significant factor $\varrho_1$ is used to update the agent positions based on a gradient-specified direction, as depicted in *Eq.* (6–8). Starting with a random set of initial solutions, this method seeks to explore the significant search space regions while also making use of the search to arrive at near-optimal and global points.

$$\varrho_1 = 2 \times \mathbf{rand} \times \delta - \delta \tag{6}$$

$$\delta = \left| \vartheta \times \sin\left(\frac{3\pi}{2} + \sin\left(\vartheta \times \frac{3\pi}{2}\right)\right)\right| \tag{7}$$

$$\vartheta = \vartheta_{\min} + (\vartheta_{\max} - \vartheta_{\min}) \times \left(1 - \left(\frac{\upsilon}{\mathbf{M}}\right)^3\right)^2 \tag{8}$$

In order to balance exploitation and exploration, the GBO algorithm adjusts the value of the parameter, $\varrho_1$ using a sine function $\delta$. In the initial optimization iterations, the parameter value is set high to increase population diversity, and then it is lowered to quicken population convergence. Within a range of iterations, the parameter value increases to enhance solution diversity and avoid local sub-regions, leading to the exploration of more solutions and convergence around the best solution. The GSR is determined by $\vartheta_{\min}$ and $\vartheta_{\max}$, which have constant values of 1.2 and 0.2, respectively. The parameter values of $\varrho_1$ change based on the sine function $\delta$, where $\mathbf{M}$ is the total number of iterations and $\upsilon$ is the current iteration number.

$$GSR = \mathbf{randn} \times \varrho_1 \times \frac{2\Delta \mathrm{j} \times j_n}{(j_{\mathrm{worst}} - j_{\mathrm{best}} + \varepsilon)} \tag{9}$$

The GSR concept in the GBO algorithm promotes exploration behaviour and prevents getting stuck in local optima by introducing random behaviour across iterations. *Eq. (9–11)* defines the factor $\Delta j$, which calculates the difference between the randomly selected solution $j_{r1}^{m}$and **best solution** $(j_{\mathrm{best}})$. To ensure that $\Delta j$ varies throughout the iterations, the $\partial$ parameter is calculated using Eq. (12). Additionally, randn is introduced to the Equation

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

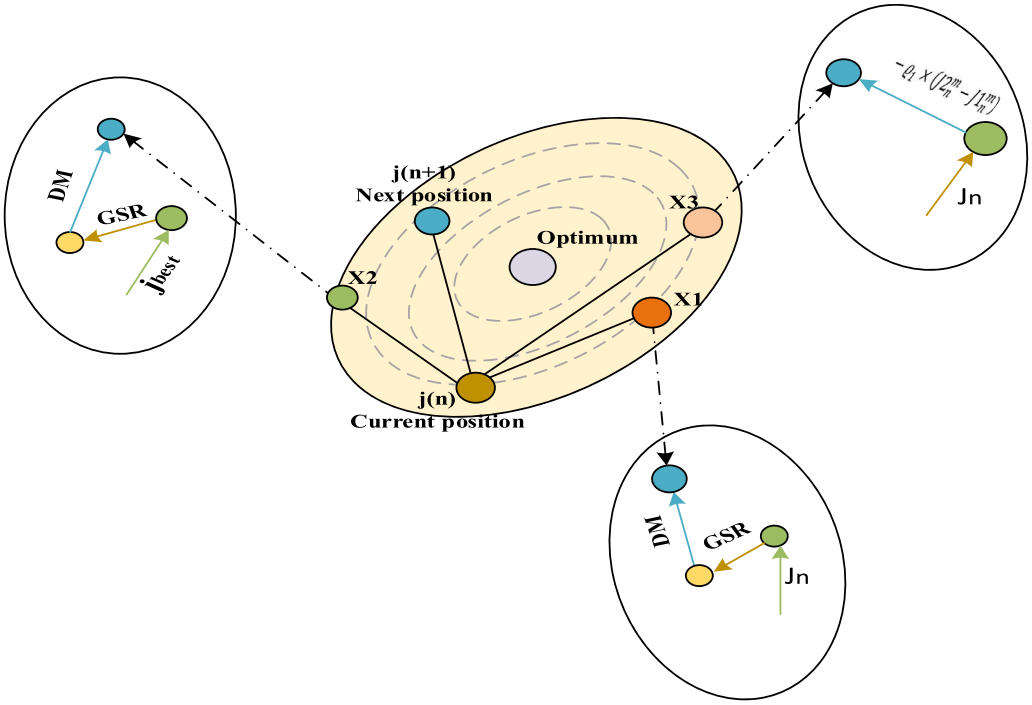



Figure 3. The generic layout of GBO algorithm.

further to enhance exploration.

$$\Delta j = \mathbf{rand}\,(1:\mathbb{N}) \times |\,\mathbf{step}\,| \tag{10}$$

$$\mathbf{step} = \frac{(j_{\mathbf{best}} - j_{i1}^{m}) + \partial}{2} \tag{11}$$

$$\partial = 2 \times \mathbf{rand} \times \left( \left| \frac{j_{i1}^{m} + j_{i2}^{m} + j_{i3}^{m} + j_{i4}^{m}}{4} - j_{n}^{m} \right| \right) \tag{12}$$

where the vector $\mathbf{rand}\,(1:\mathbb{N})$ contains $\mathbb{N}$ random values between $\epsilon\ [0,1]$. i1, i2, i3, and i4 are various integers that were randomly selected from $[1, N]$ such that $(i1 \neq i2 \neq i3 \neq i4 \neq n)$. $j_{best}$ and $j_{i1}^{m}$ determine the step size, which is denoted by *step*.

The solution area, $j_n$, is converged around using the direction of movement (DM). The mechanism adjusts the direction of $j_n$, focusing on the difference between the current vector $(j_n)$ and the best vector $(j_{\mathbf{best}})$. The vector heading from $(j_n)$ towards $(j_{\mathbf{best}})$ is indicated by the adjustment, denoted as $(j_{\mathbf{best}} - j_n)$. As a result, the equation directs $(j_n)$ motion toward the ideal solution denoted by $(j_{\mathbf{best}})$, supporting convergence inside the solution space.

This procedure offers a practical local search tendency that has a substantial impact on GBO convergence. The following Eq. (13) is used to calculate the DM:

$$DM = \mathbf{rand} \times \varrho_2 \times (j_{\mathbf{best}} - j_n) \tag{13}$$

where, $\varrho_2$ is used as a random parameter to change the step size of each vector agent, and rand is a uniformly distributed number within $[0, 1]$. The, $\varrho_2$ parameter takes important factors into account when exploring GBOs. $\varrho_2$ parameter is calculated using Eq. (14):

$$\varrho_2 = 2 \times \mathbf{rand} \times \delta - \delta \tag{14}$$

Lastly, based on these terms DM and GSR, Eq. (15) and Eq. (16) modified the vector's current position, $(j_n^m)$.

$$J1_n^m = j_n^m - GSR + DM \tag{15}$$

where, $J1_n^m$ is the fresh vector created through updating $j_n^m$. According to Eq. (8) and (13), the $J1_n^m$ is reformulated as:

$$J1_n^m = j_n^m - \mathbf{randn} \times \varrho_1 \times \frac{2\Delta j \times j_n^m}{(yp_n^m - yq_n^m + \varepsilon)} + \mathbf{rand} \times \varrho_2 \times (j_{\mathbf{best}} - j_n^m) \tag{16}$$

where $yp_n^m, yq_n^m$ are equal to $y_n + \Delta j$, and the $y_n - \Delta j, y_n$ is same as the combination of the two vectors:$z_{n+1}$ vector and the current solution $j_n$. This is determined by Eq. (17) as follows:

$$z_{n+1} = j_n - \mathbf{randn} \times \frac{2\Delta j \times j_n}{(j_{\mathbf{worst}} - j_{\mathbf{best}} + \varepsilon)} \tag{17}$$

while $j_{\mathbf{worst}}$ and $j_{\mathbf{best}}$ represent worst and best solutions, random solution vector of dimension $n$ is denoted as randn, $j_n$ is the current solution vector, and $\Delta j$ is given by Eq. (10). By using the best solution vector, $j_{\mathbf{best}}$, in the previous formula, we get $J2_n^m$ subsequently using the $j_n^m$ current solution vector.

$$J2_n^m = j_{\mathbf{best}} - \mathbf{randn} \times \varrho_1 \times \frac{2\Delta j \times j_n^m}{(yp_n^m - yq_n^m + \varepsilon)} + \mathbf{randn} \times \varrho_2 \times (j_{i1}^m - j_{i2}^m) \tag{18}$$

It aims to improve the exploitation and exploration phases by using Eq. (16) and Eq. (18) to enhance global search during exploration and local search during exploitation, respectively. For the next iteration, the new solution is made as depicted in Eq. (19):

$$j_n^{m+1} = r_a \times (r_b \times J1_n^m + (1 - r_b) \times J2_n^m) + (1 - r_a) \times J3_n^m \tag{19}$$

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

where random numbers are determined by $r_a$, and $r_b$ in range [0, 1], and $J3_n^m$ is defined by Eq. (20)

$$J3_n^m = J_n^{m+1} - \varrho_1 \times \left(J2_n^m - J1_n^m\right) \tag{20}$$

### C. Local escaping operator stage

The Local Escaping operator (LEO) is a tool designed to enhance the effectiveness of optimization algorithms that deal with complex problems. It accomplishes this by enabling the efficient updating of a solution's position, which in turn helps the algorithm escape local optima points and achieve convergence more quickly. To generate new solutions with improved performance (represented by $(J_{LEO}^m)$, the LEO uses targets. These targets include randomly selected solutions such as $(J_{\text{best}}), J1_n^m, J1_n^m)$ and $(J, J_{r1}^m, J1_{r2}^m)$. It successfully updates existing solutions, and the strategy is carried out using the following framework as shown in *Eq. (21)*:

If rand $<$ *pr*

$$J_{LEO}^m = \begin{cases} J_n^{m+1} + f_1\left(u_1 j_{\text{best}} - u_2 j_k^m\right) \\ \quad + f_2 \rho_1 \left(u_3 \left(J2_n^m - J1_n^m\right)\right) \\ \quad + u_2 \left(j_{i1}^m - j_{i2}^m\right)/2, \\ J_n^{m+1} + f_1\left(u_1 j_{\text{best}} - u_2 j_k^m\right) \\ \quad + f_2 \rho_1 \left(u_3 \left(J2_n^m - J1_n^m\right)\right) \\ \quad + u_2 \left(j_{i1}^m - j_{i2}^m\right)/2, \qquad \text{otherwise End} \end{cases} \tag{21}$$

The values of $f_1$ and $f_2$ are randomly generated from a uniform distribution within the range $\in [-1,1]$. Additionally, three random values, denoted as $u_1, u_2$, and $u_3$, are generated. The probability value pr is equal to 0.5.

$$u_1 = \begin{cases} 2 \times \text{rndm} & \text{if } \mu_1 < 0.5 \\ 1 & \text{otherwise} \end{cases} \tag{22}$$

$$u_2 = \begin{cases} \text{rndm} & \text{if } \mu_1 < 0.5 \\ 1 & \text{otherwise} \end{cases} \tag{23}$$

$$u_3 = \begin{cases} \text{rndm} & \text{if } \mu_1 < 0.5 \\ 1 & \text{otherwise} \end{cases} \tag{24}$$

Here $\mu_1$ is a number in the [0, 1], and 'rndm' symbolises a random number $\in$ [0, 1]. The previous *Eq. (22–24)* represent $u_1, u_2, u_3$, which is simply explain as follows from *Eq. (25–27)*:

$$u_1 = L_1 \times 2 \times \text{rand} + (1 - L_1) \tag{25}$$

$$u_2 = L_1 \times \text{rand} + (1 - L_1) \tag{26}$$

$$u_3 = L_1 \times \text{rand} + (1 - L_1) \tag{27}$$

if a parameter $\mu_1 < 0.5$, then the value of $L_1 = 1$, $L_1$ is a binary parameter that takes value 0,1, otherwise $L_1 = 0$. The solution $j_k^m$ is created as follows using Eq. (28):

$$j_k^m = \begin{cases} j_{\text{rand}} & \text{if } \mu_2 < 0.5 \\ j_p^m & \text{otherwise} \end{cases} \tag{28}$$

Randomly generated solution $j_{\text{rand}}$is generated using Eq. (29):

$$j_{\text{rand}} = J_{\min} + \text{rand}\,(0, 1) \times (J_{\max} - J_{\min}) \tag{29}$$

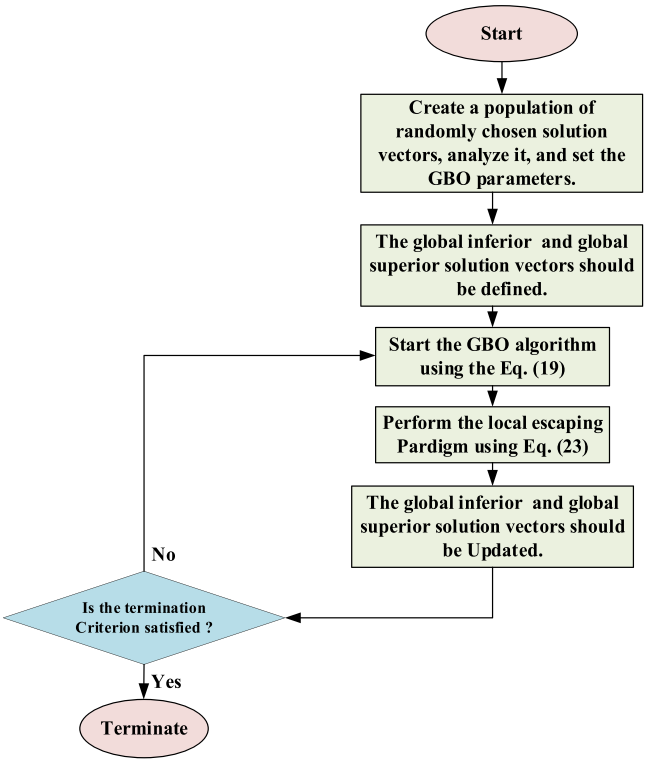



Figure 4. The proposed workflow of GBO algorithm.

$\mu_2$ is a random number between $\in$ [0, 1] ,

$j_p^m$ is a solution randomly chosen from the population

## 3.6. Self-adaptive strategy based GBO

From the above analysis, it can be inferred that the low convergence rate and accuracy of the standard GBO are due to certain factors or limitations identified in the analysis. A new proposed SA-GBO has been designed as a result to address the aforementioned issues. According to the analysis's findings, the gradient search rule value of the parameter $New_\delta$ must satisfy Eq. (31) in order for the algorithm to converge. To strike a better balance between exploration and exploitation abilities during an iteration, the parameter value should decrease gradually. As a result, a new self-adaptive strategy has been proposed to adjust the parameter value accordingly. Figure 4 depicts the working flow of GBO.

$$\varrho_2 = 2 \times \text{rand} \times New_\delta - New_\delta \tag{30}$$

$$New_\delta = \left(\frac{\text{rand}\,(1 + \vartheta_n) \times \sin^{\left(\frac{\pi\vartheta_n}{2}\right)}}{\text{rand}\left(\frac{1+\vartheta_n}{2}\right) \times \vartheta_n \times 2\left(\frac{\vartheta_n - 1}{2}\right)}\right) \tag{31}$$

$$\vartheta_t = \vartheta_0 - \vartheta_0 e^{-(1-\frac{t}{T})} \tag{32}$$

where T stands for the most iterations. Naturally, as t increases, the parameter, $New_\delta$, decreases.

From the *Eq. (30–32)*, the self-adaptive GBO algorithm replaces the parameter of the traditional GBO by introducing new efficient parameters. The parameters such as $\varrho_2$, $New_\delta$, and $\vartheta_t$ are utilized to ensure the balance between the exploration and exploitation phases during training. Moreover, due to advantage of the strategy,

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

the local convergence rate can be increased for traditional GBO. Finally, the proposed self-adaptive GBO provides high-level performance than the existing GBO. Therefore, the proposed model has adopted the SA-GBO for key generation problem of cryptography algorithm specifically ECC algorithm. The following section shows the detailed description about the SA-GBO algorithm for key generation.

**Algorithm 1. The Pseudo Code of the Self Adaptive-Gradient-Based Optimizer**

**Step 1** Initialization Put values on the parameters ε, pr, M
**Step 2** Create the first population $X0 = [x0,1, x0,2, \ldots, xo, D]$
  Evaluate the objective function value $f(X0) =, n = 1, \ldots, N$
  Replace the parameter $\varrho_2$ in gradient search rule value Eq. (6) with $New_\delta$ in Eq. (30)
  Specify the worst and best solutions $j^m_{worst}$ and $j^m_{best}$
**Step 3** Main loop
  **while** m < M **do**
    **for** n = 1 to N **do**
      **for** n = 1 to D **do**
        Select randomly r1 ≠ r2 ≠ r3 ≠ r4 ≠ n in [1, ℕ]
  Calculate the position $j^{m+1}_{n,i}$ using Eq. (19)
  **end for**
  **if** rand< pr then ▷ Local escaping operator
  Calculate the position $J^m_{LEO}$ using Eq. (21)
  $j^{m+1}_{n,} = J^m_{LEO}$
  **end if**
  Update the positions $j^m_{best}$ and $j^m_{worst}$
  **end for**
  $m = m + 1$
  **end while**
**Step 4** Return $j^m_{best}$

## 3.7. SA-GBO-based optimal key generation process for ECC Curve25519

In ECC, a randomly generated private key is essential for both encryption and decryption processes to ensure security. In the proposed model, the private key is generated using a Self-Adaptive Gradient-Based Optimizer (SA-GBO) algorithm. The private key generated by SA-GBO significantly influences the encryption quality in Curve25519 by optimizing the randomness in the cryptographic key space, thereby strengthening protection for strategic data during ECC operations.

### 3.7.1. *Optimization process of GBO for private key generation*

The GBO algorithm initiates by searching local regions within the key space for identifying potential solutions. In this work, the search space corresponds to the field over which ECC Curve25519 operates, with a targeted private key length of 256 bits. The search begins by exploring a localized region with candidate vectors representing potential keys, with iterative updates aimed at refining these vectors to approach the optimal private key solution.

### 3.7.2. *Gradient search rule and local escaping operator*

Within the Gradient Search Rule stage, a new candidate vector $J1^m_n$ is generated, which is calculated using Eq. (33), and compared to the existing vector $j^m_n$ to determine whether it enhances the solution within the exploitation phase. If the new vector improves performance, it updates the current best vector. However, if it does not provide a superior solution, the LEO stage generates an alternative candidate, which ensures that the search continues to escape local minima and find other optimal vector.

The population of GBO vectors in this search space is modelled as a $J$-dimensional matrix, where $N$ denotes the total number of vectors, and $J$ signifies the number of dimensions in the search space. The candidate vector $J_{in}$ in the $i-th$ position is initialized using Eq. (33):

$$J_{in} = J_{min} + \mathbf{rand}\,(0,1) \times (J_{max} - J_{min}) \tag{33}$$

$$Where\ i = 1, 2, \ldots, N\ and\ n = 1, 2, \ldots\ldots, J.$$

### 3.7.3. *Binarization of GBO vectors for ECC Curve25519 private key generation*

Since Curve25519 requires a binary private key represented as a 256-bit integer, a binarization process is applied to convert GBO vectors into binary value which is compatible with ECC. The binarization process for each vector $J_{in}$ is defined by Eq. (34):

$$J_{in} = \begin{cases} 1 & if\ J_{in} \geq 0 \\ 0 & if\ J_{in} < 0 \end{cases} \tag{34}$$

The length of each binary vector corresponds to the initial key length required by Curve25519, with each bit contributing to the 256-bit private key size. Eq. (35) provides a further refinement based on vector indices and an offset constant $b = 91$ :

$$K_{in} = J_{in} \times \left[(b+n) \times 2\right] + 12 \tag{35}$$

Where $n = 1, 2, \ldots\ldots j$, and $K_{in}$ defines the intermediate transformation for each vector element.

### 3.7.4. *Fitness evaluation and optimal key selection*

The fitness of each vector is assessed using Eq. (36), which calculates the fitness $F$ by binarizing and inserting $J = (0, 1)$ into each vector and evaluating through a binary XOR operation. This allows for an efficient assessment of the vector's suitability as a private key for ECC Curve25519. Specifically, Fis minimized based on the binary XOR between the candidate vector $J_i$ and the transformed vector $K_i$:

$$F = H(J, K) = \sum_{i=1}^{n} (J_i \oplus K_i) \tag{36}$$

If the calculated value of $K$ exceeds 12, a binary 1 is used in Eq. (36) to reduce the fitness function $F$; otherwise, a binary 0 is assigned. The optimization process iterates over a predefined number of cycles, gradually refining the GBO vectors until the most suitable 256-bit key is identified.

The SA-GBO optimization framework for ECC Curve25519 efficiently generates a high-quality private key by iterating over candidate vectors within a defined search space, which uses both gradient-based exploitation and local escaping mechanisms. This approach ensures that the private key space is robustly explored that leads to a securely randomized private key optimized for use in ECC-based cryptographic systems. The combination of SA with GBO enhances the search by escaping local minima by producing a private key that meets ECC Curve25519 security requirements.

## 3.8. Disease diagnosis using ODBN

As a DL algorithm, the Optimized Deep Belief Network (ODBN) has significantly gained worldwide popularity. This algorithm can effectively learn complicated representations from data and handle hidden layers effectively, enabling them to find the best

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

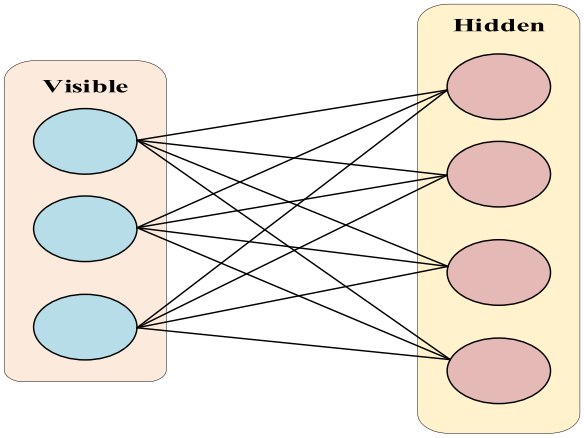



Figure 5. Restricted Boltzmann machine.

possible outcomes even when dealing with significant, complex network architecture. An ODBN has two main components: (i) a restricted Boltzmann machine (RBM)-based unsupervised learning (UL) section that serves as the basic building block, and (ii) a layer for prediction using logistic regression that is placed on top.

The double-layer neural network (DLNN), also called the RBM, plays a substantial role in the deep belief network (DBN). The RBM has two layers: a visible layer, which serves as the input layer, and a hidden layer, which comprises hidden neurons with Boolean values, as illustrated in Fig. 5.

### 3.8.1. *Pretraining*

The connection weights and offsets between the RBM layers are initialized during pretraining using an unsupervised greedy layer-by-layer method. The neurons in these two layers are connected bidirectionally and symmetrically; however, within the same layer, there are no connections between any of the neurons. The layer-wise configuration follows the energy function of these two layers, denoted by Eq. (37), to determine their probability distribution. Eq. (38) provides the hidden layer's probability distribution.

$$E(v,h) = -\sum_{i=1}^{n_v} a_i v_i - \sum_{j=1}^{n_h} b_j h_j \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_j W_{j,iv_i} \tag{37}$$

$$P(v,h) = \frac{e^{-E(v,h)}}{\sum_v \sum_h e^{-E(v,h)}} \tag{38}$$

The RBM, illustrated in Fig. 5, is a DLNN, serving as a fundamental component of the DBN. It consists of a visible layer, which is the input layer, and a hidden layer comprising Boolean hidden neurons. In Eq. (30), the variables hold the following significance: $n_k$ symbolizes the total number of neurons in the hidden layer, $h_i$ refers to the binary state of the $i^{th}$ Boolean neuron in the hidden layer, $n_v$ indicates the total number of neurons in the visible layer, $v_i$ stands for the binary state of the ith neuron in the visible layer, $W_{j,i}$ denotes the weight matrix linking the hidden and visible layers, and $a_i$ and $b_i$ are the bias vectors for the hidden and visible layers, respectively. The activation functions for the hidden and visible layers are specified by Eq. (39) and Eq. (40), respectively.

$$P(v_i = 1 \,|h) = sig\left(\alpha_i + \sum_{j=1}^{n_h} W_{j,i} h_j\right) \tag{39}$$

$$P(h_i = 1 \,|v) = sig\left(b_i + \sum_{i=1}^{n_v} W_{j,i} v_j\right) \tag{40}$$

where a sigmoid function is denoted as *sig* ().

As stated, it is insufficient to provide accurate forecasting using a single hidden RBM as a unit. The DBN uses stacked layers of RBMs with a logistic regression layer to overcome this limitation and enable gradual feature extraction from the data. Within the DBN, the training process is carried out layer-wise across various RBMs. The first RBM (RBM1) is pre-trained independently using the training data, as shown in Fig. 5. The weights between the first RBM's hidden layer (h1) and visible layer are calculated. After that, the second RBM (RBM2), which also has a hidden layer (h1) and a visible layer, receives the first RBM's output as input (h2), as depicted in Fig. 6.

After several RBMs are trained using the layer-wise method mentioned above, a logistic regression layer is added as the output layer. This layer is trained using the supervised learning (SL) approach. The parameters of the DBN algorithm are then adjusted using the Backpropagation algorithm.

### 3.8.2. *Feature selection and hyperparameter tuning*

It is a vital process of designing ML models. This phase aims to identify the most informative or relevant feature from the dataset. This section discusses the feature selection process employed in our proposed framework for managing and diagnosing medical data:

- The filter technique evaluates the significance of each feature without considering the ML model. We assess the features using statistical metrics such as variance thresholding and correlation analysis to identify the most useful features.
- The wrapper approach includes continuous evaluation of feature subsets throughout the model's training and testing process. Techniques like forward selection and backward elimination are employed to categorize the optimal features that maximize the model's performance.
- The embedded approach incorporates feature selection directly into the model's training process. We employed regularization techniques such as L1 regularization and tree-based feature significance during model training to automatically choose the most significant features.
- The proposed model's optimized hyperparameter tuning is systematically done by considering values, as depicted in Table 3, which helps achieve outstanding performance.

As stated in the previous section, the suggested model improves accuracy by analyzing loss functions using the SA-GBO technique. The dense layer employs the SoftMax activation function to produce the final output.

This research suggested an ODBN with precisely determined hyperparameters and an effective optimizer to attain outstanding results by utilizing different activation functions. Every function introduces a distinct form of non-linearity, affecting representation capabilities and learning dynamics. SA-GBO is used to find the appropriate combination of additional hyperparameters for the maximum number of epochs, batch size, number of hidden layers, and loss function (RMSE), which can be calculated using Eq. (41). As stated in the previous section, the proposed model is based on improving accuracy through analysis of the loss functions using the SA-GBO technique. This optimization algorithm is employed to update the parameters of an ML model during the training process. Utilizing appropriate

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

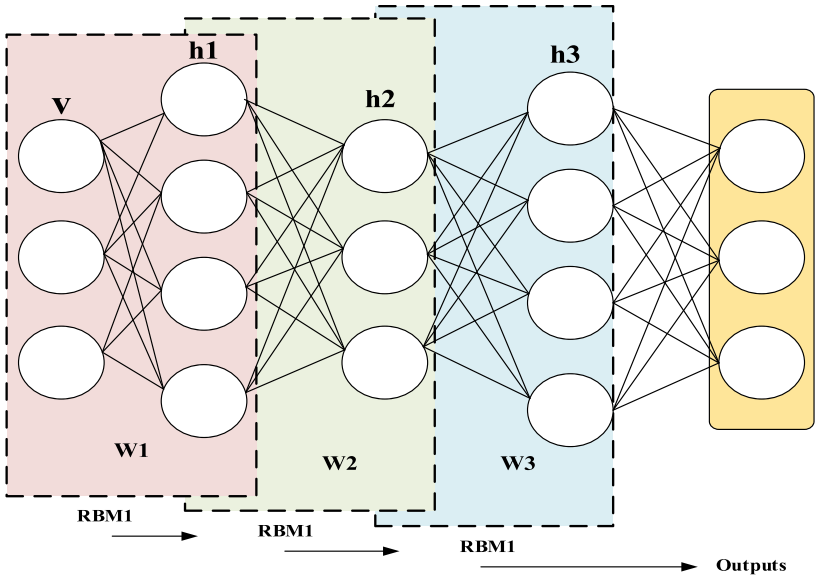



Figure 6. Layers of ODBN.

Table 3. Hyperparameter tuning of proposed DL model

| Hyperparameters | Values |
|---|---|
| Number of Dense Layers | 2 |
| Number of Neurons per Layer | 100 (1st Layer), 10 (2nd Layer) |
| Activation Function | ReLU (1st Layer), Softmax (2nd Layer) |
| Loss Function | Sparse Categorical Crossentropy |
| Number of Epochs | 100 |
| Number of Convolutional Layers | 1 |
| Optimizer | Adam |
| Batch Size | 64 |
| Number of Filters per Layer | 64 |
| Kernel Size | (1,1) |
| Pooling Size | (2,2) |
| Max Pooling Layer Dropout Rate | 0.2 |

computational approaches is essential for resolving the optimization issues of hyper-parameters. It is essential to employ suitable methods, especially when addressing the optimization challenge of hyperparameters. This approach enables an enhanced searching method, leading to optimal performance and outcomes. Table 4 depicts the Notation and its description.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - p_i)^2} \tag{41}$$

## 3.9. Proposed intrusion detection model

In this section, we propose an intrusion detection model using Random Forest (RF), Decision Tree (DT), and XGBoost on WUSTL EHMS 2020 Dataset [38]. In our work, we utilize this dataset which incorporates two types of man-in-the-middle attacks: spoofing and data injection. The spoofing attack is designed to intercept packets between the gateway and the server, thereby breaching the confidentiality of patient data. Conversely, the data injection attack actively alters packets on-the-fly, which compromises the integrity of the transmitted data. The objective is to classify network traffic as normal, or attack based on extracted features. The model is optimized using hyperparameter tuning and evaluated using standard performance metrics. Table 5 shows the feature of the dataset.

### 3.9.1. *Data preprocessing*

Data preprocessing is a crucial step in building an effective machine learning model for intrusion detection. It involves several stages to prepare the dataset for training. Data normalization is applied to scale the numerical features using Min-Max Normalization, ensuring uniformity across all features. This transformation is calculated using Eq. (42).

$$X' = \frac{X - X_{min}}{X_{max} - X_{\text{min}'}} \tag{42}$$

Where $X'$ is the normalized feature, $X_{min}$ is the minimum value, and $X_{max}$ is the maximum value.

Additionally, Label Encoding is used to convert categorical labels, such as “Normal” and “Attack”, into numerical values to make the dataset compatible with machine learning algorithms. Together, these preprocessing steps ensure that the dataset is optimized for model training, leading to better performance and accuracy.

### 3.9.2. *Machine learning models*

The classification models used in the proposed ML- based IDS are explained as follows:

*(i) DT model.*

A DT is a recursive, hierarchical structure where each node represents a feature, and each edge represents a decision. The best feature for splitting is chosen using either Gini Impurity or Entropy.

- Gini Impurity: For a dataset $D$ with $c$ classes, Gini impurity is defined using Eq. (43):

$$Gini(D) = 1 - \sum_{I=1}^{C} p_i^2 \tag{43}$$

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 4. Notation and its description

| Notations | Descriptions |
|---|---|
| $nAG_r$ | Anne's private key |
| $nB$ | Andrew's private key |
| $P_A$ | Anne's public key |
| $P_B$ | Andrew's public key |
| $G_r$ | Generator point on the elliptic curve |
| $P_m$ | Message to be sent |
| $P_c$ | Cipher text |
| $k$ | Random integer chosen for encryption |
| $kG_r$ | Intermediate value used in decryption |
| $J_{in}$ | Initialized value for the decision variable J |
| $J_{min}$ | Minimum limit of the decision variable J |
| $J_{max}$ | Maximum limit of the decision variable J |
| **Rand** [0, 1] | Random number in the interval [0,1] |
| $\varrho_1$ | Significant factor used to update agent positions |
| $\delta$ | Gradient-specified direction adjustment factor |
| $\vartheta$ | Parameter for the gradient direction computation |
| $\vartheta_{\mathbf{min}}$ | minimum value for $\delta$ |
| $\vartheta_{\mathbf{max}}$ | maximum value for $\delta$ |
| $\upsilon$ | current iteration number |
| $M$ | maximum number of iterations |
| $GSR$ | Gradient Search Radius used to balance exploration and exploitation |
| **randn** | A random value drawn from a standard normal distribution |
| $\Delta j$ | Difference between the best solution and a randomly selected solution |
| **randn** (1 : **N**) | A vector containing $N$ random values between [0, 1] |
| $j_{\mathbf{n}}$ | Current solution in the population |
| $j_{\mathbf{best}}$ | best solution in the current population |
| $j_{\mathbf{worst}}$ | worst solution in the current population |
| $\varepsilon$ | A small positive constant to avoid division by zero |
| $DM$ | Differential Modification term for updating vector agents |
| $r_a$ | A random number within the range [0, 1] |
| $r_b$ | A random number within the range [0, 1] |
| $J^m_{LEO}$ | Updated vector agent using the Local Exploitation Operator (LEO) |
| $\rho_2$ | Parameter controlling step size variability |
| $u_1, u_2, u_3$ | Random parameters used in the LEO process |
| $u_1, u_2$ | Random numbers between 0 and 1 for decision-making |
| **rndm** | Random number uniformly distributed in [0,1] |
| $f_1, f_2$ | Random coefficients generated from uniform distribution in [−1,1] |
| $J_n^{m+1}$ | New solution for the next iteration |
| $j_k^m$ | A solution vector chosen from the population or random solution |
| $L_1$ | Binary parameter used for setting $u_1, u_2, u_3$ |
| $j_{\mathbf{rand}}$ | Randomly generated solution vector |
| $j_p^m$ | A solution vector randomly chosen from the population |
| $New_\delta$ | New value for the $\delta$ parameter used in the update process |
| $\vartheta_n$ | A dynamic parameter influencing the update process |
| $\vartheta_{\mathrm{t}}$ | Time-varying parameter for the algorithm's iteration control |
| $\vartheta_0$ | Initial value of $\vartheta_{\mathrm{t}}$ |
| $t$ | Current iteration number |
| $T$ | Maximum number of iterations for the algorithm |
| $K_{in}$ | Computes the value for the dimension based on $J_{in}$, $b$, and $n$. |
| $b$ | A constant, usually defined as part of the algorithm's parameters |
| $n$ | The dimension index, ranging from 1 to J |
| $H$ $(J, K)$ | function that takes two vectors J and K as inputs and produces a scalar output by summing the bitwise XOR results of corresponding elements. |
| $J$ | Number of decision variables or dimensions |
| $J_{\mathrm{i}}$ | $\mathrm{i}^{\mathrm{th}}$ dimension value of the vector J |
| $K_{\mathrm{i}}$ | $\mathrm{i}^{\mathrm{th}}$ dimension value computed |
| $F$ | Evaluate the function using the XOR operation and summing the results: the resulting value of the function H(J,K) |
| $\upsilon_{\mathrm{i}}$ | value of the $\mathrm{i}^{\mathrm{th}}$ visible unit |
| $a_{\mathrm{i}}$ | Bias term for the $\mathrm{i}^{\mathrm{th}}$ visible unit |
| $b_j$ | Bias term for the $\mathrm{j}^{\mathrm{th}}$ hidden unit |
| $h_j$ | value of the $\mathrm{j}^{\mathrm{th}}$ hidden unit |
| $W_{j,\mathrm{i}}$ | weight connecting the $\mathrm{i}^{\mathrm{th}}$ visible unit to the $\mathrm{j}^{\mathrm{th}}$ hidden unit |
| $d_{\mathrm{i}}$ | actual observed value for the $\mathrm{i}^{\mathrm{th}}$ instance |
| $p_{\mathrm{i}}$ | predicted value for the $\mathrm{i}^{\mathrm{th}}$ instance |
| $n$ | total number of instances in the dataset |
| **rand** (1 : **N**) | Random values from a uniform distribution |

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 5. Description of dataset features

| Measurements | Values |
|---|---|
| Dataset Size | 4.4 MB |
| Attack samples | 2046 |
| Normal samples | 14,272 |
| Total number of samples | 16,318 |

Where $p_i$ is the probability of a sample belonging to class i. Lower values indicate purer nodes.

- Entropy and Information Gain: Entropy measures the uncertainty of a dataset using Eq. (44):

$$Entropy(D) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{44}$$

The Information Gain (IG) is used to determine the best feature split using Eq. (45):

$$IG(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} . Entropy(D_v) \tag{45}$$

Where $D_v$ is the subset of $D$ after splitting on attribute $A$. The feature with the highest IG is selected for splitting.

*(ii) RF model*

A RF is an ensemble of multiple decision trees, where each tree is trained on a random subset of data. The final classification is based on the majority vote of the trees which is calculated using Eq. (46).

$$P(y) = \frac{1}{N} \sum_{i-1}^{N} P_i(y) \tag{46}$$

Where $P(y)$ is the final predicted probability for class y, $P_i(y)$ is the predicted probability from the $i - th$tree, and $N$ is the total numbers of trees. Each tree uses a random subset of features, reducing overfitting and improving generalization.

*(iii) XGBoost model*

XGBoost is an optimized gradient boosting model that improves classification accuracy by training trees sequentially, minimizing a defined loss function. The overall loss function for XGBoost is calculated using Eq. (47):

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{47}$$

Where, $l(y_i, \hat{y}_i)$ is the loss function, and $\Omega(f_k)$ is the regularization term to control model complexity. Each tree is built sequentially to correct errors made by the previous trees. The update step is calculated using Eq. (48):

$$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + \eta f_t(x_i) \tag{48}$$

Where $\eta$ is the learning rate, $f_t(x_i)$ is the newly trained weak learner at step $t$, and $\hat{y}_i^{(t)}$ is the predicted value at iteration $t$. Figure 7 presents the Intrusion detection System using ML based Models.

Table 6. Hyperparameter tuning of proposed intrusion detection models

| Model | Hyperparameter | Tuned Value |
|---|---|---|
| RF | Number of Trees ($n$) | 100 |
| | Max Depth ($d$) | 20 |
| | Min Samples Split ($s$) | 2 |
| DT | Max Depth ($d$) | 10 |
| | Min Samples Split ($s$) | 2 |
| XGBoost | Learning Rate ($\eta$) | 0.1 |
| | Number of Estimators ($n$) | 150 |
| | Max Depth ($d$) | 6 |
| | Subsample | 0.8 |

To enhance model performance, Grid Search and Random Search were employed to fine-tune hyperparameters. The optimized value of the proposed model is shown in Table 6.

## 4. RESULTS ANALYSIS AND FINDINGS

This section offers a comprehensive description of the dataset, an experimental outline, and the performance analysis of the suggested method along with other existing models using standard parameter metrics. This work is the first effort to calculate the error analysis of all the approaches to recognizing Epileptic Seizures from the dataset. Furthermore, the suggested approach is compared with state-of-the-art methods, which shows that our method performs superiorly and generates outstanding results.

### 4.1. Experiment setup

The experiment has been conducted using Dell Inspiron 14 Plus Laptop with configuration of Intel® Core™ Ultra 7 processor 155H (24 MB cache, 16 cores, 22 threads, up to 4.8 GHz), Windows 11 platform, Intel® Arc™ Graphics card, and 16GB of RAM (2x8GB, LPDDR5X, 6400MT/s). To train and evaluate the ML and DL models, programming tools such as PyTorch, commonly employed for developing deep neural networks (DNN) and Scikit-learn [37] were utilized in the implementation. Also, for incorporating blockchain technology into the implementation, Hyperledger Fabric frameworks have been considered to ensure secure and decentralized data management during the training and testing process.

### 4.2. Experiment outline

The experiment has been conducted in two phases as depicted in Fig. 8. In phase 1, the EGG data generated using smart devices has been encrypted using the ECC algorithm. Before the encryption process, the sender computes the SHA-256 hash of the plaintext data, which serves as a unique identifier. It ensures its integrity during data transmission. Then, the encrypted data gets stored in the cloud and can only be accessed by the authorized user whenever required. It delivers security for the data and access control so that no unauthorized user can access it. Using blockchain mechanisms has minimized the risk of data breaches and offers secure transmission. After that, the ECC technique was used to create a robustness key for the encryption and decryption process. Finally, the hybrid SA-GBO approach helps to select the optimal key so that the entire process can be successfully completed. After this, the decryption process takes place by the hospital administrator to decrypt the cypher text into plaintext using an optimal key. Then, the working of Phase 2 begins. After decryption, the patient data is pre-trained using different techniques in this

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

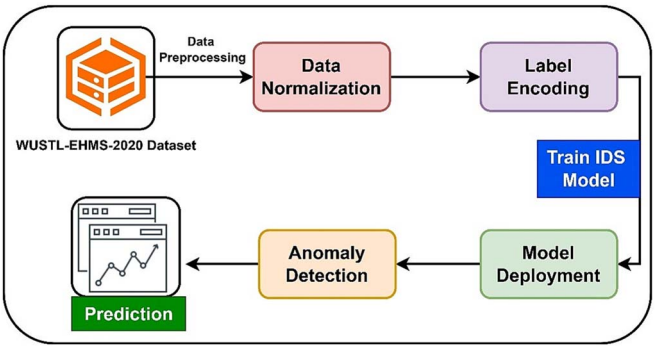



Figure 7. Intrusion detection system using ML based models.

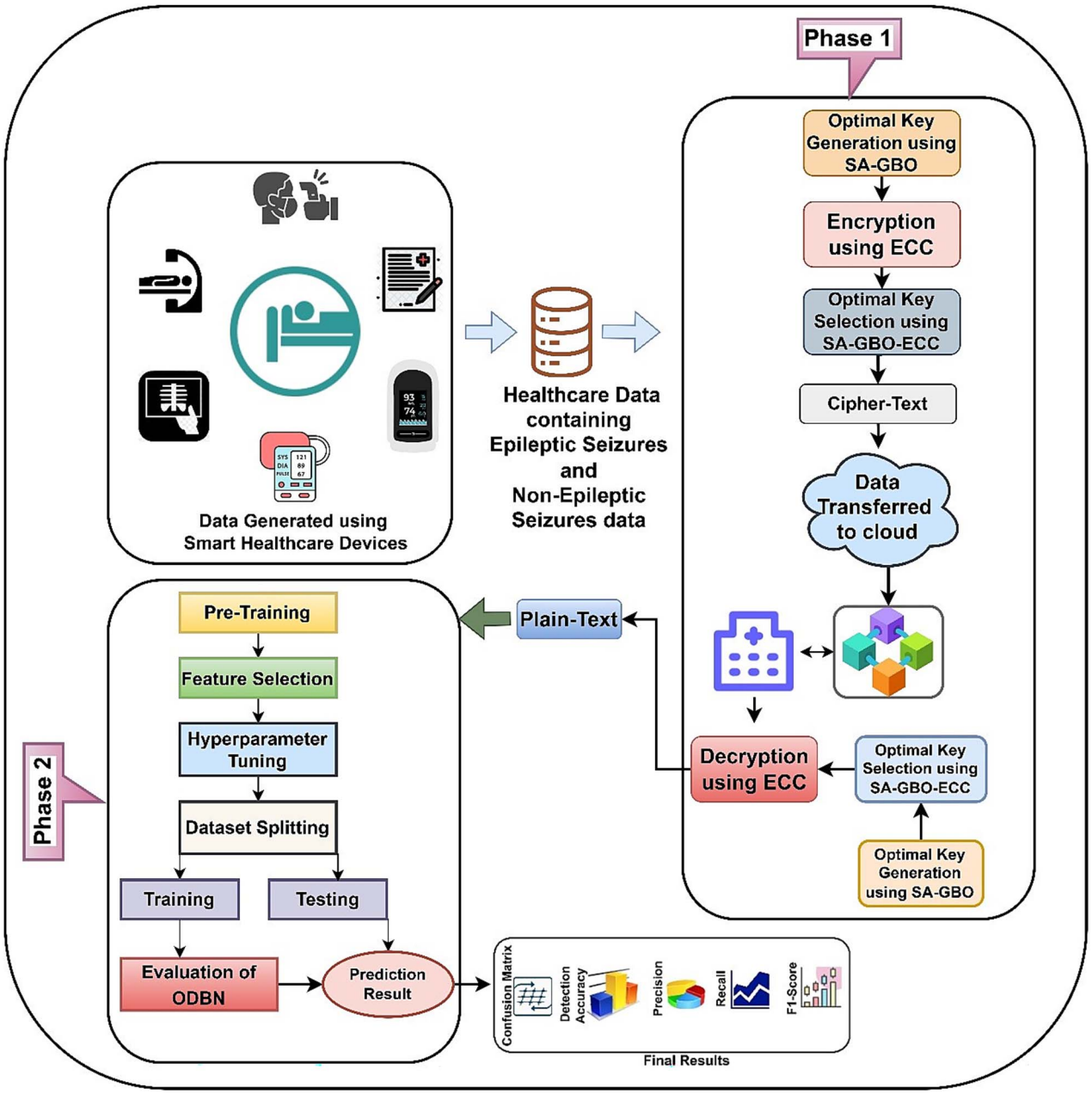



Figure 8. Experimental outline.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

phase. The proposed method has been applied to detect epilepsy disease using EEG datasets.

The First phase outcome displays a detailed performance analysis of the proposed method (SA-GBO-ODBN model) with other techniques like Hill Climbing–Camellia Encryption approach (HC-CE), Side channel analysis–Advanced Encryption Standard (SCA-AES), Gradient-Based Optimizer–Rivest-Shamir-Adleman (GBO-RSA), and Proposed model i.e. SA-GBO-ECC using Dataset 1 and Dataset 2 for securely encrypting and decrypting healthcare data. The second phase outcome compares various models for diagnosing epilepsy diseases using real-time datasets. The various models like RNN (Recurrent Neural Network), ANN (Artificial Neural Network), SVM (Support Vector Machine), and LSTM (Long Short-Term Memory) have been applied along with the Proposed approach (ODBN Model) and are further compared using standard metrics.

## 4.3. Dataset description

The research uses two different publicly available standard datasets to train and test the proposed SA_GBO-ODBN model which are explained as follows:

### 4.3.1. *Bonn University EEG dataset (D1: Dataset 1)*

The dataset (D1) represents the University of Bonn EEG dataset [39]. It comprises **five** diverse sets of single-channel EEG recordings labelled A, B, C, D, and E. Each set involves 100 single-channel EEG readings, spanning 4097 samples with a sampling frequency of 173.61 Hz and a duration of 23.6 s. Set A-B are obtained from **five** healthy people. Set A is recorded while their eyes are open, and Set B is recorded while their eyes are closed using surface EEG recording. Set C and D are collected from the hippocampal formation of the opposite hemisphere of the brain. Finally, Set E was sourced from **five** epileptic patients with seizure activity. Each and every signal was captured using the identical amplifier technique. Every class has 100 signals, each of which is made up of 4097 consecutive samples. These EEG signals are all digitalized at 17 361 samples per second with a 12-bit resolution.

### 4.3.2. *CHB-MIT EEG dataset (D2: Dataset 2)*

The dataset (D2) represents the second dataset, i.e. the CHB-MIT Scalp EEG dataset [40] gathered at the Children's Hospital Boston and MIT. It consists of multi-channel EEG data obtained from 23 epilepsy patients (5 males and 18 Females of ages 1.5–22). The dataset comprises a total of 664 EEG records, with each subgroup containing 9–42 long-term EEG records taken from a single individual. The recordings follow the worldwide 10–20 electrode placement standard and have a sampling frequency of 256 Hz. There were 198 seizures during the 916-hour process that were classified as clinical seizures. The expert physiologists manually record the beginning and end of the seizure following a visual inspection. According to the data, the majority of patients had 23 bipolar EEG signals, which were acquired using the electrodes of the international 10–20 positioning system.

## 4.4. Performance measures

This research has been carried out to offer data security and early epilepsy disease detection using D1 and D2. The performance of the proposed approach has been analyzed using two different metrics. The first phase evaluation metrics include Encryption Time (ET), Decryption Time (DT), Key Generation Time (KGT), Restoration Efficiency (RE), Response Time (RT), and Shared Record Time (SRT), which are explained as follows:

- ***ET:*** It is defined as a time metric that calculates the time duration required by the technique to convert plain text into ciphertext is called ET.
- ***DT:*** It is defined as a time metric that calculates the time duration required by the technique to convert ciphertext into plain text is called DT.
- ***KGT:*** It is defined as the time taken by the technique to generate a cryptographic key for performing secure transactions or other operations. This timeframe has been influenced by factors such as the targeted security level, the key generation method's complexity, and the hardware processing capabilities.
- ***RE:*** It is defined as the metric that measures the model's speed and effectiveness in restoring or recovering data after a system failure or disruption. It is an effective parameter that guarantees network data recovery even after facing potential challenges like data corruption or node failures while managing uninterrupted functioning.
- ***RT:*** It is defined as a time metric that calculates the time the system takes to receive, validate, and respond to a user's request or query. It is affected by key factors like consensus mechanism, network latency, request complexity, and computational efficiency. In blockchain applications, a faster RT is preferred to improve the system's performance and user experience.
- ***SRT:*** It is defined as the time metric that measures the time duration requested by the blockchain network to record and share data or transactions. It is responsible for confirming and validating transactions before they are added to the distributed ledger. Optimizing SRT is crucial for attaining fast transaction processing and managing the overall efficacy of the network.

For the second phase, the following standard metrics, as depicted in Table 7 from *Eq. (49–63)*, are used to assess the performance of the DL model with error analysis to identify disease using two different datasets. This would help analyze the similarities and differences between the various approaches to determine which approach is best for identifying disease.

## 4.5. Time and space complexity analysis

This section highlights the time and space complexity analysis of the proposed model on both datasets which is explained as follows:

***(i) Time complexity***

The time complexity of the proposed model is influenced by the convolutional layers, dense layers, activation functions, and the overall training process. For the convolutional layer with 64 filters and a $1\times1$ kernel, the time complexity is $O(H \times W \times C \times 64)$, where $H$ and $W$ are the height and width of the input, and $C$ is the number of input channels. The ReLU activation function applied to the convolutional output has a time complexity of $O(H \times W \times 64)$, while the Softmax function applied to the dense layer's 10 neurons has a time complexity of $O(10)$. The pooling layer, with a $2 \times 2$ pooling size has a time complexity of $O(H \times W \times 64)$. For the dense layers, the time complexity is $O((H \times W \times 64 \times 100) + (100 \times 10))$. Considering 100 epochs with a batch size of 64, the overall time complexity for training is

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 7. The standard metrics for measuring performance and error analysis on EEG dataset (D1 & D2)

| Metric | Description | Expressions | |
|---|---|---|---|
| Ac | The proportion of correctly classified instances among all instances, providing an overall measure of correctness. | $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ | (49) |
| Pr | The proportion of correctly identified positive instances among all instances classified as +ve, indicating the exactness of the classification. | $Precision = \frac{TP}{TP+FP}$ | (50) |
| Sen | It measures the ability to correctly identify +ve instances. | $Recall/Sensitivity = \frac{TP}{TP+FN}$ | (51) |
| Spe | It measures the ability to correctly identify -ve instances. | $Specificity = \frac{TN}{TN+FP}$ | (52) |
| F1 | The harmonic mean of pr and sen, offering a balanced measure of both metrics. | $F1-Score = \frac{2\times Sen\times \mathbf{Pr}}{Sen+\mathbf{Pr}}$ | (53) |
| MCC | A measure of the correlation between observed and predicted classifications, indicating the effectiveness of the model. | $MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)\ (TN+FP)(TN+FN)}}$ | (54) |
| NPV | The proportion of TN instances among all instances predicted as negative, representing the reliability of -ve predictions. | $NPV = \frac{TN}{TN+FN}$ | (55) |
| FDR | The proportion of FP instances among all instances classified as positive, indicating the rate of false discoveries. | $FDR = \frac{FP}{FP+TP}$ | (56) |
| FOR | The proportion of FN instances among all instances classified as negative, representing the rate of false omissions. | $FOR = \frac{FN}{FN+TN}$ | (57) |
| FPR | The proportion of FP instances among all actual -ve instances, indicating the rate of false alarms. | $FPR = \frac{FP}{FP+TN}$ | (58) |
| FNR | The proportion of FN instances among all actual +ve instances, indicating the rate of missed detections. | $FNR = \frac{FN}{FN+TP}$ | (59) |
| LR+ | The ratio of true positive rate to FPR, indicating the likelihood of a positive test result given the presence of the condition. | $LR+ = \frac{TPR}{FPR}$ | (60) |
| LR- | The ratio of FNR to TNR, indicating the likelihood of a -ve test result given the absence of the condition. | $LR- = \frac{TPR}{FPR}$ | (61) |
| BM | The sum of sen and spe minus one, providing a measure of the model's ability to make correct predictions. | $BM = TPR + TNR - 1$ | (62) |
| MK | The sum of NPV and pre minus one, offering a measure of the relationship between TP and FP rates. | $MK = \mathbf{Pr} + NPV - 1$ | (63) |

Where, Ac- Accuracy, Pr- Precision, Sen- Sensitivity, Spe- Specificity, F1- F1-Score, MCC- Matthews correlation coefficient, NPV- Negative Predictive Value, FDR- False Discovery Rate, FOR- False Omission Rate, FPR- False Positive Rate, FNR- False Negative Rate, LR + - Positive Likelihood Ratio, LR- - Negative Likelihood Ratio, BM- Bookmaker Informedness, MK- Markedness

depicted using *Eq.* (64):

$$O\left(Number\ of\ epochs \times Number\ of\ batches \times [(H \times W \times C \times 64) + (H \times W \times 64) + (100 \times (H \times W \times 64)) + (1000)]\right) \quad (64)$$

***(ii) Space complexity***

The Space Complexity includes the memory required for weights, activations, and intermediate computations. For the convolutional layer, the weight storage is $O(C \times 64)$ and the activations require $O(H \times W \times 64)$. The pooling layer's activations require $O\left(\left(\frac{H}{2}\right) \times \left(\frac{W}{2}\right) \times 64\right)$. In the dense layer, the weight storage is $O((H \times W \times 64 \times 100) + (100 \times 10))$. The total space complexity is calculated using *Eq.* (65):

$$O\left((H \times W \times C \times 64) + (H \times W \times 64) + \left(\left(\frac{H}{2}\right) \times \left(\frac{W}{2}\right) \times 64\right) + (100 \times (H \times W \times 64)) + (1000)\right) \quad (65)$$

The time complexity of the model is determined by the computational demands of the convolutional and dense layers over multiple epochs, while the space complexity is driven by the memory requirements for weights and activations across the network's layers. Table 8 presents Computational Time Analysis of Training and Testing Sets Performance on Dataset D1 and D2.

## 4.6. Performance evaluation of proposed approach with other techniques (Phase 1)

This section presents a comparison of metrics for four different approaches: HC-CE, SCA-AES, GBO-RSA, and Proposed model i.e. SA-GBO-ECC. Cryptographic algorithms like CE, AES, and RSA are fused with other optimization techniques to secure data. These approaches are evaluated based on various metrics for two datasets: Dataset 1 and Dataset 2.

Table 8. Computational time analysis: Training and testing sets performance on dataset D1 and D2

| Datasets | Performance Analysis | Computational Time |
|---|---|---|
| Bonn University EEG Dataset (D1) | Train Set | 44 s 23 ms/Steps |
| | Test Set | 23 s 8 ms/Steps |
| CHB-MIT EEG Dataset (D2) | Train Set | 18 s 21 ms/Steps |
| | Test Set | 8 s 3 ms/Steps |

### 4.6.1. *Performance analysis for dataset 1 (Bonn University EEG dataset)*

To begin with ET, it measures each approach's time to encrypt D1. The Proposed approach has the lowest encryption time of 0.117215 s, indicating that it performs the encryption process quickly compared to other approaches such as HC-CE (0.120665 s), SCA-AES (0.132583 s), GBO-RSA (0.122707 s). Additionally, DT represents the time each approach requires to decrypt the encrypted data. The Proposed approach demonstrates the lowest decryption time of 0.109006 s, indicating that it can decrypt the file more quickly compared to the other techniques like HC-CE (0.120665 s), SCA-AES (0.132583 s), GBO-RSA (0.122707 s). Moreover, the KGT metric represents the time taken by the technique to generate a cryptographic key for performing secure transactions or other operations. The Proposed approach took a minimum key generation time of 0.226221 s against other methods like

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 9. Comparison of the proposed model with other cryptographic approaches for dataset 1 (Bonn University EEG dataset)

| Performance Metrics | RSA | ECC | HC-CE | SCA-AES | GBO-RSA | GBO-ECC | Proposed Method |
|---|---|---|---|---|---|---|---|
| *ET (Sec.)* | 0.135508 | 0.131307 | 0.120665 | 0.132583 | 0.122707 | 0.118923 | **0.117215** |
| *DT (Sec.)* | 0.115768 | 0.113723 | 0.112214 | 0.118381 | 0.113752 | 0.111888 | **0.109006** |
| *KGT (Sec.)* | 0.249898 | 0.237975 | 0.232879 | 0.250964 | 0.236458 | 0.229525 | **0.226221** |
| *RE* | 0.794088 | 0.800186 | 0.808583 | 0.762972 | 0.805311 | 0.811552 | **0.814524** |
| *RT (Sec.)* | 210.0188 | 207.500 | 208.645 | 217.413 | 204.318 | 203.931 | **202.876** |
| *SRT(Sec.)* | 13.9871 | 13.0268 | 13.0403 | 13.5883 | 12.7699 | 12.7199 | **12.6797** |

HC-CE, SCA-AES, and GBO-RSA, which took 0.232879, 0.250964, and 0.236458 s respectively. Subsequently, the RE metric measures the effectiveness of the restoration process for dataset 1, where HC-CE, SCA-AES, and GBO-RSA achieve 0.808583, 0.762972, and 0.805311 values, respectively. The Proposed approach achieves a restoration efficiency of 0.814524, the highest among all the other approaches. This implies that the Proposed approach is more successful in restoring the original records from the encoded data while facing issues related to damages or errors.

Furthermore, RT represents the total time required for the complete encryption, decryption, and restoration processes of dataset 1, where HC-CE, SCA-AES, and GBO-RSA took 208.645, 217.413, and 204.318 s, respectively. The Proposed approach has the lowest response time of 202.876 s, which provides the fastest response processing time for dataset 1. Finally, the SRT parameter indicates the time each approach took to share the encryption key records for dataset 1, where HC-CE, SCA-AES, and GBO-RSA took 13.0403, 13.5883, and 12.7699 s respectively. The Proposed approach has the lowest SRT of 12.6797 s, implying a faster key-sharing process, as depicted in Table 9.

#### 4.6.2. *Performance analysis for dataset 2 (CHB-MIT EEG dataset)*

To begin with ET, it measures each approach's time to encrypt D2. The Proposed approach has the lowest encryption time of 0.117851 s, indicating that it performs the encryption process quickly compared to other approaches such as HC-CE (0.129497 s), SCA-AES (0.128422 s), GBO-RSA (0.120799 s). Additionally, DT represents the time each approach requires to decrypt the encrypted data D2. The Proposed approach demonstrates the lowest decryption time of 0.109597 s, demonstrating that it can decrypt the file more quickly compared to the other techniques like HC-CE (0.126389 s), SCA-AES (0.110944 s), GBO-RSA (0.112338 s). Moreover, The Proposed approach took a minimum key generation time of 0.227449 s against other methods like HC-CE, SCA-AES, and GBO-RSA, which took 0.255886, 0.239366, and 0.233137 s, respectively. Subsequently, the RE metric measures the effectiveness of the restoration process for Dataset 2, where HC-CE, SCA-AES, and GBO-RSA achieve 0.875357, 0.805848, and 0.915728 values, respectively. The Proposed approach achieves a restoration efficiency of 0.934503, the highest among all the approaches. This implies that the Proposed approach is more successful in restoring the original records from the encoded data while facing issues related to damages or errors.

Furthermore, RT represents the total time required for the complete encryption, decryption, and restoration processes of Dataset 2, where HC-CE, SCA-AES, and GBO-RSA took 239.004, 230.876, and 238.22 s, respectively. The suggested approach has the lowest response time of 227.395 s, demonstrating that it provides the fastest overall processing time for File 1. Finally, the SRT parameter indicates the time taken by each approach to share the encryption key records for Dataset 2, where HC-CE, SCA-AES, and GBO-RSA took 14.9378, 14.4297, and 14.8887 s, respectively. The Proposed approach has the lowest SRT of 14.2122 s, implying a faster key-sharing process, as depicted in Table 10.

The Proposed approach consistently outperforms the other approaches regarding ET, DT, KGT, RE, RT, and SRT for both datasets, as depicted in Figs. 9 and 10. It indicates its superiority in various aspects of the encryption and decryption processes. SRT represents the time to share the encryption key records. The Proposed approach demonstrates slightly lower sharing record times than the other approaches for both dataset 1 and dataset 2.

The proposed SA-GBO-ECC model demonstrates clear superiority over the GBO-RSA model due to its enhanced security with shorter key lengths, reducing time and space complexity, which is especially beneficial in resource-constrained environments. Incorporating the SA-GBO enables more adaptive and fine-tuned cryptographic operations, ensuring that ECC processes are optimized in real-time according to computational load and security needs. This dynamic optimization results in consistently lower total processing times and higher restoration efficiency, as demonstrated in experimental datasets. These improvements underscore the inherent advantages of ECC over RSA, particularly when paired with a self-adaptive optimizer, leading to more efficient and secure cryptographic operations, thus justifying the superior performance of the proposed model.

The proposed SA-GBO-ECC model significantly outperforms the GBO-RSA approach, mainly when we analyzed through key performance metrics, as shown in Tables 9 and 10. ECC inherently offers superior security with shorter key lengths than RSA, reducing time and space complexity. This reduction is crucial in environments with limited resources, allowing faster and more efficient cryptographic operations. For instance, the encryption time (ET) for the proposed method is consistently lower than GBO-RSA across both datasets, 0.117215 s against 0.122707 s for Dataset 1 and 0.117851 s in contradiction of 0.120799 s for Dataset 2. The decryption time (DT) follows a similar pattern, with the proposed model achieving faster decryption at 0.109006 s for Dataset 1 and 0.109597 s for Dataset 2, compared to GBO-RSA's 0.113752 s and 0.112338 s, respectively.

Additionally, the Key Generation Time (KGT) is also reduced in the proposed approach, emphasizing its efficiency in generating cryptographic keys, which is a critical factor in secure transactions. For Dataset 1, the KGT is 0.226221 s, while for Dataset 2, it is 0.227449 s, both outperforming GBO-RSA's times of 0.236458 s and 0.233137 s, respectively. RE is another metric where the proposed model stands out, indicating its superior capability to recover original data from encrypted information. The proposed model achieves an RE of 0.814524 for Dataset 1 and an even higher 0.934503 for Dataset 2, surpassing GBO-RSA's 0.805311 and 0.915728, respectively.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 10. Comparison of the proposed model with other cryptographic approaches for dataset 2 (CHB-MIT EEG dataset)

| Performance Metrics | RSA | ECC | HC-CE | SCA-AES | GBO-RSA | GBO-ECC | Proposed Method |
|---|---|---|---|---|---|---|---|
| *ET (Sec.)* | 0.129889 | 0.128503 | 0.129497 | 0.128422 | 0.120799 | 0.118786 | **0.117851** |
| *DT (Sec.)* | 0.118761 | 0.113788 | 0.126389 | 0.110944 | 0.112338 | 0.110855 | **0.109597** |
| *KGT (Sec.)* | 0.245618 | 0.239588 | 0.255886 | 0.239366 | 0.233137 | 0.230044 | **0.227449** |
| *RE* | 0.890582 | 0.900858 | 0.875357 | 0.805848 | 0.915728 | 0.929989 | **0.934503** |
| *RT (Sec.)* | 241.898 | 239.578 | 239.004 | 230.876 | 238.22 | 230.786 | **227.395** |
| *SRT(Sec.)* | 15.8765 | 14.9802 | 14.9378 | 14.4297 | 14.8887 | 14.4168 | **14.2122** |

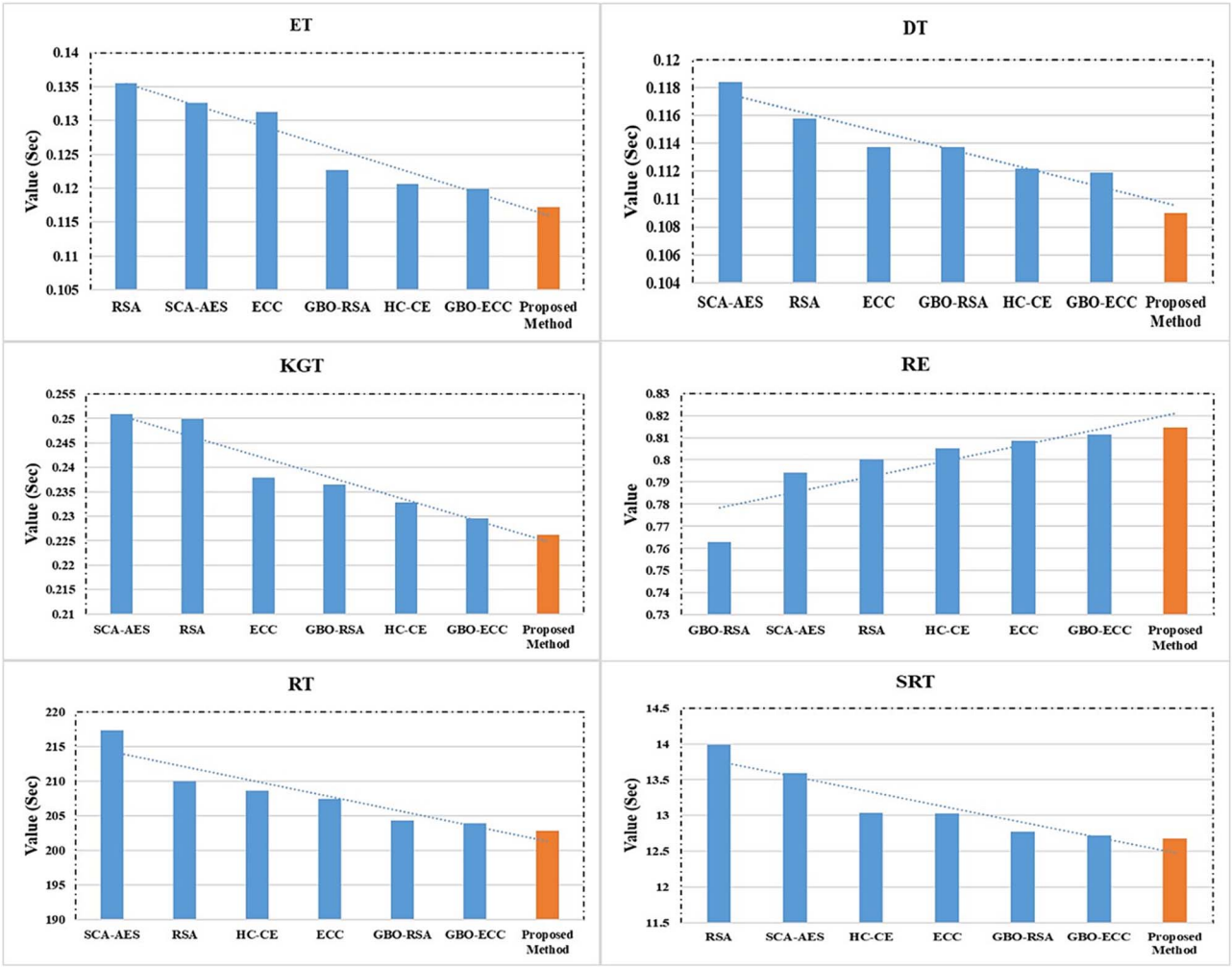



Figure 9. Performance comparison of the proposed approach with other techniques on D1.

Furthermore, the Total Processing Time (RT), which includes the entire process from encryption to decryption and restoration, is minimized with the proposed SA-GBO-ECC model, achieving 202.876 s for Dataset 1 and 227.395 s for Dataset 2, compared to GBO-RSA's 204.318 and 238.22 s. This reduction in RT demonstrates that the proposed approach not only matches but surpasses GBO-RSA in terms of overall processing efficiency. Lastly, the SRT is shorter in the proposed method, reflecting a more efficient key-sharing process essential for timely data communication in secure systems.

Overall, while the performance of GBO-RSA is robust, the proposed SA-GBO-ECC model offers a more advanced and efficient solution, using ECC's inherent advantages and the dynamic optimization provided by SA-GBO to achieve superior results in terms of computational overhead, node complexity, and cryptographic performance. This comprehensive analysis justifies the proposed model's enhanced efficiency and security, making it a superior choice for modern cryptographic applications.

Table 11 highlights the fundamental differences between ECC (using Curve25519) and RSA (3072-bit) across critical cryptographic operations. ECC utilizes the elliptic curve equation $y^2 = x^3 + ax^2 + x$, where $a = 486662$ for Curve25519. This allows for efficient scalar multiplication, point addition, and other operations on elliptic curves. In contrast, RSA operates on modular arithmetic and relies on the generation of two large

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

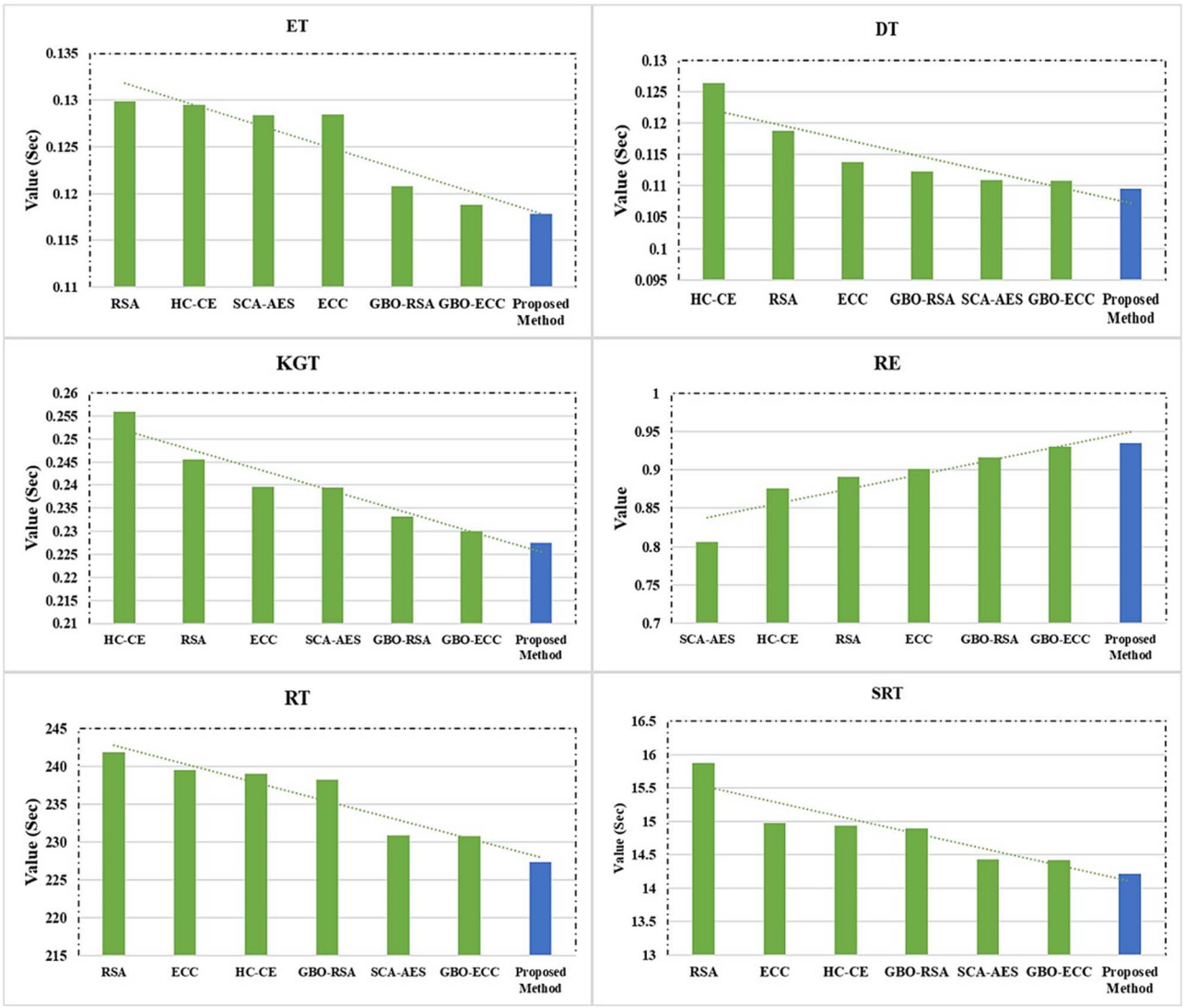



Figure 10. Performance comparison of the proposed approach with other techniques on D2.

prime numbers, which makes it inherently more computationally intensive.

The Key generation in ECC involves selecting a random private key scalar $d$ and deriving the public key as $P = d.G$, where $G$ is a base point on the curve. RSA, however, requires modulus computation, which calculates the totient function and finds modular inverses that significantly increase complexity. Similarly, encryption and decryption processes in ECC using the Elliptic Curve Integrated Encryption Scheme, which is faster and less resource-intensive compared to RSA modular exponentiation ($C = M^e \bmod n$ for encryption and $M = C^d \bmod n$ for decryption). Table 11 demonstrates ECC's efficiency in achieving equivalent security to RSA with much smaller key sizes (256-bit for ECC against 3072-bit for RSA). This not only reduces computational overhead but also makes ECC better suited for resource-constrained environments like IoT.

Table 12 provides a detailed comparison of the time taken by various cryptographic operations in ECC (Curve25519) and RSA (3072-bit). ECC operations, such as point addition, point doubling, and scalar multiplication, are orders of magnitude faster than corresponding modular arithmetic operations of RSA. For example, ECC key generation takes ~2 ms compared to RSA ~4.2 ms due to the latter reliance on prime factorization and modular inverses.

The Encryption and decryption in ECC involve scalar multiplication and shared secret derivation, which is also faster (~2.5 ms) as compared to RSA modular exponentiation, which is ~6.1 and ~5.9 ms for encryption and decryption, respectively. Digital signing and signature verification show a similar trend, with ECC consistently outperforming RSA due to the computational simplicity of elliptic curve arithmetic. Additionally, ECC native support for key agreement protocols like ECDH further highlights its efficiency, requiring only ~2 ms compared to RSA ~4.3 ms when combined with hybrid schemes.

Table 12 highlights the computational advantages of ECC, particularly emphasizing its suitability for environments that require high performance combined with minimal resource utilization. While RSA demonstrates competitive performance in certain operations, owing to its optimized modular arithmetic implementations, ECC consistently outperforms RSA overall.

Table 13 focuses on specific modular arithmetic operations, comparing ECC implicit use of modular arithmetic in point operations with RSA explicit reliance on modular exponentiation and inversion. For ECC, operations like point addition, point

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 11. Comparison of key aspects and operations between ECC-256 (Curve25519) and RSA-3072 cryptographic algorithms

| Aspects | ECC-256 (Curve25519) | RSA-3072 |
|---|---|---|
| ***Elliptic Curve Equation*** | $y^2 = x^3 + ax^2 + x$ (for Curve25519, $a = 486662$) | N/A (RSA does not use curves) |
| ***Key Generation*** | 1. Choose random scalar $d$ as the private key.<br>2. Compute public key $P = d \cdot G$ where $G$ is the base point on Curve25519. | 1. Choose two large primes $p$ *and* $q$.<br>2. Compute $n = p \times q$ (modulus).<br>3. Compute $\varnothing(n) = (p-1)(q-1)$.<br>4. Choose e such that $1 < e < \varnothing(n)$ and $\gcd(e, \varnothing(n)) = 1$.<br>5. Compute $d$ as the modular inverse of $e \bmod \varnothing(n)$. |
| ***Public Key*** | $P = d \cdot G$ (Point on the curve derived by multiplying the base point<br>$G$ with the private scalar $d$) | $(n, e)$ where $n = p \times q$ and $e$ is the public exponent. |
| ***Private Key*** | $d$ (a randomly chosen integer for Curve25519) | $d$ (Calculated from $e$ and $\varnothing(n)$) |
| ***Encryption*** | ***ECIES (Elliptic Curve Integrated Encryption Scheme):***<br>1. Generate a random ephemeral key $k$<br>2. Calculate ephemeral public key<br>$R = k \cdot G$.<br>3. Derive shared secret $S = k \cdot P$.<br>4. Use Sto derive symmetric encryption key for plaintext $M$.<br>5. $C = (Encrypted\ M, R)$ | ***RSA Encryption:*** $C = M^e \bmod n$ m where $M$ is the plaintext message, $C$ is the ciphertext, and $(n, e)$ is the recipient's public key. |
| ***Decryption*** | 1. Derive shared secret $S = d \cdot R$ using recipient's private key $d$ and ephemeral public key $R$.<br>2. Use $S$ to derive the symmetric key to decrypt $C$ and recover $M$. | $M = C^d \bmod n$ m where $C$ is the ciphertext and $(n, d)$ is the recipient's private key. |
| ***Ciphertext*** | $C = (Encrypted\ M, R)$, includes both encrypted message and ephemeral public key. | $C = M^e \bmod n$, a single integer value representing the encrypted message. |

Table 12. Performance comparison of cryptographic operations: ECC (Curve25519) vs. RSA (3072-bit)

| Operation | ECC (Curve25519) | RSA (3072-bit) | Time (ms) ECC | Time (ms) RSA | Difference in Time (ms) |
|---|---|---|---|---|---|
| ***Point Addition*** | $P + Q$, points on the curve | Not applicable | ∼0.001 | N/A | N/A |
| ***Point Doubling*** | $2P$, where $P$ is a point on the curve | Not applicable | ∼0.001 | N/A | N/A |
| ***Scalar Multiplication*** | $k \cdot P$ | Not applicable | ∼1.5 | N/A | N/A |
| ***Key Generation*** | Random scalar and scalar multiplication | Generate primes, calculate $n$, modular inverse | ∼2 | ∼4.2 | ∼2.2 |
| ***Digital Signing*** | ECDSA: involves scalar multiplication | RSA: Modular exponentiation | ∼2.5 | ∼3.7 | ∼1.2 |
| ***Signature Verification*** | Scalar multiplication and point addition | Modular exponentiation | ∼2 | ∼2.8 | ∼0.8 |
| ***Encryption*** | ECIES: Uses shared secret, symmetric encryption | Modular exponentiation with public key | ∼2.5 | ∼6.1 | ∼3.6 |
| ***Decryption*** | Scalar multiplication with private key | Modular exponentiation with private key | ∼2.5 | ∼5.9 | ∼3.4 |
| ***Key Agreement (ECDH)*** | Scalar multiplication for shared secret | Not native; requires DH or hybrid scheme | ∼2 | ∼4.3 (Hybrid) | ∼2.3 |
| ***Hashing to Curve*** | Maps hash to elliptic curve point | RSA does not require this | ∼0.5 | N/A | N/A |

doubling ($2P$), and scalar multiplication ($k.P$) are lightweight and computationally efficient, with times as low as ∼0.001 ms for addition or doubling and ∼ 1.5 ms for scalar multiplication. Conversely, RSA requires operations like modular exponentiation and modular inversion, which are significantly more time-consuming (∼0.07 ms and ∼ 1.2 ms, respectively).

The efficiency of ECC is attributed to its use of smaller key sizes and curve-based arithmetic, which inherently reduces the computational load. While RSA modular arithmetic is robust, the larger key sizes required for equivalent security (3072-bit or higher) result in a higher computation time. For example, the necessary modular inversion for private key computation in RSA takes ∼1.2 ms compared to ∼0.5 ms in ECC.

Table 13 underscores ECC advantage in environments where computational efficiency and low resource consumption are critical. It also highlights why ECC is preferred in IoT, despite RSA competitive performance in modular arithmetic due to algorithmic optimizations.

Table 11 explains the major differences in cryptographic operations, which highlights that ECC (Curve25519) achieves equivalent security to RSA (3072-bit) with significantly smaller key sizes and reduced computational complexity. While RSA relies on resource-intensive modular arithmetic (e.g. modular exponentiation and inversion), ECC employs efficient point-based operations on elliptic curves, which enables faster key generation, encryption, and decryption. Table 12 further quantifies the computational time for various cryptographic operations, which shows consistent advantage of ECC across most operations, such as key generation (∼2 ms for ECC vs. ∼4.2 ms for RSA) and encryption (∼2.5 ms for ECC vs. ∼6.1 ms for RSA). However, the similarities in

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 13. Performance analysis of arithmetic operations in ECC (Curve25519) and RSA (3072-bit)

| Operation | ECC (Curve25519) | RSA (3072-bit) | Time (ms) ECC | Time (ms) RSA | Difference in Time (ms) |
|---|---|---|---|---|---|
| ***Point Addition*** | $P + Q$, points on the curve | Not applicable | ∼0.001 | Not applicable | N/A |
| ***Point Doubling*** | $2P$, where $P$ is a point on the curve | Not applicable | ∼0.001 | Not applicable | N/A |
| ***Scalar Multiplication*** | $k \cdot P$ (Multiply point by integer $k$) | Not applicable | ∼1.5 | Not applicable | N/A |
| ***Modular Exponentiation*** | Not applicable | $M^e \bmod n$ | Not applicable | ∼0.07 | ∼5.5 |
| ***Modular Multiplication*** | Not explicit, done through point operations | $a.b \bmod n$ | Not applicable | ∼0.01 | N/A |
| ***Modular Inversion*** | Inverse of point on the curve (used in decryption) | Inverse of $e \bmod \varnothing(n)$ | ∼0.5 | ∼1.2 | ∼0.7 |

Table 14. Comparison of key and Ciphertext sizes between ECC and RSA algorithms

| Aspect | ECC (256-bit) | RSA (3072-bit) | Difference | Reduction Percentage |
|---|---|---|---|---|
| ***Key Size*** | 32 bytes | 384 bytes | 352 bytes | 91.67% |
| ***Ciphertext Size*** | 64 bytes | 384 bytes | 320 bytes | 83.33% |

performance for certain operations, such as encryption and decryption, arise from the optimized modular arithmetic implementations in RSA, particularly in modern hardware environments. Table 13 narrows the focus to modular arithmetic, which emphasizes the efficiency of ECC in operations like scalar multiplication and modular inversion compared to RSA computationally heavy modular arithmetic. Tables 11–13 collectively justifies the comparative performance metrics, demonstrating ECC efficiency while acknowledging the optimized capabilities of GBO-RSA, explaining the observed performance similarity in specific scenarios.

Table 14 demonstrates the significant reduction in key and ciphertext sizes when using ECC compared to RSA. However, the real advantage of using ECC in our proposed framework lies in its direct impact on performance and resource optimization, especially in resource-constrained environments typical of IoT devices and HSs.

By reducing the key size by 91.67% and the ciphertext size by 83.33%, ECC enables a more efficient use of bandwidth and storage, which are critical in environments with limited resources. This reduction not only improves communication efficiency but also significantly lowers cryptographic operation times, making ECC a superior choice for high-performance systems.

In our proposed framework, ECC is used to ensure secure data transmission without burdening devices with excessive computation or memory requirements. For example, in IoT healthcare devices, where devices often operate on limited battery power and processing capacity, ECC's smaller size allows for faster encryption and decryption times, ensuring real-time secure communication without overloading the system.

Further, in blockchain-based HSs, where each transaction involves securing and verifying sensitive patient data, ECC's smaller ciphertext size enhances transaction throughput and scalability. With reduced storage requirements, the system can handle more transactions per second, thus improving overall system performance and cost-effectiveness.

Therefore, the use of ECC in our framework not only minimizes computational overhead but also enhances scalability and efficiency, offering tangible improvements over RSA in both security and resource utilization. These improvements are particularly significant in the context of IoT and blockchain applications, where both security and efficiency are paramount.

#### 4.6.3. *Theoretical cost analysis of the proposed work*

To illustrate the computational costs and performance metrics of each cryptographic component using SHA-256 and ECC with SA-GBO as shown in Table 15. We evaluate the key operational metrics, including time and space complexity, encryption and decryption times, and key generation efficiency, which highlights the benefits and trade-offs of each algorithm within the framework of secure healthcare data management.

In the proposed work, SHA-256 primarily serves to ensure data integrity, which offers a predictable processing cost that scales with the input data size. This hashing function exhibits low and stable computational overhead with a time complexity of $0(n)$ where $n$ represents the number of 512-bit blocks. This efficiency is particularly strategic in scenarios where rapid data verification is required. ECC using Curve25519 with SA-GBO (Self-Adaptive Gradient-Based Optimizer) further optimizes encryption and decryption by significantly reducing computational overhead. This optimization enhances the efficiency of key generation, encryption, and decryption processes, which yields substantial time and resource savings especially for resource-constrained healthcare IoT environments. Both SHA-256 and ECC with SA-GBO shows high RE and demonstrating minimal data integrity loss after encryption and decryption. Additionally, the bandwidth and storage benefits of ECC arise from its shorter key sizes (256-bit), which makes ECC more suitable for bandwidth-limited and storage-limited devices. This combination of SHA-256 and optimized ECC, as summarized in Table 15, illustrates the computational efficiency and practicality of the model for secure healthcare data management by balancing the robust security with manageable computational demands.

### 4.7. Qualitative comparison of proposed approach with other techniques for disease detection (Phase 2)

This section provides a detailed performance analysis of the proposed approach (ODBN) with other techniques like RNN, ANN, SVM, and LSTM to identify epilepsy using the real-time EEG dataset. The performance has been evaluated using standard parameters like accuracy, precision, recall or sensitivity, specificity, F1-Score, and MCC, depicted in Tables 16 and 17.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 15. Theoretical cost analysis of the proposed work

| Metric | SHA-256 | ECC with SA-GBO (Curve25519) |
|---|---|---|
| ***Key Generation Time (KGT)*** | Not applicable (hashing only) | Optimized: Significantly reduced due to SA-GBO |
| ***Encryption Time (ET)*** | Not applicable (hashing only) | Low: 0 $(\log k)$, optimized by SA-GBO. |
| ***Decryption Time (DT)*** | Not applicable (hashing only) | Low: Same as encryption time, optimized by SA-GBO |
| ***Processing Time per Block*** | 64 rounds per 512-bit block | Optimized: 0 $(\log k)$due to SA-GBO. |
| ***Time Complexity*** | $O(n)$, where $n$ is the number of blocks | 0 $(\log k)$ for both encryption and decryption. |
| ***Space Complexity*** | Constant space (256-bit hash) | Low: Small key sizes (256-bit) |
| ***Security Level*** | 256-bit output (high data integrity) | 256-bit keys |
| ***Bandwidth and Storage*** | Constant 256-bit hash output | Smaller key sizes (256-bit) |
| ***Transmission Cost*** | Low (constant size output) | Low: Reduced key size (256-bit) |
| ***Restoration Efficiency (RE)*** | High (hash integrity verification) | High (maintains data integrity post-encryption and decryption) |
| ***Suitability for IoT Devices*** | Efficient, suitable for resource-constrained devices | Optimized for IoT with small key sizes and low computational overhead |

Table 16. Performance evaluation of the proposed model in contrast to other approaches on dataset 1 (Bonn University EEG dataset) for disease detection

| Qualitative Testing Metrics (Dataset 1) | | | | | |
|---|---|---|---|---|---|
| **Metrics** | **RNN** | **ANN** | **SVM** | **LSTM** | **Proposed Method (ODBN)** |
| **Accuracy** | 0.907739 | 0.889108 | 0.923192 | 0.951231 | **0.989337** |
| **Precision** | 0.869507 | 0.851661 | 0.968802 | 0.943412 | **0.981193** |
| **Sensitivity** | 0.954124 | 0.934541 | 0.874523 | 0.959241 | **0.997656** |
| **Specificity** | 0.861969 | 0.844277 | 0.971217 | 0.943328 | **0.981128** |
| **F-1 Score** | 0.880449 | 0.862378 | 0.892232 | 0.921603 | **0.989281** |
| **MCC** | 0.867076 | 0.85928 | 0.917509 | 0.942388 | **0.987002** |

Table 17. Comparison of the proposed model with other approaches on dataset 2 (CHB-MIT EEG dataset) for disease detection

| Qualitative Testing Metrics (Dataset 2) | | | | | |
|---|---|---|---|---|---|
| **Metrics** | **RNN** | **ANN** | **SVM** | **LSTM** | **Proposed Method (ODBN)** |
| **Accuracy** | 0.947480 | 0.931620 | 0.912499 | 0.976257 | **0.994082** |
| **Precision** | 0.880258 | 0.862192 | 0.955077 | 0.950781 | **0.982722** |
| **Sensitivity** | 0.947530 | 0.979225 | 0.959127 | 0.974477 | **0.992357** |
| **Specificity** | 0.906768 | 0.884646 | 0.866489 | 0.968145 | **0.995917** |
| **F-1 Score** | 0.915705 | 0.903612 | 0.885066 | 0.945849 | **0.987476** |
| **MCC** | 0.862598 | 0.862167 | 0.844472 | 0.937053 | **0.971177** |

#### 4.7.1. *Comparison of the proposed model with other approaches on Dataset 1 for disease detection*

This section has applied five techniques to detect epilepsy using the Bonn dataset. Table 16 depicts the testing metrics for Dataset 1, which shows the performance of all models (namely, RNN, ANN, SVM, LSTM, and ODBN) based on various evaluation metrics.

To begin with, Accuracy measures the percentage of correctly predicted instances, and all models achieve relatively high Accuracy, with ODBN performing the best by achieving an accuracy of 98.93%. In contrast, the other techniques, such as RNN, ANN, SVM, and LSTM, attained 90.77%, 88.91%, 92.31%, and 95.12%, respectively. An accuracy improvement of 8.16%, 10.02%, 6.62% and 3.81% shows that the proposed model performs better than other approaches. Additionally, ODBN achieved 98.11% precision against the RNN (86.95%), ANN (85.16%), SVM (96.88%), and LSTM (94.34%). The proposed approach has attained improved precision of 11.16%, 12.95%, 1.23%, and 3.77% against RNN, ANN, SVM, and LSTM. Moreover, SVM (87.45%) attained the lowest sensitivity rate among other approaches like RNN (95.41%), ANN (93.45%), and LSTM (95.92%). The proposed approach surpassed the result with a high sensitivity rate of 99.76%, with an improvement of 12.31%, 4.35%, 6.31%, and 3.84%.

Subsequently, ODBN (98.11%) achieved the highest specificity with improvements of 11.92%, 13.69%, 0.99%, and 3.78% against RNN (86.19%), ANN (84.42%), SVM (97.12%), and LSTM (94.33%) respectively. Furthermore, with the enhanced rate of 10.88%, 12.69%, 9.7%, and 6.76% against RNN (88.04%), ANN (86.23%), SVM (89.22%), and LSTM (92.16%) respectively, our proposed approach performed better and gained 98.92% of F1-Score. Finally, MCC considers both false positives and true negatives, ranging from −1 to 1, where 1 is the perfect prediction. To rate the effectiveness of classification, our model MCC rate achieved 98.70% with an improvement of 12%, 12.78%, 6.95%, and 4.47% concerning RNN (86.70%), ANN (85.92%), SVM (91.75%), and LSTM (94.23%) respectively.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

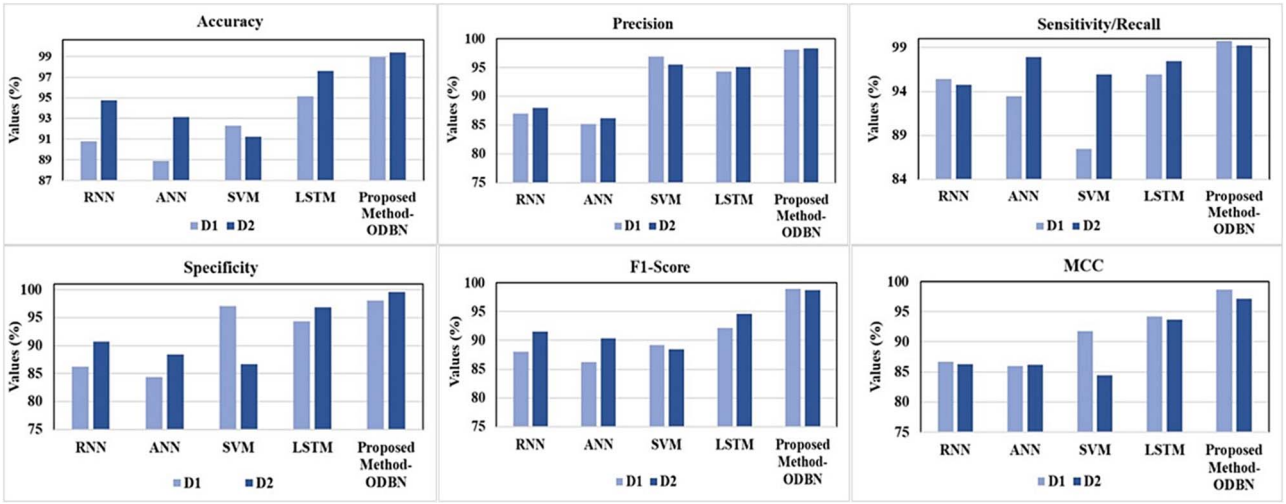



Figure 11. Qualitative evaluation of proposed model in contrast to other techniques in terms of accuracy, precision, sensitivity, specificity, F1-score, and MCC.

### 4.7.2. *Comparison of the proposed model with other approaches on Dataset 2 for disease detection*

This section has applied five techniques to detect epilepsy using the CHB-MIT EEG Dataset. Table 17 depicts the testing metrics for Dataset 2, which shows the performance of all models (namely, RNN, ANN, SVM, LSTM, and ODBN) based on various evaluation metrics.

To begin with, Accuracy measures the percentage of correctly predicted instances, and all models achieve relatively high Accuracy, with ODBN performing the best by achieving an accuracy of 99.40%. In contrast, the other techniques, such as RNN, ANN, SVM, and LSTM, attained 94.74%, 93.16%, 91.24%, and 97.62%, respectively. An accuracy improvement of 4.66%, 6.24%, 8.16%, and 1.78% shows that the proposed model performs better than other approaches. Additionally, ODBN achieved 98.27% precision against the RNN (88.02%), ANN (86.21%), SVM (95.50%), and LSTM (95.07%). The proposed approach has attained improved precision of 10.25%, 12.06%, 2.77%, and 3.2% against RNN, ANN, SVM, and LSTM. Moreover, RNN (94.75%) attained the lowest sensitivity rate among other approaches like ANN (97.92%), SVM (95.91%), and LSTM (97.44%). The proposed approach surpassed the result with a high sensitivity rate of 99.23%, with an improvement of 4.48%, 1.31%, 3.32%, and 1.79%. Subsequently, ODBN (99.59%) achieved the highest specificity with improvements of 8.92%, 11.13%, 12.95%, and 2.78% against RNN (90.67%), ANN (88.46%), SVM (86.64%), and LSTM (96.81%) respectively. Furthermore, with the enhanced rate of 7.17%, 8.38%, 10.24%, and 4.16% against RNN (91.57%), ANN (90.36%), SVM (88.50%) and LSTM (94.58%) respectively, our proposed approach performed better and gained 98.74% of F1-Score. Finally, MCC considers both false positives and true negatives, ranging from −1 to 1, where 1 is the perfect prediction. To rate the effectiveness of classification, our model MCC rate achieved 97.11% with an improvement of 10.86%, 10.9%, 12.67%, and 3.41% concerning RNN (86.25%), ANN (86.21%), SVM (84.44%), and LSTM (93.70%) respectively. Figure 11 depicts the performance of Proposed model against other approaches on both datasets.

### 4.7.3. *Quantitative error rate analysis*

This section demonstrates the error rate analysis of the proposed method against other approaches on Datasets 1 and 2.

***(a) Quantitative error analysis for Dataset 1 (Bonn University)***

This sub-section presents the error analysis of proposed models with other techniques on D1, considering performance metrics like NPV, FDR, FOR, FPR, FNR, LR+, LR-, BM, and MK, as depicted in Table 18, which demonstrates the results attained by different approaches.

From the result as shown in Table 18, the following observations have been noted: To begin with NPV, ODBN achieved a high NPV rate (97.52%) with improvement of 7.05%, 8.91%, 12.44%, and 6.51% against different techniques like RNN (90.47%), ANN (88.61%), SVM (85.08%), and LSTM (91.01%) respectively. Additionally, ODBN with the lowest FDR value of 0.019, FOR of 0.0248, FPR of 0.0199, and FNR of 0.0034 surpasses the performance of other approaches. Moreover, our proposed model performed better by achieving the highest LR+ (52.47) and lowest LR- rate (0.0034) than the other approaches. Subsequently, the ODBN model attained a high BM rate of 97.8% with an improvement of 16.3%, 20%, 13.3%, and 7.6% among RNN (81.5%), ANN (77.8%), SVM (84.5%), and LSTM (90.2%) models respectively. Furthermore, with an improvement of 15.4%, 19%, 11.9%, and 7.4% against RNN (77.3%), ANN (73.7%), SVM (80.8%), and LSTM (85.3%) techniques, respectively, our proposed model surpassed the performance and attained a high MK of 92.7%.

***(b) Quantitative Error Analysis for Dataset 2 (CHB-MIT EEG Dataset)***

This subsection presents the error analysis of proposed models with other techniques on D2 by considering performance metrics, as depicted in Table 18, demonstrating the results of different approaches.

From the result as shown in Table 19, the following observations have been noted: to begin with NPV, ODBN achieved a high NPV rate (97.56%) with improvement of 13.96%, 7.6%, 9.45%, and 7.07% against different techniques like RNN (83.60%), ANN (89.96%), SVM (88.11%), and LSTM (90.49%) respectively. Additionally, ODBN with the lowest FDR value of 0.0173, FOR of 0.0244, FPR of 0.0196, and FNR of 0.0034 surpasses the performance of other approaches. Moreover, our proposed model performed better by achieving the highest LR+ (50.62) and lowest LR- rate (0.0034) than the other approaches. Subsequently, the ODBN model attained a high BM rate of 98.82% with an improvement of 13.4%, 12.44%, 16.27%, and 4.57% among RNN (85.42%), ANN (86.38%), SVM (82.55%), and LSTM (94.25%) models respectively.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 18. Quantitative evaluation of the proposed model in contrast to other approaches on dataset 1 (Bonn University EEG dataset) for disease detection

| Quantitative evaluation (Dataset 1) | | | | | |
|---|---|---|---|---|---|
| **Metrics** | **RNN** | **ANN** | **SVM** | **LSTM** | **Proposed Method (ODBN)** |
| **NPV** | 0.90476 | 0.88619 | 0.8508 | 0.910145 | **0.975215** |
| **FDR** | 0.131 | 0.149 | 0.032 | 0.057 | **0.019** |
| **FOR** | 0.0953 | 0.1138 | 0.1492 | 0.0898 | **0.0248** |
| **FPR** | 0.025041 | 0.024527 | 0.020485 | 0.02671 | **0.019943** |
| **FNR** | 0.006708 | 0.00657 | 0.004241 | 0.00572 | **0.003495** |
| **LR+** | 38.10 | 38.91 | 43.7 | 36.86 | **52.47** |
| **LR-** | 0.0077 | 0.0076 | 0.0043 | 0.0060 | **0.0034** |
| **BM** | 0.815 | 0.778 | 0.845 | 0.902 | **0.978** |
| **MK** | 0.773 | 0.737 | 0.808 | 0.853 | **0.927** |

Table 19. Quantitative evaluation of proposed model in contrast to other approaches on dataset 2 (CHB-MIT EEG dataset) for disease detection

| Quantitative evaluation (Dataset 2) | | | | | |
|---|---|---|---|---|---|
| | **RNN** | **ANN** | **SVM** | **LSTM** | **Proposed Method (ODBN)** |
| **NPV** | 0.836040 | 0.899638 | 0.881173 | 0.904992 | **0.975657** |
| **FDR** | 0.1198 | 0.1379 | 0.045 | 0.0493 | **0.0173** |
| **FOR** | 0.164 | 0.1004 | 0.1189 | 0.0951 | **0.0244** |
| **FPR** | 0.020809 | 0.025437 | 0.024914 | 0.027131 | **0.019634** |
| **FNR** | 0.004308 | 0.006814 | 0.006674 | 0.005810 | **0.003441** |
| **LR+** | 45.552 | 38.551 | 38.518 | 35.955 | **50.622** |
| **LR-** | 0.0047 | 0.0076 | 0.0076 | 0.0059 | **0.0034** |
| **BM** | 0.8542 | 0.8638 | 0.8255 | 0.9425 | **0.9882** |
| **MK** | 0.7162 | 0.7617 | 0.8361 | 0.8599 | **0.9583** |

Furthermore, with an improvement of 24.21%, 19.66%, 12.22%, and 9.84% against RNN (71.62%), ANN (76.17%), SVM (83.61%), and LSTM (85.99%) techniques respectively, our proposed model surpassed the performance and attained high MK of 95.83%.

Figure 12 shows performance analysis using metrics FDR, FOR, FPR, FNR, BM, and MK, where pairs of error analysis metrics have been compared with each other, such as FDR vs FOR, FPR vs FNR, and BM vs MK for both the datasets. For FDR vs FOR, the proposed model performs better than other approaches. The proposed model achieved a lower value of 0.019 for FDR and 0.0248 for FOR against the other method, as depicted in Table 20 using Dataset 1. It gained a lower value of 0.0173 for FDR and 0.0244 for FOR using Dataset 2. Similarly, the proposed approach achieved a lower value for FPR vs FNR and a higher value for BM vs MK compared to other models for both datasets as depicted in Fig. 13. This comprehensive comparison provides a clear understanding of the effectiveness of the ODBN models related to the other applied approaches.

Overall, the metrics suggest that ODBN is the best-performing model, followed by LSTM, among the other models tested on both the datasets D1 and D2. SVM also performs well in terms of precision, while RNN and ANN have lower performance in most metrics. Considering these findings, we can say that the Proposed Optimized DBN (ODBN) has performed admirably compared to other peer Neural Network approaches.

#### 4.7.4. *Qualitative and quantitative comparison of the proposed method with existing approaches*

This section offers a comparative analysis of the proposed model, ODBN, against various existing approaches, which are explained as follows:

Aristizabal *et al.* [41] proposed a unique approach to categorize EEG time series using RNNs with LSTM networks, inspired by the latest advances in deep learning. To facilitate neurologists in the diagnosis of epilepsy, Thara *et al.* [42] suggested automated systems for epilepsy detection. Neurologists use these automated tools as support systems to diagnose epilepsy quickly and accurately. This research attempted to create an automated DNN-based seizure detection system and attained an accuracy of 97.12%. Abdelhameed and Bayoumi [43] introduced an innovative DL method to identify seizures in paediatric patients by categorizing nominally pre-processed raw multi-channel EEG data recordings. The new method utilizes a 2D-DCAE linked to a neural network-based classifier to automatically learn features and create a unified system. Yao *et al.* [44] presented an attention mechanism using BiLSTM to exploit spatial and temporal discriminating features and overcome seizure variabilities. The attention mechanism captures spatial characteristics based on seizure-causing brain areas. Forward and backward temporal features are extracted by BiLSTM. Zhang *et al.* [45] suggested an automatic seizure detection system based on a Bi-GRU neural network (Bi-GRU-NN) to aid epilepsy diagnosis and treatment. Khan *et al.* [46] suggested a low-complexity epileptic seizure detection algorithm for wearable health systems. Single-channel analysis can be used to build a modular multi-channel seizure detector. The algorithm design is more straightforward without feature selection, feature raking, or other feature extraction stage design complexity. Jaishankar *et al.* [47] suggested a new learning method for predicting epileptic seizures. To learn temporal and spatial features from raw EEG data, an adaptive grey wolf optimizer is used for UL. Classification is performed by an adaptive auto-encoder with a genetic algorithm (aADGA) to improve prediction accuracy.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

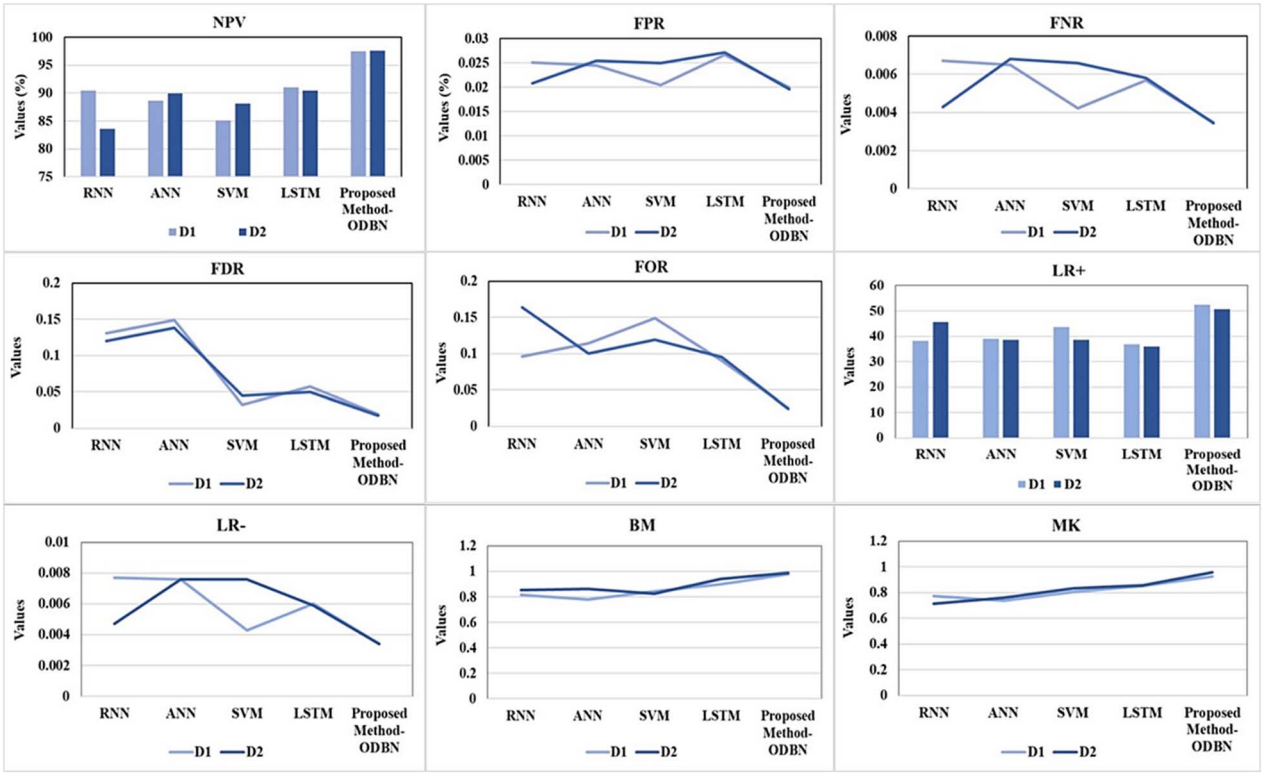



Figure 12. Quantitative evaluation of proposed model in contrast to other techniques in terms of NPV, FPR, FNR, FDR, FOR, LR+, LR-, BM, and MK.

Table 20. Quantitative analysis of different models in terms of specified metrics, including FDR vs FOR, FPR vs FNR, and BM vs MK for D1 and D2

| Parameter | Dataset | RNN | ANN | SVM | LSTM | Proposed Method (ODBN) |
|---|---|---|---|---|---|---|
| ***FDR vs FOR*** | D1 | 0.131 | 0.149 | 0.032 | 0.057 | **0.019** |
| | | 0.0953 | 0.1138 | 0.1492 | 0.0898 | **0.0248** |
| | D2 | 0.1198 | 0.1379 | 0.045 | 0.0493 | **0.0173** |
| | | 0.164 | 0.1004 | 0.1189 | 0.0951 | **0.0244** |
| ***FPR vs FNR*** | D1 | 0.025041 | 0.024527 | 0.020485 | 0.02671 | **0.019943** |
| | | 0.006708 | 0.00657 | 0.004241 | 0.00572 | **0.003495** |
| | D2 | 0.020809 | 0.025437 | 0.024914 | 0.027131 | **0.019634** |
| | | 0.004308 | 0.006814 | 0.006674 | 0.005810 | **0.003441** |
| ***BM vs MK*** | D1 | 0.815 | 0.778 | 0.845 | 0.902 | **0.978** |
| | | 0.773 | 0.737 | 0.808 | 0.853 | **0.927** |
| | D2 | 0.8542 | 0.8638 | 0.8255 | 0.9425 | **0.9882** |
| | | 0.7162 | 0.7617 | 0.8361 | 0.8599 | **0.9583** |

The proposed ODBN model demonstrates superior performance across most metrics as shown in Table 21, especially regarding accuracy, precision, recall and F1-Score, as depicted in the Fig. 14 and quantitative measures comparison in Fig. 15. This highlights its effectiveness in accurately diagnosing diseases in the given medical datasets, making it a promising approach for enhancing healthcare diagnosis systems.

## 4.8. Results analysis of intrusion detection system

This subsection presents the performance evaluation of machine learning-based Intrusion Detection System (IDS) models, specifically RF, DT, and XGBoost. The evaluation is based on key classification metrics, including Accuracy, Precision, Recall, F1-Score, FPR, FNR, and ROC_AUC.

Table 22 shows that XGBoost outperforms RF and DT across all evaluation metrics. XGBoost achieves the highest accuracy of 99.1%, demonstrating its superior classification ability. It also records the highest precision (98.7%), recall (98.9%), and F1-score (98.8%), highlighting its balanced performance in correctly identifying both normal and malicious activities. Additionally, XGBoost exhibits the lowest FPR (0.90%) and FNR (1.10%), minimizing both false alarms and missed detections.

RF also performs well, attaining 98.5% accuracy, with a precision of 97.8%, recall of 98.2%, and an F1-score of 98.0%. However, it has slightly higher FPR (1.50%) and FNR (1.80%) than XGBoost, indicating a marginal increase in misclassification rates.

In comparison, DT exhibits the lowest performance among the three models, with an accuracy of 96.2%, precision of 95.1%, recall of 95.6%, and an F1-score of 95.3%. It also records the highest FPR (3.80%) and FNR (4.40%), suggesting

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

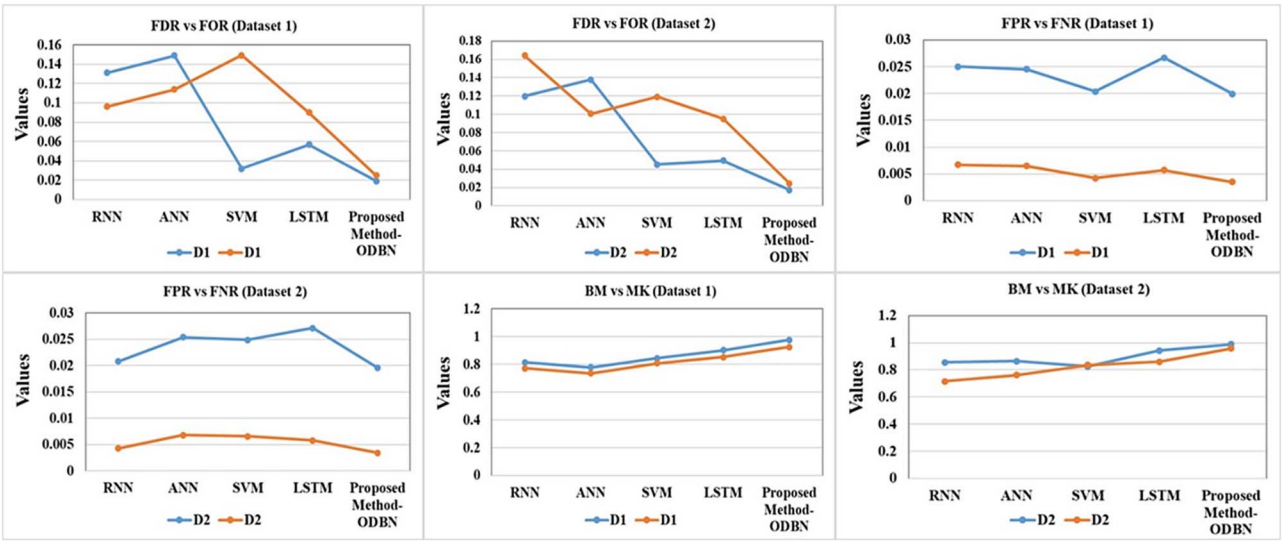



Figure 13. Performance evaluation of proposed model in contrast to other techniques in terms of FDR vs FOR (D1), FDR vs FOR (D2), FPR vs FNR (D1), FPR vs FNR (D2), BM vs MK (D1), and BM vs MK (D2).

Table 21. Qualitative and quantitative comparison of the proposed method with existing approaches

| Ref | Dataset | Proposed Model | Qualitative Metrics | | | | | Quantitative Metrics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ac | Pr | Re | F1 | Spe | FDR | FPR | FNR | LR+ | LR− |
| Aristizabal *et al.* [41], | Bonn | RNN-LSTM | 91.25 | 91.50 | 91.83 | – | 90.50 | 0.0875 | 0.095 | 0.0817 | 9.666 | 0.0962 |
| Thara *et al.* [42] | Bonn | DNN | 97.21 | 94 | 98.17 | 98.48 | 91.47 | 0.060 | 0.0599 | 0.0183 | 16.38 | 0.0200 |
| Abdelhameed and Bayoumi [43], | CHB-MIT | 2D deep convolution autoencoder + Bi-LSTM | 98.7 | 97.7 | 98.8 | – | 97.86 | 0.023 | 0.0214 | 0.012 | 46.16 | 0.0121 |
| Yao *et al.* [44], | CHB-MIT | Attention Mechanism and Bi-LSTM | 83.89 | 85.36 | 83.72 | 83.63 | 84.06 | 0.1464 | 0.1594 | 0.1628 | 5.2521 | 0.1936 |
| Zhang *et al.* [45] | CHB-MIT | Bi-GRU-NN | 98.49 | – | 93.89 | – | 97.49 | – | 0.0251 | 0.0611 | 37.40 | 0.0620 |
| Khan *et al.* [46], | CHB-MIT | Shallow auto encoder framework | 98.85 | 97.8 | 99.1 | – | 97.86 | 0.022 | 0.0214 | 0.0089 | 46.30 | 0.0090 |
| Jaishankar *et al.* [47] | CHB-MIT | aADGA | 97.49 | 96.90 | 95.90 | 98.2 | 96.90 | | 0.031 | 0.041 | 30.93 | 0.0423 |
| **Proposed Method** | **BONN University (D1)** | **ODBN** | **98.93** | **98.11** | **99.76** | **98.92** | **98.11** | **0.019** | **0.0199** | **0.0034** | **52.47** | **0.0034** |
| | **CHB-MIT (D2)** | | **99.40** | **98.27** | **99.23** | **98.74** | **99.59** | **0.0173** | **0.0196** | **0.0034** | **50.62** | **0.0034** |

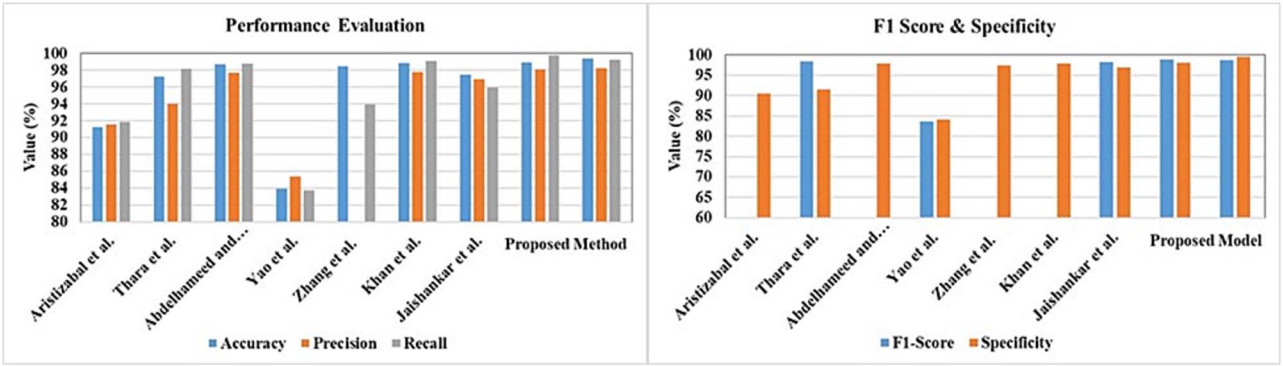



Figure 14. Qualitative comparison of the proposed method with existing approaches.

a greater susceptibility to misclassifying benign and malicious instances.

The ROC_AUC score measures the overall classification capability of the models. XGBoost achieves the highest ROC_AUC of 99.00%, indicating its strong ability to distinguish between normal and attack traffic. RF follows closely with 98.35%, while DT has the lowest ROC_AUC of 95.90%, reinforcing its relatively lower classification performance. Figure 16 depicts the graphical representation of ML-Based Intrusion Detection Models in detecting attacks.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

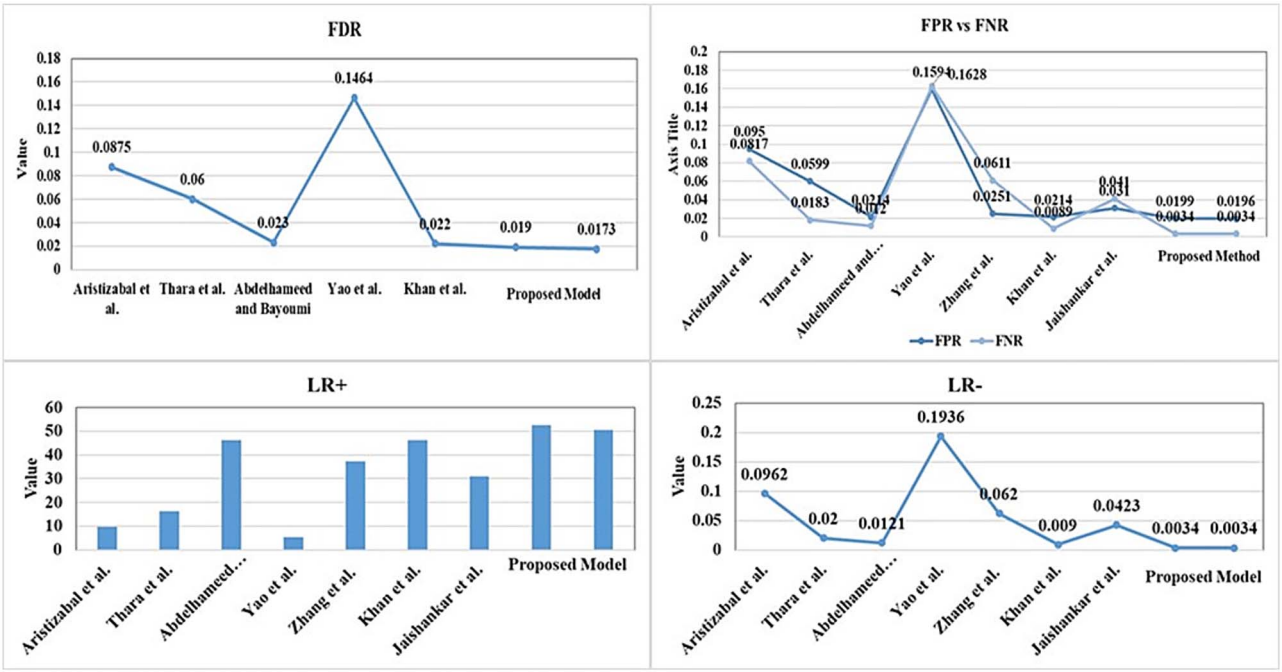



Figure 15. Quantitative comparison of the proposed method with existing approaches.

Table 22. Performance comparison of ML-based intrusion detection models

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | FPR (%) | FNR (%) | ROC_AUC (%) |
|---|---|---|---|---|---|---|---|
| RF | 98.5 | 97.8 | 98.2 | 98 | 1.50 | 1.80 | 98.35 |
| DT | 96.2 | 95.1 | 95.6 | 95.3 | 3.80 | 4.40 | 95.90 |
| XGBoost | 99.1 | 98.7 | 98.9 | 98.8 | 0.90 | 1.10 | 99 |

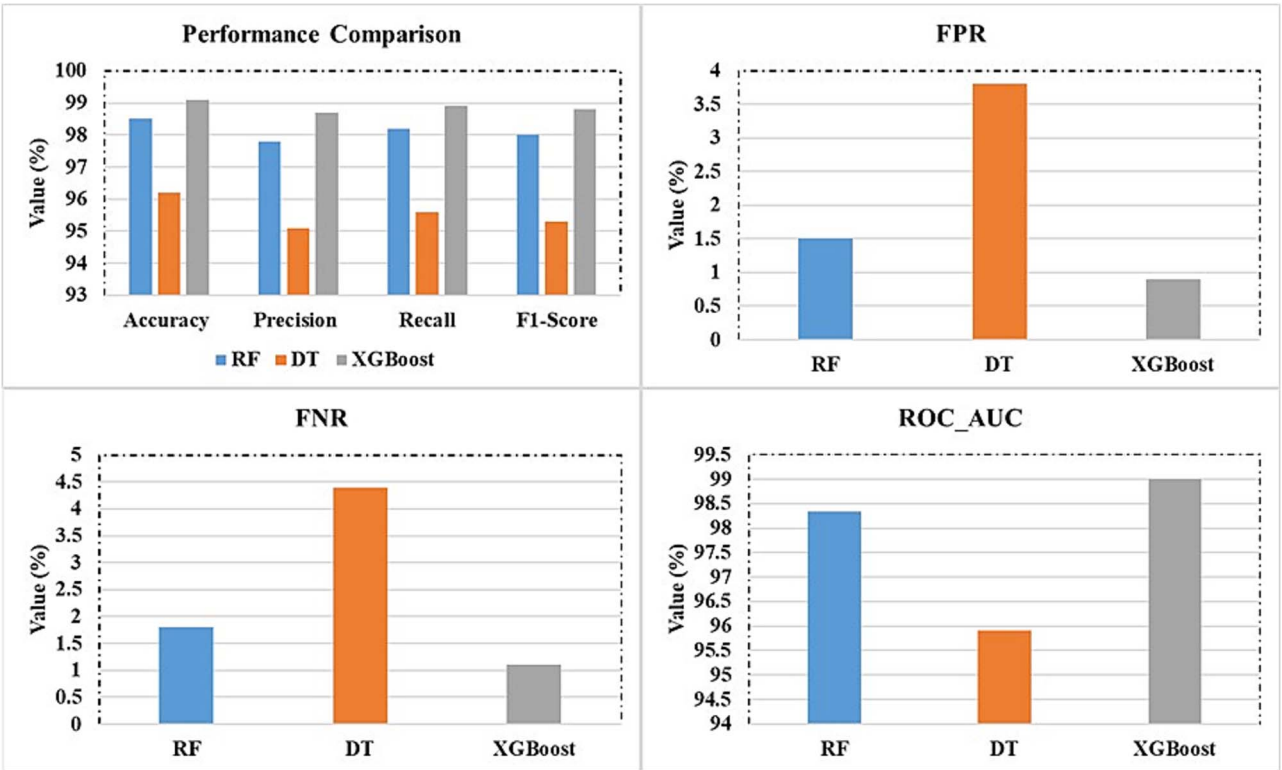



Figure 16. Performance comparison of ML-based intrusion detection models.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

## 4.9. Security and privacy analysis proofing

This section presents the security and privacy verification for the proposed framework and ensures it is effectively secure against potential cyber-attacks.

### 4.9.1. *Security analysis*

We analyzed the security of the framework against common cyber threats, including impersonation attacks, insider attacks, and Man-In-The-Middle (MITM) and replay attacks.

***(i) Impersonation attack***

An impersonation attack occurs when an adversary attempts to authenticate as a legitimate node by providing fraudulent credentials. The verification process follows the following steps:

*a) Request authentication*

The adversary submits a request $R$ with temporary credentials $C_t$, identification $ID_t$, and the sensor's MAC address $MAC_t$ as shown in Eq. (66). A timestamp $T_r$ is generated at the time of request:

$$R = \{C_t, ID_t, MAC_t, T_r\} \tag{66}$$

*b) Verifier authentication*

The verifier checks if the received credentials exist in the system using Eq. (67):

$$(C_t, ID_t, MAC_t) \in D \tag{67}$$

Where $D$ is the database of legitimate users. If $(C_t, ID_t, MAC_t) \notin D$, authentication fails.

*c) Timestamp validation*

The verifier checks whether $T_r$ matches the recorded timestamp $T_{db}$ within an allowable threshold $\delta$ as represented in Eq. (68).

$$|T_r - T_{db}| \leq \delta \tag{68}$$

If the timestamp does not match, the request is rejected.

*d) PBFT consensus validation*

The request is broadcasted to $n$ nodes participating in the PBFT consensus. Each node verifies the identity and signs the request with its key $K_i$ using Eq. (69).

$$S_i = Sign_{K_i}(C_t, ID_t, MAC_t, T_r) \tag{69}$$

If $f+1$ nodes where$f$ is the number of Byzantine nodes as shown in Eq. (70), agree on the request validity, it is accepted:

$$\sum_{i=1}^{n} 1(S_i = valid) > f \tag{70}$$

By ensuring multi-stage authentication and consensus verification, the framework prevents impersonation attacks.

***(ii) Insider attack***

In an insider attack, the adversary has access to valid credentials but still cannot bypass authentication due to strict timestamp validation and PBFT consensus.

*a) Credential verification*

A legitimate user mayj possess $C_t, ID_t, MAC_t$, but authentication is incomplete without a valid timestamp as represented in Eq. (71):

$$\forall((C_t, ID_t, MAC_t) \in D), \quad T_r \notin \mathcal{T} \Rightarrow Authentication\ Failure \tag{71}$$

*b) PBFT-based consensus enforcement*

Even if a malicious insider submits a request, it requires approval from at least $f + 1$ honest nodes as shown in Eq. (72):

$$\sum_{i=1}^{n} 1(S_i = valid) > f \tag{72}$$

Without consensus, the insider attack is thwarted. Thus, insider attacks are mitigated by timestamp-based authentication and multi-node validation.

***(iii) MITM and replay attack***

MITM and replay attacks involve intercepting and reusing messages to gain unauthorized access.

*a) Attack formulation*

The adversary intercepts using Eq. (73).

$$R = (C_t, ID_t, MAC_t, T_r\} \tag{73}$$

To execute the attack, the adversary needs to compute a valid $T'_r$ using brute-force as shown in Eq. (74):

$$T'_r = T_r + \Delta T \tag{74}$$

*b) Computational complexity of timestamp forger*

The computational complexity of timestamp forgery is determined by the entropy of the timestamp, denoted as $H(T)$. To successfully forge a valid timestamp, an adversary must attempt $2^{H(T)}$ possible timestamps, making brute-force attacks highly infeasible. The required computation time for brute-force is calculated using Eq. (75):

$$T_{comp} = \frac{2^{H(T)}}{R} \tag{75}$$

Where $R$ is the adversary's computational rate. As $H(T)$ increases, the number of required computations grows exponentially, making the attack computationally impractical even for powerful adversaries. This ensures that the proposed framework remains secure against timestamp forgery attempts.

*c) PBFT-based validation*

Since each transaction undergoes verification by multiple nodes, the probability of a successful replay attack is calculated using Eq. (76):

$$P_{replay} = \left(\frac{1}{2^{H(T)}}\right)^n \tag{76}$$

For large n, $P_{replay} \approx 0$, making the attack infeasible. Thus, the framework effectively mitigates MITM and replay attacks using PBFT and timestamp validation.

### 4.9.2. *Privacy analysis*

The proposed framework ensures data privacy by preventing unauthorized access and modification using blockchain and PBFT consensus.

*(i) Privacy preservation in PBFT*

*a) Consensus process*

Each transaction $T_i$ undergoes a multi-stage verification as shown in Eq. (77):

$$T_i \xrightarrow{Pre-Prepare} T_i \xrightarrow{Pre-Prepare} T_i \xrightarrow{Pre-Prepare} T_i \tag{77}$$

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Honest nodes $H_n$ reach consensus based on the following rule as shown in Eq. (78):

$$\sum_{i=1}^{n} 1\left(T_i valid\right) > f \tag{78}$$

*b) Blockchain immutability*

Each block $B_i$ contains following attributes as shown in Eq. (79):

$$B_i = \left\{H\left(T_i\right), H\left(B_{i-1}\right), S_i\right\} \tag{79}$$

Altering any transaction requires modifying all subsequent hashes as shown in Eq. (80):

$$H\left(B_i'\right) = H\left(B_i\right) + \epsilon \tag{80}$$

Where $\epsilon$ is an inconsistency that triggers block rejection.

*c) Security against modifications*

A malicious modification requires recalculating all hashes as shown in Eq. (81):

$$\sum_{i=1}^{m} 2^{H(B_i)} \tag{81}$$

This requires infeasible computational power, ensuring privacy.

By employing blockchain's immutability and PBFT consensus, the proposed framework preserves privacy against unauthorized modifications. Thus, by integrating the PBFT consensus mechanism into our framework, we provide a robust privacy-preserving mechanism that maintains the security and trustworthiness of the blockchain network. This approach significantly strengthens the integrity and immutability of the data and makes it extremely difficult for any single entity to modify information without detection.

## 4.10. Performance evaluation of the proposed model in contrast to state-of-the-art

This section provides a detailed comparison between the proposed methodology and State-of-the-art technologies designed for detecting epilepsy disease using publicly available real-time EEG data (Datasets 1 and 2) as depicted in Table 23.

For dataset 1, the following author has proposed different models which are further compared with the suggested approach to detect epilepsy disease. Akyol *et al.* [48] proposed a stacking ensemble-based DNN (SEA-DNN) that addresses the binary detection problem of epilepsy versus non-epileptic for identifying epilepsy. Hassan *et al.* [49] identify epileptic seizures using a signal processing approach called complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN). Initially, the author employed CEEMDAN to break down EEG signal segments into intrinsic mode functions. After that, normal inverse Gaussian (NIG) pdf parameters were utilized to model the mode functions. This work is the first to use NIG modelling in conjunction with CEEMDAN for the identification of epileptic seizures. Abiyev *et al.* [50] recommended DL-based convolutional neural networks (CNN) to improve epileptic seizure identification. Dash and Kolekar [51] examined HMM (Hidden Markov model) classifier-based auto-seizure detection. This study considers seizure and non-seizure episodes as consecutive activities using the HMM classifier. Torse *et al.* [52] suggested a novel approach that uses signal Ensemble Empirical Mode Decomposition (EEMD) and Least Squares Support Vector Machine (LSSVM) to categorize EEG signals into two classes that is seizure and non-seizure. Chowdhury *et al.* [53] classified EEG signals of seizure and non-seizure activities into five groups: seizure start, action, and brain signal recording site. EEG data of gamma-band (40–80 Hz) and theta-band (4–8 Hz) oscillations were recorded for empirical mode decomposition (EMD). Jing *et al.* [54] offer signal enhancement-based categorization and recognition. First, the new sliding window weighting method reduces noise and boosts the target signal. Then, discrete wavelet transform (DWT) is employed to select the suitable frequency band data, and non-zero processing is conducted to enhance features in feature extraction. The retrieved characteristics are then fed into an SVM classifier for classification and recognition. Lenkala *et al.* [55] investigated EEG-based automatic ML techniques for epilepsy diagnosis. A CAD system technique for seizure detection was developed by CIMR *et al.* [56] that maximizes the complexity of the necessary solution while also being reusable on various applications. The data classification is performed by an 8-layer deep CNN specifically created for this purpose. Mishra *et al.* [57] introduced a novel classification method based on DWT and Moth Flame Optimization-based Extreme Learning Machine (DM-ELM) for classifying EEG brain signals for epilepsy detection. A differential attention ResNet-LSTM network was proposed by Qiu *et al.* [58]. ResNet and LSTM networks exhibit spatial correlations and temporal interdependence in the proposed model.

Additionally, a difference layer automatically mines epileptic seizure data. Dastgoshadeh and Rabiei [59] suggest an ML and data mining approach to identify epilepsy seizures. The suggested detection system has three primary steps: DWT is used to pre-process the input signals in the first stage, after which usable information is retrieved from sub-bands. The characteristics of each sub-band are extracted in the second stage using sample entropy and approximate entropy, and the ANOVA test is used to score the features. Lastly, the FSFS technique is used to pick features. The third phase is the classification of seizures using three different algorithms: the Naive Bayes model (NB), KNN, and LS-SVM.

The research proposed the following approaches to detect epilepsy seizures using dataset 2. Park *et al.* [60] suggested a Deep Convolutional Network (DCN) for multi-channel EEG signals that uses 1D and 2D convolutional layers to detect epileptic seizures using spatiotemporal correlation. 2D convolutional layers analyze spatial interactions between EEG channels, while 1D layers consider the temporal evolution of each channel's EEG data. Based on the tensor decomposition of the time-frequency representation (TFR) of EEG signals, Taherished *et al.* [61] offer a novel approach for dimensionality reduction that lowers the dimension of CNN inputs. An extensive collection of the input tensor's slices is transformed into a condensed set of super-slices using the suggested tensor decomposition-based dimensionality reduction technique. Using super-slices reduces the size of the CNN's training inputs and addresses the artefacts and redundancy of the EEG data. Yao *et al.* [62] suggested the independently RNN as an innovative seizure and non-seizure classification model. The novel method gradually extends time scales to extract temporal and spatial information from local time duration to the complete record.

Yang *et al.* [63] introduced a new feature extraction method named MinMaxHist for assessing the topological characteristics of EEG signals. Consequently, MinMaxHist and various other feature extraction methods optimize the EEG signals. Borhade and

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

Table 23. Performance comparison of proposed methodology with state of the art using dataset 1 and dataset 2

| **Dataset 1: BONN University** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Author & Year** | **Method** | ***Ac*** | ***Pr*** | ***Re*** | ***F1*** | ***Spe*** | **MCC** |
| Aristizabal *et al.* [41], 2018 | RNN-LSTM | 91.25 | 91.50 | 91.83 | – | 90.50 | – |
| Thara *et al.* [42], 2019 | DNN | 97.21 | – | 98.17 | 98.48 | 94.93 | 91.96 |
| Akyol *et al.* [48], 2020 | SEA-DNN | 97.17 | – | 93.11 | – | 98.18 | – |
| Hassan *et al.* [49], 2020 | CEEMDAN | 98.67 | – | 98.67 | – | 98.72 | – |
| Abiyev *et al.* [50], 2020 | Deep CNN | 98.67 | – | 97.67 | – | 98.83 | – |
| Dash *et al.* [51], 2020 | HMM | 97 | – | – | – | – | – |
| Torse *et al.* [52], 2021 | EEMD and LSSVM | 94.7 | – | 96.5 | – | 98.3 | – |
| Chowdhury *et al.* [53], 2021 | Gamma & theta band EMD operation, bimodal Gaussian parameters of first IMFs, ANN classifier | 90% | – | – | – | – | – |
| Jing *et al.* [54], 2022 | DWT, non-zero processing, SVM | 96.59 | – | 96.21 | – | 97.39 | – |
| Lenkala *et al.* [55], 2023 | AutoGluon | 0.76 | 0.69 | 0.76 | 0.72 | – | – |
| | Auto-Sklearn | 0.85 | 0.74 | 100 | 0.95 | – | – |
| | Amazon Sagemaker | 0.93 | 0.96 | 0.90 | 0.96 | – | – |
| Cimr *et al.* [56], 2023 | CAD System | 98 | – | 98.5 | – | 98 | – |
| Mishra *et al.* [57], 2023 | DM-ELM | 96 | – | 97 | 94 | 95 | – |
| Qiu *et al.* [58], 2023 | ResNet-LSTM | 90.17 | 90 | 90.16 | 90.05 | – | – |
| Dastgoshadeh and Rabiei [59], 2023 | Approximate and sample entropy + LS-SVM | 98 | – | – | – | – | – |
| **Proposed Method** | **ODBN** | **98.93** | **98.11** | **99.76** | **98.92** | **98.11** | **98.70** |
| **Dataset 2: CHB-MIT dataset** | | | | | | | |
| **Ref** | **Method** | ***Ac*** | ***Pr*** | ***Re*** | ***F1*** | ***Spe*** | **MCC** |
| Park *et al.* [60], 2018 | DCN | 90.5 | – | – | – | – | – |
| Taherisadr *et al.* [61], 2019 | 2D CNN | 89.63 | – | – | – | – | – |
| Yao *et al.* [62], 2019 | CNN | 82.90 | – | – | – | – | – |
| | RNN, | 86.25 | – | – | – | – | – |
| Yang *et al.* [63], 2020 | MinMaxHist technique | 91.80 | – | 93.93 | – | 89.67 | 84.28 |
| Borhade and Nagmode [64], 2020 | MASO algorithm | 94.12 | – | 96.52 | – | 93.73 | – |
| Abdelhameed and Bayoumi [43], 2021 | 2D deep convolution autoencoder + Bi-LSTM | 98.7 | 98.7 | 98.8 | – | – | – |
| Yao *et al.* [44], 2021 | Attention Mechanism and Bi-LSTM | 83.89 | – | 83.72 | – | 84.06 | – |
| Cura and Akan [65], 2021 | SST | 95.1 | – | 90.3 | – | – | – |
| Alharthi *et al.* [66], 2022 | DL model of 1D-CNN, Bi-LSTM | 96.8 | 96.9 | 96.8 | – | – | – |
| Ru *et al.* [67], 2022 | Improved sample entropy and phase synchronization indices | 91.78 | – | 91.27 | – | 93.61 | – |
| Zhang *et al.* [45], 2022 | Bi-GRU-NN | 98.49 | – | 93.89 | – | 98.49 | – |
| Guo *et al.* [68], 2022 | FBAD, AS, and SCBE | 92.62 | – | 95.55 | – | 92.57 | – |
| Mishra *et al.* [57], 2023 | DM-ELM | 96.99 | – | 97.06 | – | 96.89 | – |
| Khan *et al.* [46], 2023 | Shallow auto encoder framework | 98.8 | 98.8 | 99.2 | – | – | – |
| Jaishankar *et al.* [47], 2023 | aADGA | 97.49 | – | 95.9 | – | 96.9 | – |
| Feizbakhsh and Omranpour [69], 2023 | DL + Cluster-based phase space density feature | – | – | 94.94 | – | 94.94 | – |
| **Proposed Method** | **ODBN** | **99.40** | **98.27** | **99.23** | **98.74** | **99.59** | **97.11** |

Nagmode [64] introduced a Modified Atom Search Optimization (MASO) algorithm that combines the Squirrel Search Algorithm with Atom Search Optimization to achieve early and precise seizure prediction based on EEG data. Cura *et al.* [65] utilized the Synchro-squeezed Transform (SST), a high-resolution TFR, to classify epileptic EEG signals. The SST matrices for the EEG data of 16 epilepsy patients during seizures and pre-seizure periods have been computed. Alharthi *et al.* [66] introduced a framework that involves many functions, such as selecting the dominant channel and applying a new algorithm to analyze XLtek EEG data. Ru *et al.* [67] proposed enhanced sample entropy and phase synchronization indices of two intrinsic mode functions (IMFs) using variational mode decomposition, which could mitigate noise's impact on detection performance. Guo *et al.* [68] proposed a hybrid system with an UL module and a SL module, reducing the data labelling workload substantially. UL uses EEG extraction, forest-based anomaly detection, adaptive segmentation (AS), and silhouette coefficient-based evaluation (SCBE) for preliminary seizure screening. Feizbakhsh *et al.* [69] proposed new features to classify distinct brain states in the EEG data of patients with epileptic disease. Each cluster's data densities are regarded as features, and a classifier uses these features as inputs to categorize various brain signals.

## 5. CONCLUSION

This paper proposed a novel SA-GBO-ODBN model using blockchain and deep learning technology for secure medical data handling and diagnosis. It offers a secure and efficient solution for managing and analyzing sensitive medical data, which

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

addresses the challenges associated with traditional healthcare data management systems. The combination of Hyperledger Fabric for secure data storage, ECC for optimal key generation, and SA-GBO for efficient key management provides a robust foundation for data security. Using real-time EEG datasets, the ODBN-based diagnostic model demonstrates higher performance in disease diagnosis. The resulting finding demonstrates that our proposed model is superior to other existing.

approaches in terms of both classification accuracy and security. It achieved an accuracy of 98.93% for the BONN EGG dataset and 99.40% for the CHB-MIT dataset, with an F1-Score of 98.92% and 98.74%, respectively, to detect epilepsy diseases. It offers the lowest FPR value of 0.0199 for Dataset 1 and 0.01963 for Dataset 2, with a lower FNR value of 0.00349 and 0.003441, respectively. While the SA-GBO-ODBN model offers promising results, we will explore its applicability in future diagnosing a more comprehensive range of medical conditions. Moreover, it can be used with federated learning techniques, making it possible to diagnose diseases collaboratively without compromising the data's privacy. Furthermore, we will explore integrating explainable AI (XAI) methods into the DL approaches, which could enhance transparency and trust in the diagnostic process. By focusing on these parameters, we can continuously enhance the reliability and effectiveness of B-AI-powered (Blockchain and AI) HSs.

## AUTHORS' CONTRIBUTIONS

All authors have contributed equally.

## CONFLICT OF INTEREST

Not applicable.

## FUNDING

Not applicable.

## DATA AVAILABILITY

Data will be made available on reasonable request.

## ETHICAL APPROVAL

Not applicable.

## REFERENCES

1. Lai KY, Webster C, Kumari S. *et al.* The nature of cities and the covid-19 pandemic. *Curr Opin Environ Sustain* 2020;**46**:27–31. https://doi.org/10.1016/j.cosust.2020.08.008.
2. Morin O, Vallières M, Braunstein S. *et al.* An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. *Nature Cancer* 2021;**2**:709–22. https://doi.org/10.1038/s43018-021-00236-2.
3. soufiene BO, Bahattab AA, Trad A. *et al.* PEERP: an priority-based energy-efficient routing protocol for reliable data transmission in healthcare using the IOT. *Procedia Computer Science* 2020;**175**:373–8. https://doi.org/10.1016/j.procs.2020.07.053.
4. Xie R, Khalil I, Badsha S. *et al.* Collaborative extreme learning machine with a confidence interval for P2P learning in healthcare. *Comput Netw* 2019;**149**:127–43. https://doi.org/10.1016/j.comnet.2018.11.002.
5. Mani N, Singh A, Nimmagadda SL. An IOT guided healthcare monitoring system for managing real-time notifications by fog computing services. *Procedia Computer Science* 2020;**167**:850–9. https://doi.org/10.1016/j.procs.2020.03.424.
6. Chen Y, Ding S, Xu Z. *et al.* Blockchain-based medical records secure storage and medical service framework. *J Med Syst* 2018;**43**:5. https://doi.org/10.1007/s10916-018-1121-4.
7. Gao S, Piao G, Zhu J. *et al.* TrustAccess: A trustworthy secure ciphertext-policy and attribute hiding access control scheme based on Blockchain. *IEEE Trans Veh Technol* 2020;**69**:5784–98. https://doi.org/10.1109/tvt.2020.2967099.
8. Xu Z, Tang J, Yin C. *et al.* Experience-driven congestion control: when multi-path TCP meets deep reinforcement learning. *IEEE J Sel Areas Commun* 2019;**37**:1325–36. https://doi.org/10.1109/jsac.2019.2904358.
9. Zhao R, Yan R, Chen Z. *et al.* Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing* 2019;**115**:213–37. https://doi.org/10.1016/j.ymssp.2018.05.050.
10. Attaran M. Blockchain technology in healthcare: challenges and opportunities. *International Journal of Healthcare Management* 2020;**15**:70–83. https://doi.org/10.1080/20479700.2020.1843887.
11. Kassab M, DeFranco J, Malas T. *et al.* Exploring research in blockchain for healthcare and a roadmap for the future. *IEEE Trans Emerg Top Comput* 2021;**9**:1835–52. https://doi.org/10.1109/tetc.2019.2936881.
12. Tanwar S, Parekh K, Evans R. Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *J Inf Secur Appl* 2020;**50**:102407. https://doi.org/10.1016/j.jisa.2019.102407.
13. Thanh Nguyen P, Dang Bich Huynh V, Dang Vo K. *et al.* Deep learning based optimal multimodal fusion framework for intrusion detection systems for healthcare data. *Computers, Materials & Continua* 2021;**66**:2555–71. https://doi.org/10.32604/cmc.2021.012941.
14. Thamilarasu G, Odesile A, Hoang A. An intrusion detection system for internet of medical things. *IEEE Access* 2020;**8**:181560–76. https://doi.org/10.1109/access.2020.3026260.
15. Akkaş MA, Sokullu R, Ertürk Çetin H. Healthcare and patient monitoring using IOT. *Internet of Things* 2020;**11**:100173. https://doi.org/10.1016/j.iot.2020.100173.
16. Islam N, Faheem Y, Din IU. *et al.* A blockchain-based fog computing framework for activity recognition as an application to e-healthcare services. *Future Gener Comput Syst* 2019;**100**:569–78. https://doi.org/10.1016/j.future.2019.05.059.
17. Wang Y, Zhang A, Zhang P. *et al.* Cloud-assisted EHR sharing with security and privacy preservation via consortium blockchain. *IEEE Access* 2019;**7**:136704–19. https://doi.org/10.1109/access.2019.2943153.
18. Xu J, Xue K, Li S. *et al.* Healthchain: A blockchain-based privacy preserving scheme for large-scale health data. *IEEE Internet Things J* 2019a;**6**:8770–81. https://doi.org/10.1109/jiot.2019.2923525.
19. Wang Z, Luo N, Zhou P. GuardHealth: Blockchain empowered secure data management and graph convolutional network enabled anomaly detection in smart healthcare. *J Parallel Distrib Comput* 2020;**142**:1–12. https://doi.org/10.1016/j.jpdc.2020.03.004.
20. Zhou T, Shen J, He D. *et al.* Human-in-the-loop-aided privacy-preserving scheme for smart healthcare. *IEEE Transactions on Emerging Topics in Computational Intelligence* 2022;**6**:6–15. https://doi.org/10.1109/tetci.2020.2993841.

Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

21. Hossain MS, Muhammad G. Deep learning based pathology detection for smart connected healthcare. *IEEE Network* 2020;**34**:120–5. https://doi.org/10.1109/mnet.011.2000064.
22. Mohammad Hossein K, Esmaeili ME, Dargahi T. *et al.* BCHealth: A novel blockchain-based privacy-preserving architecture for IOT healthcare applications. *Comput Commun* 2021;**180**:31–47. https://doi.org/10.1016/j.comcom.2021.08.011.
23. Zaabar B, Cheikhrouhou O, Jamil F. *et al.* HealthBlock: A secure blockchain-based healthcare data management system. *Comput Netw* 2021;**200**:108500. https://doi.org/10.1016/j.comnet.2021.108500.
24. Zhang Y, Wei X, Cao J. *et al.* Blockchain-enabled decentralized attribute-based access control with policy hiding for smart healthcare. *Journal of King Saud University—Computer and Information Sciences* 2022;**34**:8350–61. https://doi.org/10.1016/j.jksuci.2022.08.015.
25. Lin G, Wang H, Wan J. *et al.* A blockchain-based fine-grained data sharing scheme for e-healthcare system. *J Syst Archit* 2022;**132**:102731. https://doi.org/10.1016/j.sysarc.2022.102731.
26. Ahmad I, Abdullah S, Ahmed A. IOT-fog-based healthcare 4.0 system using blockchain technology. *The Journal of Supercomputing* 2022;**79**:3999–4020. https://doi.org/10.1007/s11227-022-04788-7.
27. Gupta A, Bhagat M, Jain V. Blockchain-enabled healthcare monitoring system for early monkeypox detection. *The Journal of Supercomputing* 2023;**79**:15675–99. https://doi.org/10.1007/s11227-023-05288-y.
28. Ahmed I, Chehri A, Jeon G. Artificial intelligence and blockchain enabled smart healthcare system for monitoring and detection of COVID-19 in biomedical images. *IEEE/ACM Trans Comput Biol Bioinform* 2024;**21**:814–22. https://doi.org/10.1109/tcbb.2023.3294333.
29. Alruwaili FF, Alabduallah B, Alqahtani H. *et al.* Blockchain enabled smart healthcare system using jellyfish search optimization with dual-pathway deep convolutional neural network. *IEEE Access* 2023;**11**:87583–91. https://doi.org/10.1109/access.2023.3304269.
30. Rahal HR, Slatnia S, Kazar O. *et al.* Blockchain-based multi-diagnosis deep learning application for various diseases classification. *International Journal of Information Security* 2023;**23**:15–30. https://doi.org/10.1007/s10207-023-00733-8.
31. Ahmad I, Zhu M, Li G. *et al.* A secure and interpretable AI for smart healthcare system: A case study on epilepsy diagnosis using EEG signals. *IEEE J Biomed Health Inform* 2024;**28**:3236–47. https://doi.org/10.1109/jbhi.2024.3366341.
32. Saidi R, Rahmany I, Dhahri S. *et al.* A privacy-enhanced framework for chest disease classification using federated learning and Blockchain. *IEEE Access* 2024;1–1. https://doi.org/10.1109/access.2024.3419084.
33. Myrzashova R, Alsamhi SH, Hawbani A. *et al.* Safeguarding patient data-sharing: Blockchain-enabled federated learning in medical diagnostics. *IEEE Transactions on Sustainable Computing* 2024;**10**:176–89. https://doi.org/10.1109/tsusc.2024.3409329.
34. Almalawi A, Khan AI, Alsolami F. *et al.* Managing security of healthcare data for a modern healthcare system. *Sensors* 2023;**23**:3612. https://doi.org/10.3390/s23073612.
35. Kumar P, Kumar R, Gupta GP. *et al.* A blockchain-orchestrated deep learning approach for secure data transmission in IOT-enabled healthcare system. *J Parallel Distrib Comput* 2023;**172**:69–83. https://doi.org/10.1016/j.jpdc.2022.10.002.
36. Hasanova H, Tufail M, Baek U-J. *et al.* A novel blockchain-enabled heart disease prediction mechanism using machine learning. *Computers and Electrical Engineering* 2022;**101**:108086. https://doi.org/10.1016/j.compeleceng.2022.108086.
37. Veeramakali T, Siva R, Sivakumar B. *et al.* An intelligent internet of things-based secure healthcare framework using blockchain technology with an optimal deep learning model. *The Journal of Supercomputing* 2021;**77**:9576–96. https://doi.org/10.1007/s11227-021-03637-3.
38. Hady AA, Ghubaish A, Salman T. *et al.* Intrusion detection system for healthcare systems using medical and network data: A comparison study. *IEEE Access* 2020;**8**:106576–84. https://doi.org/10.1109/access.2020.3000421.
39. S. A, S. SK, Mohan N. *et al.* EEG based automated detection of seizure using machine learning approach and traditional features. *Expert Systems with Applications* 2024;**251**:123991. https://doi.org/10.1016/j.eswa.2024.123991.
40. Amrani G, Adadi A, Berrada M. An explainable hybrid DNN model for seizure vs. non-seizure classification and seizure localization using multi-dimensional EEG signals. *Biomedical Signal Processing and Control* 2024;**95**:106322. https://doi.org/10.1016/j.bspc.2024.106322.
41. Ahmedt-Aristizabal D, Fookes C, Nguyen K. *et al.* Deep classification of epileptic signals. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, 332–335. https://doi.org/10.1109/embc.2018.8512249.
42. D.K. T, B.G P, Xiong F. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recogn Lett* 2019;**128**:544–50. https://doi.org/10.1016/j.patrec.2019.10.029.
43. Abdelhameed A, Bayoumi M. A deep learning approach for automatic seizure detection in children with epilepsy. *Front Comput Neurosci* 2021;**15**:1–15. https://doi.org/10.3389/fncom.2021.650050.
44. Yao X, Li X, Ye Q. *et al.* A robust deep learning approach for automatic classification of seizures against non-seizures. *Biomedical Signal Processing and Control* 2021;**64**:102215. https://doi.org/10.1016/j.bspc.2020.102215.
45. Zhang Y, Yao S, Yang R. *et al.* Epileptic seizure detection based on bidirectional gated recurrent unit network. *IEEE Trans Neural Syst Rehabil Eng* 2022;**30**:135–45. https://doi.org/10.1109/tnsre.2022.3143540.
46. Khan GH, Khan NA, Altaf MA. *et al.* A shallow autoencoder framework for epileptic seizure detection in EEG signals. *Sensors* 2023;**23**:4112. https://doi.org/10.3390/s23084112.
47. Jaishankar B, .M. A A, D. V. *et al.* A novel epilepsy seizure prediction model using deep learning and classification. *Healthcare Analytics* 2023;**4**:100222. https://doi.org/10.1016/j.health.2023.100222.
48. Akyol K. Stacking ensemble based deep neural networks modeling for effective epileptic seizure detection. *Expert Systems with Applications* 2020;**148**:113239. https://doi.org/10.1016/j.eswa.2020.113239.
49. Hassan AR, Subasi A, Zhang Y. Epilepsy seizure detection using complete ensemble empirical mode decomposition with adaptive noise. *Knowledge-Based Systems* 2020;**191**:105333. https://doi.org/10.1016/j.knosys.2019.105333.
50. Abiyev R, Arslan M, Bush Idoko J. *et al.* Identification of epileptic EEG signals using convolutional neural networks. *Applied Sciences* 2020;**10**:4089. https://doi.org/10.3390/app10124089.
51. Dash DP, Kolekar MH. Hidden markov model based epileptic seizure detection using tunable Q wavelet transform. *The Journal of Biomedical Research* 2020;**34**:170–9. https://doi.org/10.7555/jbr.34.20190006.


Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

52. Torse DA, Khanai R. Classification of epileptic seizures using ensemble empirical mode decomposition and least squares support vector machine. In: *2021 International Conference on Computer Communication and Informatics (ICCCI)*. Coimbatore, India: IEEE, 2021. https://doi.org/10.1109/iccci50826.2021.9402307.
53. Chowdhury TT, Fattah SA, Shahnaz C. Seizure activity classification based on bimodal gaussian modeling of the gamma and theta band imfs of EEG signals. *Biomedical Signal Processing and Control* 2021;**64**:102273. https://doi.org/10.1016/j.bspc.2020.102273.
54. Jing J, Pang X, Pan Z. *et al.* Classification and identification of epileptic EEG signals based on signal enhancement. *Biomedical Signal Processing and Control* 2022;**71**:103248. https://doi.org/10.1016/j.bspc.2021.103248.
55. Lenkala S, Marry R, Gopovaram SR. *et al.* Comparison of automated machine learning (automl) tools for epileptic seizure detection using electroencephalograms (EEG). *Comput Secur* 2023;**12**:197. https://doi.org/10.3390/computers12100197.
56. Cimr D, Fujita H, Tomaskova H. *et al.* Automatic seizure detection by convolutional neural networks with computational complexity analysis. *Comput Methods Programs Biomed* 2023;**229**:107277. https://doi.org/10.1016/j.cmpb.2022.107277.
57. Mishra S, Kumar Satapathy S, Mohanty SN. *et al.* A Dm-elm based classifier for EEG brain signal classification for epileptic seizure detection. *Communicative & Integrative Biology* 2022;**16**:1–9. https://doi.org/10.1080/19420889.2022.2153648.
58. Qiu X, Yan F, Liu H. A difference attention resnet-LSTM network for epileptic seizure detection using EEG signal. *Biomedical Signal Processing and Control* 2023;**83**:104652. https://doi.org/10.1016/j.bspc.2023.104652.
59. Dastgoshadeh M, Rabiei Z. Detection of epileptic seizures through EEG signals using entropy features and ensemble learning. *Front Hum Neurosci* 2023;**16**:1–13. https://doi.org/10.3389/fnhum.2022.1084061.
60. Park C, Choi G, Kim J. *et al.* Epileptic seizure detection for multi-channel EEG with deep convolutional neural network. In: *2018 International Conference on Electronics, Information, and Communication (ICEIC)*. Honolulu, HI, USA: IEEE, 2018, https://doi.org/10.23919/elinfocom.2018.8330671.
61. Taherisadr M, Joneidi M, Rahnavard N. EEG signal dimensionality reduction and classification using tensor decomposition and deep convolutional neural networks. In: *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. Pittsburgh, PA, USA: IEEE, 2019. https://doi.org/10.1109/mlsp.2019.8918754.
62. Yao X, Cheng Q, Zhang G-Q. A novel independent rnn approach to classification of seizures against non-seizures. *AMIA Symposium* 2019;1–10. https://arxiv.org/abs/1903.09326.
63. Yang S, Li B, Zhang Y. *et al.* Selection of features for patient-independent detection of seizure events using scalp EEG signals. *Comput Biol Med* 2020;**119**:103671. https://doi.org/10.1016/j.compbiomed.2020.103671.
64. Borhade RR, Nagmode MS. Modified atom search optimization-based deep recurrent neural network for epileptic seizure prediction using electroencephalogram signals. *Biocybernetics and Biomedical Engineering* 2020;**40**:1638–53. https://doi.org/10.1016/j.bbe.2020.10.001.
65. Cura OK, Akan A. Classification of epileptic EEG signals using synchrosqueezing transform and machine learning. *Int J Neural Syst* 2021;**31**:2150005. https://doi.org/10.1142/s0129065721500052.
66. Alharthi MK, Moria KM, Alghazzawi DM. *et al.* Epileptic disorder detection of seizures using EEG signals. *Sensors* 2022;**22**:6592. https://doi.org/10.3390/s22176592.
67. Ru Y, Li J, Chen H. *et al.* Epilepsy detection based on variational mode decomposition and improved sample entropy. *Comput Intell Neurosci* 2022;**2022**:1–11. https://doi.org/10.1155/2022/6180441.
68. Guo Y, Jiang X, Tao L. *et al.* Epileptic seizure detection by cascading isolation Forest-based anomaly screening and EasyEnsemble. *IEEE Trans Neural Syst Rehabil Eng* 2022;**30**:915–24. https://doi.org/10.1109/tnsre.2022.3163503.
69. Feizbakhsh B, Omranpour H. Cluster-based phase space density feature in multichannel scalp EEG for seizure prediction by deep learning. *Biomedical Signal Processing and Control* 2023;**86**:105276. https://doi.org/10.1016/j.bspc.2023.105276.


Downloaded from https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaf045/8123022 by Delhi Technological University user on 06 May 2025

© The British Computer Society 2025. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com
*The Computer Journal*, 2025, 1–38
https://doi.org/10.1093/comjnl/bxaf045
Original Article

Check for updates

*International Journal of Network Management*

WILEY

RESEARCH ARTICLE

# Traffic Engineering Optimization in Hybrid Software-Defined Networks: A Mixed Integer Non-Linear Programming Model and Heuristic Algorithm

Samiullah Mehraban | Rajesh K. Yadav

IoT lab, Dept. of Computer Science and Engineering, Delhi Technological University, Delhi, India

**Correspondence:** Samiullah Mehraban (mehraban780@gmail.com)

**Received:** 17 December 2024 | **Revised:** 7 February 2025 | **Accepted:** 4 April 2025

**Keywords:** HSDN | MLU | SDN | TE

## ABSTRACT

In contemporary networking infrastructure, the integration of software-defined networking (SDN) concepts with traditional networking methodologies has given rise to hybrid software-defined networking (HSDN); as the current network infrastructure evolves, it combines the programmability of SDN with the established protocols of traditional network infrastructure. This paradigm shift introduces novel opportunities and challenges in traffic engineering (TE), necessitating innovative solutions to enhance network performance, resource utilization, and efficiency. In this study, we examined the challenges associated with TE in a hybrid SDN environment, where traditional network devices coexist alongside SDN nodes; we investigated the routing optimization of traffic engineering in a migrated hybrid network and formulated the problem as a mixed integer non-linear programming model. We proposed a heuristic algorithm (H-STE) that optimizes both the OSPF weight setting and the splitting ratio of SDN nodes in the hybrid environment. Extensive evaluations were conducted using real network topologies to validate our method. The results demonstrate that a 30% deployment ratio of SDN nodes significantly improves traffic engineering performance. Specifically, the Maximum Link Utilization (MLU) stabilizes at this ratio, indicating near-optimal network efficiency. This research provides valuable insights for researchers, practitioners, and network architects navigating SDN, hybrid SDN, and traffic engineering optimization.

## 1 | Introduction

The primary challenge in the current network design is network management, wherein both industry and professionals require assistance in managing contemporary networks; several concepts and structures have been proposed globally for future network design. To address these challenges, the software-defined networking (SDN) design was initiated; SDN's fundamental concept is the separation of data and control layers. With its global perspective and centralized management, SDN enables flexible and reliable network management, increasing network throughput and link utilization. Furthermore, it provides a novel flow scheduling approach to enhance traffic engineering (TE). In recent years, SDN has become a prominent topic of discussion in telecommunications. It is a crucial and innovative network architecture that emerged with the advancement of technologies such as mobility, virtualization, and the Internet of Things (IoT).

In recent decades, Internet service provider (ISP) networks have experienced a significant increase in traffic due to the rapid expansion of Internet applications, including games, communication tools, and videos. ISPs have been investing substantially in constructing their network infrastructure to ensure a network with maximum bandwidth and minimum latency; however, this increase in traffic has resulted in frequent congestion, leading to a lack of guaranteed user

© 2025 John Wiley & Sons Ltd.

experience and satisfaction. Due to the exponential advancement of the IoT and the Internet of Everything (IoE), emerging Internet applications have substantially increased network traffic [1]. Consequently, global attention is being directed towards finding solutions to mitigate network congestion and achieve traffic load balance; simultaneously, there is an urgent need to advance network management technologies. Traffic engineering is a highly effective network administrative tool that balances network load and enhances the network performance by optimizing traffic routing methods [2]. TE is an efficient approach that balances the flow, prevents congestion, and improves network performance. TE enables control over how traffic is routed and has gained substantial interest from business and academia in the past few decades. However, due to the network's complex structure, the optimization of traffic engineering remains an open area for research.

SDN is a well-established network design that decouples the network's control and data planes. It offers an efficient and flexible approach to optimize traffic routing from a TE perspective. In an SDN network, policies or routing algorithms can be implemented in the control plane to provide flexible control over the distribution of traffic to any outgoing links in SDN switches, and the SDN switches can distribute traffic to various paths in a flexible and unrestricted manner in TE [3, 4]. SDN networks can achieve more user-oriented, flexible, and intelligent network control by tailoring the forwarding rules for individual flows [5]. The SDN network design is composed of three main layers: The first layer is the data layer, which consists of forwarding network devices that route data packets. The second layer is the control layer, the primary layer of the SDN network architecture; the control layer is mainly responsible for managing the network and translating SDN application layer requests to the SDN data layer. The last layer is the application layer, which is composed of applications. With a centralized controller, this network can control all the networks with programmable devices; centralized control can significantly enhance overall network performance compared to traditional networks regarding traffic engineering, link utilization, and network throughput. It has gained significant worldwide attention in recent years, and the adoption of SDN has substantially increased among network operators. For instance, Microsoft [3] and Google [4] have been striving to achieve nearly 100% utilization of the links. The SDN network provides significant advantages in traffic engineering by allowing dynamic and timely network route adjustments, effectively addressing unexpected issues such as traffic variations and link failures. SDN optimizes TE flexibly, which renders the SDN network a unique and advantageous technology.

Despite significant motivation and numerous advantages, the direct migration to a fully SDN infrastructure presents considerable challenges. Upgrading an entire conventional network to SDN appears risky and costly, particularly for large-scale networks comprising thousands of conventional network devices. This transition faces multiple obstacles, including economic, technical, and security concerns and potential service disruptions. Consequently, implementing hybrid SDN or the incremental deployment of SDN nodes emerges as the most viable approach to address these challenges. A hybrid SDN integrates the flexibility and programmability of SDN with the established reliability of traditional network infrastructure, allowing conventional IP network devices to coexist with SDN nodes. This hybrid architecture represents a network design in which decentralized and centralized networks operate concurrently and maintain communication. There is no need to upgrade all the traditional network devices to the SDN network; we can incrementally migrate to a hybrid SDN and then to a fully SDN network.

In recent years, hybrid software-defined networking has emerged as a prominent network design paradigm. Furthermore, exploring traffic engineering within hybrid SDN environments has garnered significant attention from industry and academia [6]. Open shortest path first (OSPF) in conventional networks remains one of the most widely utilized interior gateway protocols [7]. Within a traditional IP network, a weight (or cost) is assigned to each link, and traffic is routed along the shortest paths (least cost) from source to destination nodes using routing protocols that prioritize shortest path routing, such as OSPF. When multiple shortest paths are available, traffic is uniformly distributed among equal cost multiple paths (ECMPs). Consequently, the efficacy of traffic engineering in a conventional network is directly influenced by the weights assigned to the network. In hybrid SDN environments, conventional network devices are limited to supporting the OSPF protocol. In contrast, SDN nodes operate in hybrid mode, supporting the OSPF protocol for communication with legacy devices and the OpenFlow protocol for interaction with SDN controllers and programmable nodes.

This study focuses on a hybrid SDN architecture in which SDN nodes are incrementally deployed in conventional networks. We concentrate on minimizing the maximum link utilization (MLU) of the hybrid SDN by optimizing two critical aspects: optimizing the OSPF weight settings across the entire network to balance the flows originating from conventional devices and optimizing the traffic splitting ratio of SDN nodes. Optimizing both aspects presents significant challenges as these devices are closely interconnected. The OSPF weight settings are crucial in determining the shortest paths for conventional devices to their destination. When the weights are adjusted, they alter the routes that data packets traverse through the network, which affects the volume of traffic flowing into SDN nodes, subsequently influencing how the traffic is divided among different output paths. As both network devices are interconnected, optimizing both is critical for effective traffic management and ensuring a balanced load across the network. To address this challenge, we introduce a heuristic algorithm, hybrid SDN traffic engineering (H-STE), which aims to minimize MLU more effectively in a hybrid SDN environment. Considering the advantages and cost of SDN nodes, the simulation analysis indicates that with a 30% deployment ratio of SDN nodes, we can achieve nearly optimal performance compared to a fully SDN environment.

This paper addresses the routing optimization in a hybrid SDN environment; we proposed a heuristic algorithm to minimize MLU by optimizing the traffic splitting ratio of SDN nodes and the OSPF weight setting of conventional network devices. Lastly, we evaluate and examine the effects of our suggested method on various topologies and real network traffic datasets. The key contributions of our study are summed up as follows:



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

- For the routing optimization in a hybrid SDN environment, we formulated the problem as a mixed integer non-linear programming problem.
- To minimize MLU in a hybrid SDN, we introduced a heuristic algorithm that optimizes the traffic splitting ratio of SDN nodes and the OSPF weight setting of conventional network devices.
- After evaluating and comparing our proposed method on various topologies and network traffic datasets, our simulation analysis demonstrates that considering cost and SDN features, with minimum SDN node deployment, we can get the most significant benefits, resulting in nearly optimal performance.

The rest of the article is structured as follows: Section 2 presents the related work. Section 3 presents the system description. Section 4 presents the network model and problem formulation. Section 5 introduces the algorithms. Section 6 discusses the experiments and analysis of our method on various network topologies. Finally, Section 7 presents the conclusion.

## 2 | Related Work

This section introduces related works on traffic engineering and routing optimization challenges in different network architectures. Researchers and industries have recently concentrated on optimizing different aspects of hybrid SDN, as different aspects can be optimized during migration to a hybrid SDN. For conventional networks, the traffic is routed on the shortest paths between source and destination under distributed routing protocols such as OSPF and IS-IS. TE is limited in the conventional network as the routing is restricted to the shortest paths; the primary focus of routing optimization algorithms in conventional networks is often on improving the OSPF weight settings. When network failures happen or dynamic traffic demands change, the network suffers performance degradation in the conventional network. Different heuristic algorithms deployed to improve TE performance in conventional networks. Prior research efforts have primarily focused on optimizing routing in pure conventional or fully SDN networks, while these techniques cannot be applied directly to a hybrid SDN environment. For the first time, Agarwal et al. [8] discussed traffic engineering solutions in the hybrid SDN and aimed to maximize network throughput; to determine the best traffic splitting ratio for the SDN nodes, they suggested fully polynomial time approximation schemes (FPTAS). Jia et al. [9] proposed a heuristic algorithm for the incremental deployment of a hybrid SDN to maximize network control ability while minimizing upgrade costs. The simulation result showed that with a 10% upgrading cost, it is possible to control 95% of network flow through SDN devices. Xu et al. [10] proposed a heuristic algorithm for the incremental deployment of hybrid SDN under budget constraints, and they used randomized rounding and depth-first search techniques to solve multicommodity issues in a hybrid SDN to maximize performance. Yang et al. [11] concentrate on the problem of selecting suitable SDN candidates, which involves identifying the least number of legacy routers to upgrade to SDN controllers to defend against any single point of failure within a specified network. The technique they created intends to incrementally convert traditional network devices to a hybrid SDN by limiting the repair path of damaged links while reducing delay and improving network bandwidth utilization. Kar et al. [12] studied the transition of traditional network devices to the hybrid SDN by framing and improving hop/path coverage while reducing cost. They proposed two different heuristic algorithms to model the migration sequence problem concerning path/hop coverage. Their result showed that their algorithms achieve more hop and path coverage with the same investment as previous algorithms. For routing optimization in a hybrid SDN environment, optimizing both the OSPF weight setting and the splitting ratio of SDN nodes is essential. To find the migration sequence for the layer 2 hybrid SDN, Sun et al. [13] discovered the upgrading devices problem and proposed a heuristic approach. Their finding demonstrates the superiority of their proposed approach in minimizing the forwarding path length and accelerating the migration procedure. Guo et al. suggest SOTE in [14] to optimize the OSPF weight setting and SDN node splitting ratio simultaneously in a hybrid SDN; they jointly optimized the OSPF weight setting and SDN node traffic splitting ratio. Yingya et al. [15] formulated the routing optimization problem in a hybrid SDN with path cardinality constraints and proposed H-permissible path routing scheme, which effectively routes traffic flows under path cardinality constraints. Guo et al. [16] formulated the traffic engineering problem in hybrid SDN over multiple traffic matrices by mixing online splitting ratio optimization with offline weight setting optimization. However, these algorithms can only optimize routing over a traffic matrix and are time-consuming for recomputing new routing configurations while using many computational resources.

With the rapid improvement of artificial intelligence technology, much research has been conducted recently to solve TE challenges in the network using machine learning (ML) and deep learning (DL). Xu et al. [17] developed a learning-based traffic engineering approach, TEAL, to optimize WAN traffic engineering by utilizing GPUs' parallel processing capabilities. Yinga et al. [18] proposed an reinforcement learning (RL)-based approach to handle the traffic that is constantly changing to achieve link load balancing in the hybrid SDN network. Liu et al. [19] present DRL-OR, a multi-agent deep reinforcement learning-based online routing algorithm, to enhance the quality of service (QoS) by optimizing traffic flow routing with diverse QoS requirements. Guo et al. [20] proposed a multi-agent RL-based traffic engineering method that promptly chooses the network flow path in hybrid SDNs. Lin et al. [21] proposed a model to integrate the transformer model with deep reinforcement learning to enhance traffic engineering performance in hybrid SDN with fluctuating traffic demands. However, using ML, RL, and DL methods sometimes struggles or neglects to capture temporal relationships between traffic demands during these agent trainings. These approaches struggle to respond quickly to dynamic traffic demands. The trained agent may struggle to produce an appropriate routing methodology for changing traffic demands, resulting in poor traffic engineering performance in dynamic network configuration.

For routing optimization in a hybrid SDN environment, optimizing both the OSPF weight setting and the splitting ratio of SDN nodes is essential. These hybrid network scenarios addressed in the studies differ from the hybrid SDN scenario investigated



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

in our context. In this paper, we formulated a mixed integer non-linear programming problem for routing optimization in a hybrid SDN environment. We proposed an effective heuristic algorithm that optimizes both the traffic splitting ratio of the SDN nodes and the OSPF weight setting of conventional network devices and minimizes MLU more efficiently than the previous methods. We have built the conventional network setup using Mininet and MiniEdit tools, then incrementally deployed SDN nodes into our topology. Selecting the controller in the hybrid SDN is another issue that needs to be considered; in this simulation, we have studied and tested multiple controllers; finally, the open network operating system (ONOS) controller is chosen to be the SDN controller, the integration between legacy network, and the SDN network is carried out through the SDN-IP. This program operates on the ONOS controller, allowing SDN nodes to communicate with legacy routers using BGP protocols.

## 3 | Hybrid SDN Network

This section describes the hybrid SDN discussed in our paper. A hybrid SDN integrates traditional network equipment with SDN nodes to enhance user experience and performance by regulating, modifying, configuring, and managing network behaviors. It combines SDN's centralized management and programmability with conventional networks' reliability and stability. While traditional devices ensure reliable routing, SDN nodes focus on network optimization. Hybrid SDN comprises three main components: conventional network devices, SDN nodes, and controllers. In this architecture, traditional devices manage certain network parts, such as access layers or legacy systems, while centralized controllers oversee other areas, like data centers or virtualized environments. As illustrated in Figure 1, R1–R5 are conventional legacy routers that route traffic based on the shortest path from source to destination, while S1 represents an SDN node controlled by an SDN controller.

Upgrading all conventional network devices to SDN is unnecessary; instead, we can incrementally deploy SDN nodes to transition to a hybrid SDN before fully integrating into a Fully SDN network. Traditional devices utilize distributed routing protocols like IGP and OSPF to forward packets along the shortest path defined by OSPF. In contrast, SDN devices operate in hybrid mode, supporting OSPF and OpenFlow protocols. The SDN controller oversees the forwarding operations of these devices, managing network traffic from a global perspective. It provides flow entries to the flow tables of SDN nodes, which then forward packets according to the established rules. In a hybrid SDN, some traffic is managed by conventional devices while the SDN centralized controller flexibly manages the remainder.

The SDN controller is a critical network component responsible for managing all data plane operations. Its effectiveness directly impacts overall network performance. Selecting the appropriate SDN controller based on network topology is essential, as various options are available. We have thoroughly analyzed several controllers suitable for a hybrid SDN environment, including NoX, POX, Floodlight, Ryu, ODL, and ONOS. After careful consideration, we chose the ONOS controller for our hybrid SDN topology due to its compatibility with hybrid networks and support for SDN-IP applications. SDN-IP integrates traditional and SDN devices, enabling them to connect and exchange information using Border Gateway Protocol (BGP) and Autonomous Systems (AS).

ONOS is designed as a modular-distributed SDN controller, which outperforms other controllers due to better adaptability with thread and socket processing with the provided hardware and has the highest availability among other controllers. ONOS incorporates different subsystems to provide scalability, resilience, and efficient network control. This controller can run on multiple scales and instances to ensure low latency and high availability, which is ideal for large hybrid

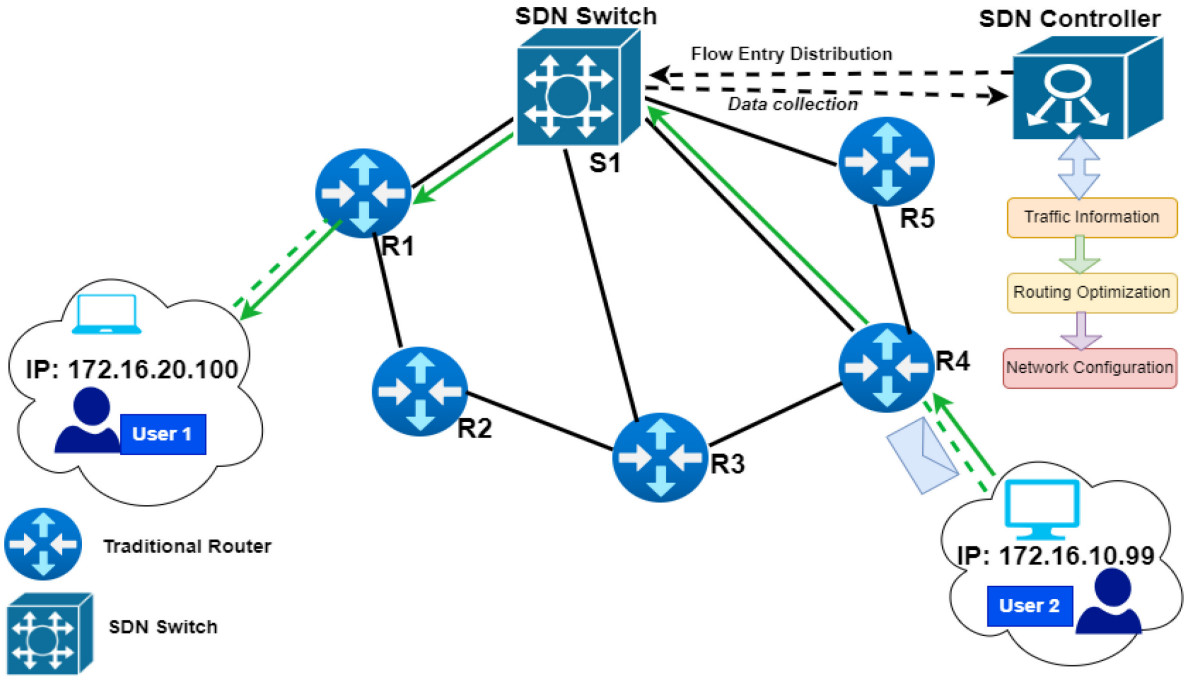



**FIGURE 1** | Hybrid SDN design.



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

SDN deployments. It supports dynamic path optimization and real-time network monitoring, ideal for minimizing network MLUs. This controller stands out due to hybrid network compatibility, scalability, high availability, and advanced TE. Compared to other controllers, ONOS provides the best built-in capabilities for real-time TE, OSPF integration, and dynamic routing.

Several factors must be considered when constructing an SDN network: load balancing, traffic engineering, network performance, security, and link failure recovery. A practical strategy to leverage these aspects is incrementally deploying SDN nodes and controllers alongside conventional networking. This approach enhances overall network performance while minimizing service disruption.

## 4 | Network Model and Problem Formulation

### 4.1 | Network Model and Variable Declaration

We modeled the network as a graph G = (V, E), where V represents the set of all network nodes and E represents the set of links connecting devices to the network. Each node can be a conventional network device (Vl) or an SDN node (Vs), where Vl∩Vs = ∅. For ∀ υ1, υ2 ∈V, υ1 ≠ υ2, for υ1 to υ2, there should be at least one path to reach. C (*e*) represents the link capacity of *e*∈ E. D represents the Traffic Matrices (TMs) set where D = (D1, D2, D3, ....,Dn), and r*i* represents a positive weight coefficient for each TM D*i*. Vc represents the subset of migrated SDN nodes. ω represents the link weight setting of the OSPF protocol for legacy devices and the variable $f$ represents the traffic splitting ratio of SDN nodes. Table 1 summarizes all the sets of notations utilized in this paper for a better understanding.

**TABLE 1** | Key notations.

| G = (V, E) | Undirected graph |
|---|---|
| V | Set of all network nodes |
| E | Set of links in the network |
| C (*e*) | Link capacity of the edge *e* |
| Vl | Legacy node set |
| Vs | SDN node set |
| $e_{ij}$ | A link with nodes for the head (*i*) and tail (*j*) |
| Vc | A subset of the migrated SDN nodes |
| D = (D*i*) | Set of traffic matrices |
| r*i* | Weight coefficient of D*i* |
| U*i* | The value of MLU under D*i* |
| T | Total periods |
| ω | OSPF weight configuration |
| $\phi$ | The weighted MLU |
| τ | Controllable flow |

In this study, we posit that all legacy devices employ conventional IP routing, such as OSPF, which transmits packets according to the shortest path. we configure all nodes as conventional network devices that implement the OSPF protocol. Our objective is to incrementally upgrade legacy nodes to SDN nodes, wherein the splitting ratios are presumed to be determined by the SDN controller to split the flows and increase controllable traffic that can be efficiently managed. The network topology and traffic matrix are considered to be fixed for this experiment; at each stage, the traffic is routed in two ways: A conventional device will route traffic along the shortest path according to the routing table, while SDN nodes will route traffic based on the optimal splitting ratio derived at that stage. We aim to upgrade most legacy devices to SDN while maintaining the overall network load balance while minimizing MLU. The most crucial metric for evaluating load balance in the hybrid SDN is to minimize MLU; in this study, we evaluate the network load balance by considering the MLU. It denotes the percentage utilization of the most heavily loaded link in the network, indicating either overutilization in cases of congestion or underutilization when capacity remains unused.

#### 4.1.1 | Definition

Given a directed graph G = (V, E), a traffic demand matrix TM D*i*, and an integer number of legacy devices; the problem is minimizing the MLU when upgrading most legacy devices to SDN nodes with adaptable flow-splitting capabilities.

### 4.2 | Problem Formulation

We formulate this problem as a mixed integer non-linear programming based on the given network model and definition. We consider the MLU for each device's migration to measure the performance of migration sequences. The primary goal is to minimize the total MLU after the entire migration process (1)–(7). In the first step, we determine the best migration sequence with the migration of each SDN node. We use an optimization algorithm to analyze different migration processes; this algorithm will calculate the MLU for each migration and, ultimately, will determine and select the nodes with a minimum of MLU to improve the overall network performance and minimize the overall network load during the migration process. For conventional network devices, the flows are restricted using the equal cost multi-path (ECMP) rule, which routes the packet based on the shortest paths. To achieve flow balancing, we optimize the OSPF weight setting. As the traffic is splitting along their outgoing links for the SDN nodes, we optimize the splitting ratio of SDN nodes to achieve load balancing.

The network performance improved by minimizing the MLU for each traffic matrix Di. In the hybrid SDN environment, the value of MLU is specified by the traffic splitting ratio of SDN nodes $f$ and the link weight setting of conventional devices ω. As a result, we can formulate the problem as follows:

$$\phi = \text{minimize} \sum_{s=1}^{N} \text{Ui}_{\{\upsilon_1, \upsilon_2, \ldots\ldots \upsilon_s\}} \tag{1}$$

$$\sum_{e \in \text{In}(c)} f^{t}_{e,D_i}(\omega) + \text{Di}(c,t) = \sum_{e \in \text{Out}(c)} f^{t}_{e,D_i}(\omega) \quad c \in \text{Vs}, t \in \text{V}, i \in [1,n] \tag{2}$$



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

$$\sum_{i=1}^{n} r_i = 1, \quad 0 \le r_i \le 1 \tag{3}$$

$$\omega(e) \in N+, \quad \forall e \in E \tag{4}$$

$$0 \le Ui_{\{\upsilon_1, \upsilon_2, \ldots\ldots \upsilon_s\}} \le 1, \forall D_i \in D, i \in [1, n] \tag{5}$$

$$f^t_{e,D_i}(\omega) \ge 0, \quad \forall e \in E, \forall t \in V, \forall D_i \in D, i \in [1, n] \tag{6}$$

$$\sum_{t \in V} f^t_{e,D_i}(\omega) \le Ui_{\{\upsilon_1, \upsilon_2, \ldots\ldots \upsilon_s\}} C(e), \forall e \in E, \forall D_i \in D, i \in [1, n] \tag{7}$$

To determine the best migration sequence, the primary objective is to minimize the total maximum link utilization after the entire migration process (1)–(7); the aim is to identify the best migration sequence S = (υ1, υ2, υ3 ... υn) with a minimum total of MLU. Equation (1) represents the primary objective of the optimization, aiming to minimize where represents the weighted sum of MLU over the entire migration process. Equation (2) represents the flow conservation mechanism in the network at each node v, indicating that the total flow of a device, combined with the flows it generates (traffic demand), should equal the flows that exit this device, which is flow conservation; this equation implies that the traffic demands should be met equivalently. Equation (3) stipulates that the weight ri assigned to each traffic matrix should be a real number (floating-point) within the range [0,1]; link weights are frequently utilized in routing algorithms to define the cost or preference of a particular link. Ensuring these weights are integers facilitates the routing calculations for specific network protocols requiring integer weights. Equation (4) mandates that the link weights are positive integers, essential for practical networks. For proper network operation, the link weights must be non-negative. Equation (5) constrains the MLU values between 0 and 1, which is realistic, where one signifies full link utilization while 0 represents no utilization. Equation (6) ensures that all traffic flows on edge (e) are non-negative; negative flows are not physically meaningful in network flow problems. This constraint guarantees that all flow values are realistic and feasible within the network. The constraints ensure proper flow conservation, respect link capacities, and enforce non-negative traffic flows while meeting traffic demands and integer weight requirements. Equation (7) represents the capacity constraint, which imposes a limitation on link capacity; this equation ensures that the total flow on any edge node should not exceed its capacity, and the flow on each device should remain within its capacity limits, preventing overload that could result in failures or congestion. This optimization problem balances network traffic across multiple traffic matrices, ensuring efficient utilization of network resources while adhering to network constraints.

### 4.3 | Problem Complexity

Minimizing MLU is classified as an NP-complete problem in a hybrid SDN environment. The SDN node's deployment ratios influence the MLU in such a network environment. Given a fixed splitting ratio and predefined link weights, the network MLU can be computed in polynomial time and assessed against a predefined threshold U, allowing for a rapid determination of whether the MLU remains within acceptable limits. However, when link weights are either unknown or constrained to integer values, the complexity of the problem increases significantly. Consequently, minimizing MLU in a hybrid SDN undergoing migration is proven to be more complex than a standard NP-complete problem.

To establish NP-completeness, we employ a polynomial-time reduction from the exact cover by 3-sets (X3C) problem, well-known as NP-complete [22]. By demonstrating that the problem of minimizing MLU in a hybrid SDN can be transformed into the X3C problem, we confirm its computational intractability within polynomial time constraints.

Definition of the X3C problem:

The X3C problem is defined as follows: Given a set 3X containing 3p elements and a collection of subsets 3C, the objective is to determine whether a subfamily of 3C exists that precisely covers all elements of 3X without overlap.

Mapping MLU to the X3C problem:

We construct a flow graph to relate the relationship between MLU minimization and the X3C problem, as illustrated in Figure 2. In this graph, S represents the source node, while t denotes the destination node. The elements of 3X and their corresponding subsets are represented as intermediate nodes. Directed edges connect the subset nodes to the element nodes, indicating which elements belong to which subsets.

This model generates unit flows from the source node S to the destination node t. The flow from S to t must match the exact cover size, denoted as q, so the total flow from S to t equals q. if a valid exact cover exists, the MLU value will be 1, signifying an even distribution of flows across the links. Achieving a total q flow from S to t with an MLU of q confirms that the chosen subset covers 3X without overlap, thereby solving the X3C problem. By demonstrating that the problem of minimizing MLU in a hybrid SDN can be transformed into solving the X3C problem, this proof establishes MLU minimization in NP-complete.

Given a hybrid SDN, if the weights of the links and the flow splitting ratio are known, we can compute the MLU in polynomial time. We can quickly determine if the MLU is below a certain threshold U. We prove that the MLU problem can be reduced to the X3C problem, which is known to be NP-complete. The X3C problem is defined as follows:

Instance: a set X = {x1, x2, ... xn} of n elements (where $n = 3q$) and a collection of C of n three-element subsets of X (each C$i$ contains exactly three elements).

Question: Can you find q pairwise disjoint subsets from C that cover all elements in X together?

We build a flow graph, as shown in Figure 2, to model the relationships between the elements and subsets in the context of the hybrid SDN. Each element $x_j$ in X becomes a node; each subset $C_i$ becomes an SDN node. We added the nodes s t as the source and destination nodes to manage the flows in the



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

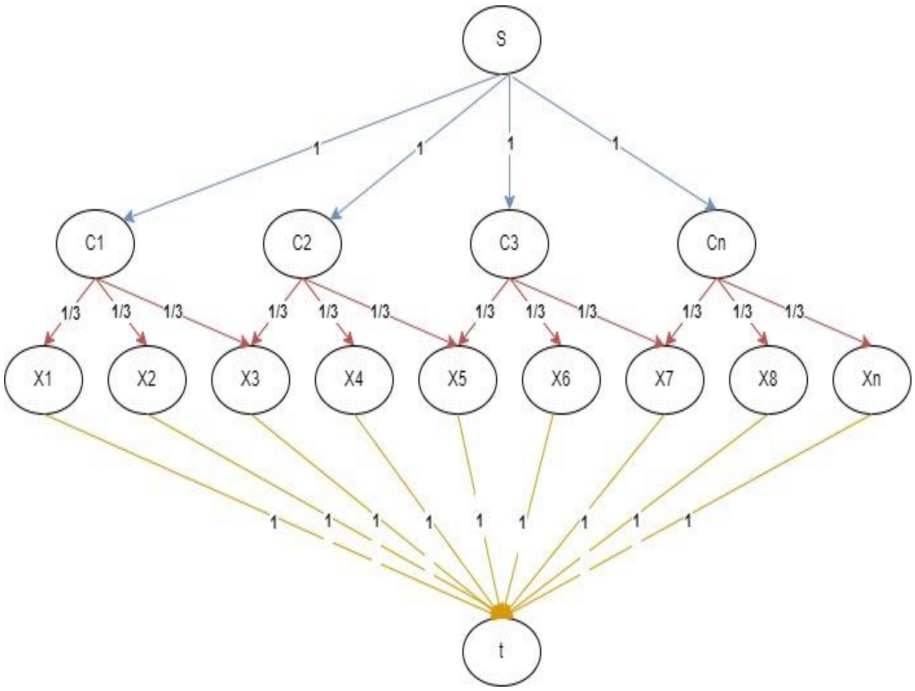



**FIGURE 2** | MLU flow graph.

hybrid SDN. Directed edges from s to each subset $C_i$ (capacity of 1), directed edges from $C_i$ to its respective elements $x_j$ (capacity of 1/3), and the edges from each $x_j$ to the destination node t (capacity of 1). For flow analysis, if a valid cover exists in X3C, we assign unit flows from node $s$ to the selected $C_i$ nodes. Since the flow is distributed evenly, the MLU will be 1, and the total flow from $s$ to $t$ will equal $q$ (the number of subsets chosen). Conversely, if the flow from $s$ to $t$ is $q$ and the MLU is 1, the chosen subsets cover all elements without overlap, confirming a solution to the X3C problem. By showing that the problem of minimizing MLU can be transformed into solving the X3C problem, we conclude that minimizing MLU in a hybrid SDN is NP-complete.

### 4.4 | Routing Optimization

Optimizing routing flows is critical after migrating conventional network devices to SDN and establishing a hybrid SDN environment. In such hybrid networks, conventional network devices mostly rely on the OSPF protocols for the routing, while SDN nodes utilize flow-based routing mechanisms. This dual environment approach necessitates the optimization of two key parameters: the OSPF weight settings for conventional network devices and the traffic splitting ratio of SDN nodes.

To address these challenges, we propose a heuristic approach, H-STE, which leverages a combination of algorithms to achieve offline optimization in the hybrid environment. The H-STE method employs the distributed Floyd–Warshall algorithm to effectively compute optimal paths between source and destination nodes. This algorithm is particularly efficient in identifying the shortest paths in a network with dynamically changing edge weights, making it perfect for hybrid SDN networks where conventional and SDN-based routing coexist.

Additionally, the H-STE method utilizes the Belman-Ford algorithm to optimize the traffic splitting ratio of SDN nodes and compute the shortest path for each node. This algorithm is suitable for scenarios where edge weights may vary dynamically, as it can adapt to traffic conditions and network topology changes. Integrating these algorithms ensures that the splitting ratio of SDN nodes and the OSPF weight settings are optimized, enhancing overall network performance and efficiency in the hybrid SDN environment.

## 5 | Algorithms

In this section, we outline the algorithms used in this experiment. First, we explain the local search heuristic approach designed to optimize flow routing in a hybrid SDN; next, we present the approximation ratios and analyze the complexity of the algorithms.

Flexible SDN nodes coexist with traditional routing protocols in a hybrid SDN environment. The H-STE method helps bridge the gap between the two by optimizing the whole network performance and adhering to the limitations imposed by legacy routing infrastructure. It iteratively adjusts the SDN node's traffic splitting ratios to reduce network congestion, and it has a multi-object focus, as it minimizes MLU and considers other network performance metrics. The algorithm dynamically switches between local refinement and occasional worldwide search perturbations to improve the chance of finding the global optimum.

For routing optimization in hybrid SDN environments, we introduce the heuristic algorithm H-STE, as described in algorithm 1. The algorithm's main objective is to optimize traffic engineering in a hybrid SDN environment by optimizing the traffic splitting ratio of SDN nodes and the OSPF weight setting. This method leverages a combination of algorithms to achieve



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

**ALGORITHM 1** | H-STE

```
          Input: G= (V, E), V_s, C(e), D,
initial_step_size, threshold, increase_
factor, decrease_factor, max_iter_time
           Output: U
 1         Initialize weight-setting ma-
trix ω$$ $$;
 2         U_current = parallel_distributed_
floyd_warshall (G, ω$$ $$, D);
 3         Set U_best = U_current;$$ ;$$
 4         Set ω_best$$ $$= ω;$$ $$
 5         Set Adaptive_step_size = ini-
tial_step_size;
 6         While iteration_times ≤ max_
iter_time do
 7             U_current = splitting_ratio
(ω_best, V_s, D, C(e), G, adaptive_step_size;
 8             If Ucurrent < U_best then
 9                   Update U_best = U_current;
 10                  Update ω_best= ω;
 11           If random (0,1) < global_
search_probability then
 12                  ω = global_pertur-
bation(ωbest);
 13           If last_improvement <
threshold then
 14                  adaptive_step_size* =
increase_factor × random (0.9, 1.1);
 15           Else
 16                  adaptive_step_size* =
decrease_factor × random (0.9, 1.1);
 17           if avg_improvement < small_
threshold then
 18                  max_iter_time -= de-
crease_factor;
 19           else if avg_improvement >
large_threshold then
 20                  max_iter_time += in-
crease_factor;
 21           if distributed_stopping_cri-
terion_met () then
 22                  break;
 23            Apply Flow_Entry_
Aggregation(Ucurrent);
 24            Iteration_times=iteration_
times +1;
 25            Return U = U_best;
```

offline optimization in the hybrid environment. The goal is to minimize MLU in the hybrid SDN environement. As it optimizes both, the optimization process is divided into two phases: the optimization of the OSPF weight setting and the optimization of the traffic Splitting ratio of SDN nodes.

This local search heuristic approach is designed to optimize the splitting ratio of SDN nodes and the OSPF weight setting to minimize MLU in a hybrid SDN environment. In the initialization phase (lines 1–4), the algorithm begins to set the initial weight setting matrix $\omega$. The starting point is based on historical traffic patterns, ensuring a better initial solution. To achieve minimum MLU, we initially utilized a distributed version of the Floyd-Warshal algorithm to find the most efficient routes between each combination of source and destination nodes faster (line 2); this algorithm is particularly efficient in identifying the shortest paths in a network with dynamically changing edge weights, making it perfect for hybrid SDN networks where conventional and SDN-based routing coexist. The current and best MLU values are set (lines 3–4). Following that, we distribute traffic demands along these optimal shortest paths.

The main loop phases (lines 6–22) iterate until the stopping criterion or a maximum iteration limit is reached. In each iteration, the splitting ratio function updates the traffic splitting ratios to split and distribute traffic more efficiently, balancing multiple objectives such as MLU and QoS (line 7). The splitting ratio plays a crucial role in the iterative process of the H-STE method. It is used to modify the traffic splitting ratios at the SDN nodes at each iteration, ensuring effective load balancing and traffic distribution over multiple paths. The splitting ratio algorithm is designed to consider multiple objectives and uses the adaptive approach to fine-tune the traffic flows. This approach ensures that traffic is routed to minimize MLU and meet service-level requirements for critical applications. If an improvement is found, the best solution will be updated (lines 8–10). To avoid local minima, the solution is periodically modified by a hybrid global search mechanism (lines 11–13). An adaptive step size adjustment ensures that the algorithm can search the solution space more aggressively when improvements are slow and more thoroughly when close to the optimal solution (lines 14–16). The iteration count is dynamically adjusted based on the convergence rate (lines 17–20). Finally, a distributed stopping condition (lines 21–22) terminates the algorithm early if improvements stagnate across the network. The distributed approach ensures scalability for large networks. To minimize the number of distinct flow entries, we applied group flow with similar paths and destination patterns (lines 23-24). The algorithm returns the best-found solution Ubest, representing the configuration with the lowest MLU (line 25). In this algorithm, we have used the distributed Floyd–Warshall algorithm to find the shortest possible path between each combination of source and destination node faster, and the time complexity of this algorithm is O(n3), as n is the number of nodes in the topology.

We present a thorough explanation of the splitting ratio in algorithm 2. The algorithm optimizes the traffic splitting ratio of SDN nodes in a hybrid SDN environment using the Bellman–Ford shortest path computation and adaptive flow distribution. It adjusts traffic flows over multiple paths while considering real-time network conditions, such as link utilization and dynamic weight updates. The main objective of this algorithm is to optimize the traffic splitting ratio of SDN nodes when OSPF weight is given, and the time complexity of the splitting ratio function is O(n2), as n is the number of nodes in the topology. Traffic demand in a hybrid SDN can be acquired through network monitoring approaches such as sFlow [23]. The algorithm operates in several phases. Lines 1–2 compute the shortest paths for each node using the Bellman–Ford algorithm, which is particularly useful in networks where edge weights may change dynamically. The weights are updated adaptively during the



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

**ALGORITHM 2** | Splitting ratio with Bellman–Ford

```
       Input: local optimal OSPF weight
setting ω_current, V_s, C(e), D, V, E
         Output: U
 1      foreach v ∈ V do in parallel:
 2              DAG[v]=Bellman_Ford_
Shortest_Path (ωcurrent, v, E, dynamic_
weight_adjustment);
 3      foreach i ∈ Vs, eij in outgoing
links of i do
4              if link eij meets heuristic
criteria then
               Add link eij to DAG where
i≠j and i, j are incomparable;
 5      foreach iteration in max_refine-
ment_iterations do
 6      foreach v in Topological_Sort(V)
do
 7              Route_Flow_Adaptively (G,
v, D, utilization_data);
 8      expr=Multi_Commodity_With_
Adjusted_Constraints (E, V, utilization_
data);
 9      U_candidate=cplex_solve(expr);
 10      if|U_candidate - U_previous|<tol-
erance then
 11             break;
 12      U_previous=U_candidate;
 13      U=U_previous;
 14      return U;
```

computation based on the current network condition, such as traffic demand and link utilization. The shortest paths are stored in a directed acyclic graph (DAG). Initially, we form DAG by considering each node as a destination node. In this algorithm, DAG refers to the matrix containing all the feasible routes from a node to a particular node v. In creating a DAG, we begin by picking a node and identifying all the shortest paths to the node using the Bellman-Ford Shortest Path algorithm, ensuring no loops in the shortest routes tree. We add all the other edges representing the SDN node outbound links into the DAG; after constructing the DAG, we can easily direct the traffic in the route flow function. The DAG is refined based on heuristic criteria (lines 3–4). For example, outgoing links with low utilization or high capacity are prioritized when constructing the DAG.

Lines 5–10 represent iterative flow distribution with adaptive adjustment: In this phase, the algorithm iteratively adjusts traffic flows across the network, routing traffic adaptively based on real-time link utilization data. We select a node based on the topological sorting sequence, such that the flows traveling through the current node are computed before being allocated to the next hop; the flows are then routed to all outbound links in the DAG (5–6). To split the flow, we may utilize the hash functions to map the flows to distinct next hops; if the selected node is a legacy node, the flows are uniformly distributed throughout the shortest pathways and directed toward the following hops. This guarantees that all the flows from the selected node to the node *v* are correctly routed (line 7). Each iteration solves a multi-commodity flow problem, adjusting constraints to better balance the load (line 8). We repeat the process this way and route the flow of each node in a topological sort order to obtain the linear constraints. And finally, we employ the linear programming solver CPLEX [24] to address the optimization problem (line 9). The algorithm stops iterating when improvement becomes negligible, ensuring computation efficiency. Line 14 presents the final output; once convergence is achieved, the algorithm outputs the final MLU, representing the most optimized traffic distribution during the iterative process.

### 5.1 | Analysis of the Algorithm

The splitting ratio with the Bellman–Ford algorithm provides a scalable and adaptive method for traffic engineering in a hybrid SDN environment. Three main factors shape the performance of this algorithm; the first one is the parallelized computation of shortest paths using the Bellman–Ford algorithm, which ensures efficiency and can handle large-scale networks. The Bellman–Ford algorithm performs very well in the network scenario with dynamically changing traffic demands, so it is the perfect option for handling dynamic changes in link weights. The second one is the heuristic-based DAG construction, which helps the algorithm focus on high capacity and minimum link utilization, ensuring the traffic is directed via the most effective paths, which improves the overall network load balance and minimizes network congestion. The last one is the multi-commodity flow optimization and iterative flow distribution, which allows the algorithm to improve the traffic distribution incrementally, and using an optimization solver (CPLEX) guarantees that the algorithm finds a worldwide optimized solution for the network. Solver CPLEX uses the dual simplex approach to solve linear programming problems, and its time complexity is in polynomial time. Considering the above algorithms, the overall time complexity of our algorithm is O(n3).

## 6 | Experiments and Analysis

In this part, we demonstrate the performance of our proposed method on different real network datasets and topologies used in our experimental analysis; we execute the algorithms mentioned in Section 5 and analyze their performance on different real network topologies. First, we will examine the environment setting, simulation topologies, and traffic datasets. Following that, we present the effectiveness of the proposed method H-STE on minimizing MLU and flow entries under different deployment ratios of SDN nodes.

### 6.1 | Environment Setting

All the simulation experiments are simulated on a high-performance computer with a 3.20GHz Intel Core i7 with 16 GB of RAM; we simulate the network environment using Mininet Simulator [25]; for configuring legacy routers, we have used free range routing (FRR) using VTY Shel to configure network protocols. The simulation experiments are coded with the C++ program to model the traffic routing procedure; we employ the CPLEX program in our C++ program



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

to determine the splitting ratio of SDN nodes. To connect different networks in the conventional network and to integrate the legacy network with SDN Autonomous Systems (AS), we have configured BGP (Border Gateway Protocols) inside all the legacy routers, as it is necessary to have a peering system that makes the BGP protocol work. The legacy and SDN networks are integrated through the SDN-IP; this program operates on top of the ONOS controller. The SDN-IP application allows SDN nodes to communicate with legacy routers using BGP protocols.

Selecting the controller in the hybrid SDN is another issue that needs to be considered; in this simulation experiment, the ONOS [26] controller is chosen to be the SDN controller, and the integration between the legacy network and the SDN network is carried out through the SDN-IP. The SDN-IP application allows SDN nodes to communicate with legacy routers using BGP protocols.

To demonstrate the capability of our proposed approach for routing optimization, MLU, and flow entry minimization with minimum SDN deployment ratios, we have evaluated our experiment on different real network topologies (Abilene, Cernet, Nobel-Germany, and GEANT) and compared it to the following methodologies.

OSPF [27]: The OSPF technique is a routing protocol primarily used in conventional legacy networks; it supports routing the flows based on the shortest path from the source to the destination to get load balanced. The shortest pathways are determined based on connection weights. To compare it with the heuristic algorithm H-STE, we set all the parameters the same as [27]; if there are multiple shortest paths, the flows are eventually split into the shortest paths using ECMP.

OSRO [8]: To thoroughly investigate and contrast the efficiency of SDN node optimization, we compare with the only splitting ratio of SDN nodes for TE performance; we compare and conduct experiments with static OSPF and optimization of only the splitting ratio of SDN nodes.

RMC [28]: Rule multiplexing scheme studied rule placement problem to minimize rule space occupation for multiple unicast sessions under QoS constraints. This method saves TCAM resources significantly and guarantees high QoS satisfaction. To efficiently use TCAM resources, the RMC method attempts to multiplex a set of routing rules deployed on nodes to reduce the use of flow entries.

HPRS [15] formulated the routing optimization problem in a hybrid SDN with path cardinality constraints and proposed H-permissible path routing scheme, which effectively routes traffic flows under path cardinality constraints.

OORO [16] formulated the traffic engineering problem in hybrid SDN over multiple traffic matrices by mixing online splitting ratio optimization with offline weight setting optimization.

H-STE: We conducted experiments on different real network topologies to optimize the splitting ratio of SDN nodes and OSPF weight setting. All of the initial weight settings of the network are set to 1. The CDF curves of the MLU are shown in Figure 4. As shown in Figure 4, compared to OSPF and OSRO under different real network topologies, our method achieves minimum MLU; with a deployment ratio of 30% SDN nodes, we can get close to the optimal result and MLU in the network.

## 6.2 | Dataset and Topologies

The dataset and network topologies used in this experiment are shown in Table 2; the network topologies datasets are available on SDNlib [29]. The Abilene topology, part of the American Research and Education Network, comprises 12 nodes and 30 links. The Cernet topology, belonging to the China education and research network, consists of 14 nodes and 32 links. The Noble-Germany network topology consists of 17 nodes and 26 links. Lastly, the GEANT topology, part of the European research and education network, consists of 23 nodes and 74 links. The traffic datasets required in the experiment analysis are real network datasets measured or created in the specific topologies every 5 or 15 min.

For Abilene topology, the total traffic provided by the TOTEM project [30] was measured from the first of March 2004 to September 2004; every 24 h, snapshots were created at 5-min intervals. For the Cernet topology, the datasets are provided by Zhang et al. [31]. The traffic matrices were measured every 5 min from 19 Feb 2013 to 26 March 2013. For the GEANT topology datasets provided by Uhling [32], the traffic matrices were measured every 15 min from May 4, 2005, to August 31, 2005.

## 6.3 | Evaluation of Network Performance

### 6.3.1 | MLU

In a hybrid SDN network, the SDN deployment ratio is defined as the ratio of SDN nodes throughout the network. Minimizing the network's MLU is a critical parameter in the hybrid SDN when considering network load balance. The MLU of each link can be calculated as the maximum ratio of the link load divided by the link capacity. The primary objective is minimizing the MLU with minimum SDN nodes while maintaining the network's load balance to enhance performance. Maintaining network load balance in a hybrid SDN refers to the equitable distribution of traffic across all network paths

**TABLE 2** | Network topologies information.

| Network | Nodes | Links | Time | Time-range |
|---|---|---|---|---|
| Abilene | 12 | 30 | 5 min | 6 months |
| Cernet | 14 | 32 | 5 min | 1 month |
| Nobel-Germany | 17 | 26 | 5 min | 1 day |
| Geant | 23 | 74 | 15 min | 4 months |



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

to ensure efficient utilization of network resources and minimize congestion on any single link. This balance helps to prevent the situation in which some links exceed their capacity (overloaded), leading to congestion and inefficient traffic handling, while other links remain unutilized or underutilized. By balancing the load, the network can provide consistent performance, minimize latency, and improve reliability. The H-STE method helps bridge the gap between both networks by optimizing the overall network performance while adhering to the limitations imposed by traditional routing protocols. This method adjusts the splitting ratio of SDN nodes to reduce network congestion while having a multi-objective focus, primarily minimizing MLU and considering other network performance metrics. Various methods and metrics can evaluate the load balance in the hybrid SDN environment, such as load variance, link utilization distribution, MLU, failover capacity, and path diversity. The most crucial metric for evaluating load balance in the hybrid SDN is MLU; in this study, the network load balance is evaluated by considering the MLU. It represents the utilization percentage of the link with overloaded or underutilized links in the network. A high MLU indicates that particular links are approaching their maximum capacity, which causes network congestion and inefficient traffic handling. Minimizing MLU is critical for load balancing because it helps prevent overload on any single link; with minimum MLU, traffic load can be distributed more efficiently across the network.

Our main objective is to minimize the MLU with the minimum number of SDN nodes while maintaining the network's load balance. To minimize the MLU, we increase the deployment ratios; when we deploy 0.3, we can see improvement of MLU as the MLU of different topologies begins to stabilize. We find that the value of MLU varies and directly depends on the increase in SDN deployment ratios; we plot the curves to show that the value of MLU becomes smaller and decreases with the increases in the SDN nodes deployment ratio under different network topologies. Figure 3 depicts the effect of the SDN nodes deployment ratio on improving the MLU under different network topologies. The result shows that the improvement of MLU directly depends on the increase of SDN deployment ratios; we can see a rapid decrease in the value of MLU with a shift from 0 to 0.3 in the SDN deployment ratios. As soon as the deployment ratio surpasses 0.3, the MLU begins to stabilize, and the curve of H-STE becomes flat. The findings indicate that increasing the deployment ratios of SDN can obtain near-optimal outcomes in optimizing MLU comprehensively. Increasing the number of SDN nodes can boost the ability to split and direct the traffic flows in hybrid SDN. The main goal is to minimize the MLU to prevent network congestion and evenly distribute the network load to enhance the overall network performance. In Figure 3, we can see that the OSPF curve remains steady and flat with decreasing or increasing SDN deployment ratios; the reason is that OSPF routes traffic based on the shortest path from the source to the destination. Figure 3e shows the impact of varying SDN deployment ratios on improving MLU. The optimal splitting ratios are obtained using the CPLEX Solver based on the optimization model. The finding indicates that as the deployment ratio increases, MLU improves correspondingly. MLU stabilizes as the deployment ratio reaches approximately 30% for different network topologies.

We consider and set the SDN deployment ratios to 0.3 based on the experimental analysis of different network topologies. As the MLU stabilizes, we can see that the curves are getting near-optimal results. To show our proposed method's priority and efficiency for routing optimization while enhancing TE performance, we compared our suggested method with different strategies under different real network topologies, as shown in Table 2. To compare it with the only optimization of the OSPF weight setting [27], we set all the parameters the same as [27] and compared it with the only splitting ratio of SDN nodes optimization with a fixed link weight setting [8]. In this study, we conducted a series of experiments to evaluate the performance of different approaches. The results of these experiments were systematically compared, and the mean MLU outcomes were analyzed. To present the findings, the average MLU results were plotted in CDF graphs displayed in Figure 4, which illustrates the performance under different network topologies. As shown in Figure 4, when the deployment ratio increases, the value of MLU becomes smaller. Compared to other methods, the H-STE method is close to the optimal result, as this method optimizes both the SDN node splitting ratio and OSPF weight setting. In contrast, the OSPF method only optimizes the OSPF weight setting, and the OSRO method optimizes only the splitting ratio of SDN nodes with fixed length weight.

In hybrid SDN, reducing the MLU and optimizing routing present significant challenges. To minimize the MLU, we increase the deployment ratios, and we can see that with a 30% deployment ratio of SDN, the MLU of different topologies begins to stabilize. The findings indicate that increasing the deployment ratios of SDN can obtain near-optimal outcomes in optimizing MLU comprehensively. Minimizing MLU is directly dependent on the increasing deployment ratio of SDN nodes. More SDN nodes can bring more flexibility to split and forward the traffic flows in hybrid SDN, as our common goal is to minimize the MLU with minimum SDN nodes while balancing the network load to enhance the overall network performance. As a result, compared to other optimization methods, the H-STE method obtains a much lower MLU under different network topologies; with a minimum deployment ratio of SDN nodes, we can reduce the MLU and get close to the optimal result to get maximum benefits from the concept of a hybrid SDN network.

A 30% deployment of SDN nodes in a real-world hybrid SDN network means that only a portion of the network is managed through SDN principles. At the same time, the rest operate using conventional distributed network protocols. The finding is insightful and suggests a balance point where SDN control is applied to control traffic without heavily depending on SDN resources. It brings several practical implications across different network dimensions, such as scalability, cost-effectiveness, operational complexity, and network performance; it can significantly reduce hardware replacement costs. Depending on the organization's requirements, they can incrementally migrate to a hybrid SDN and then to a fully SDN network.



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

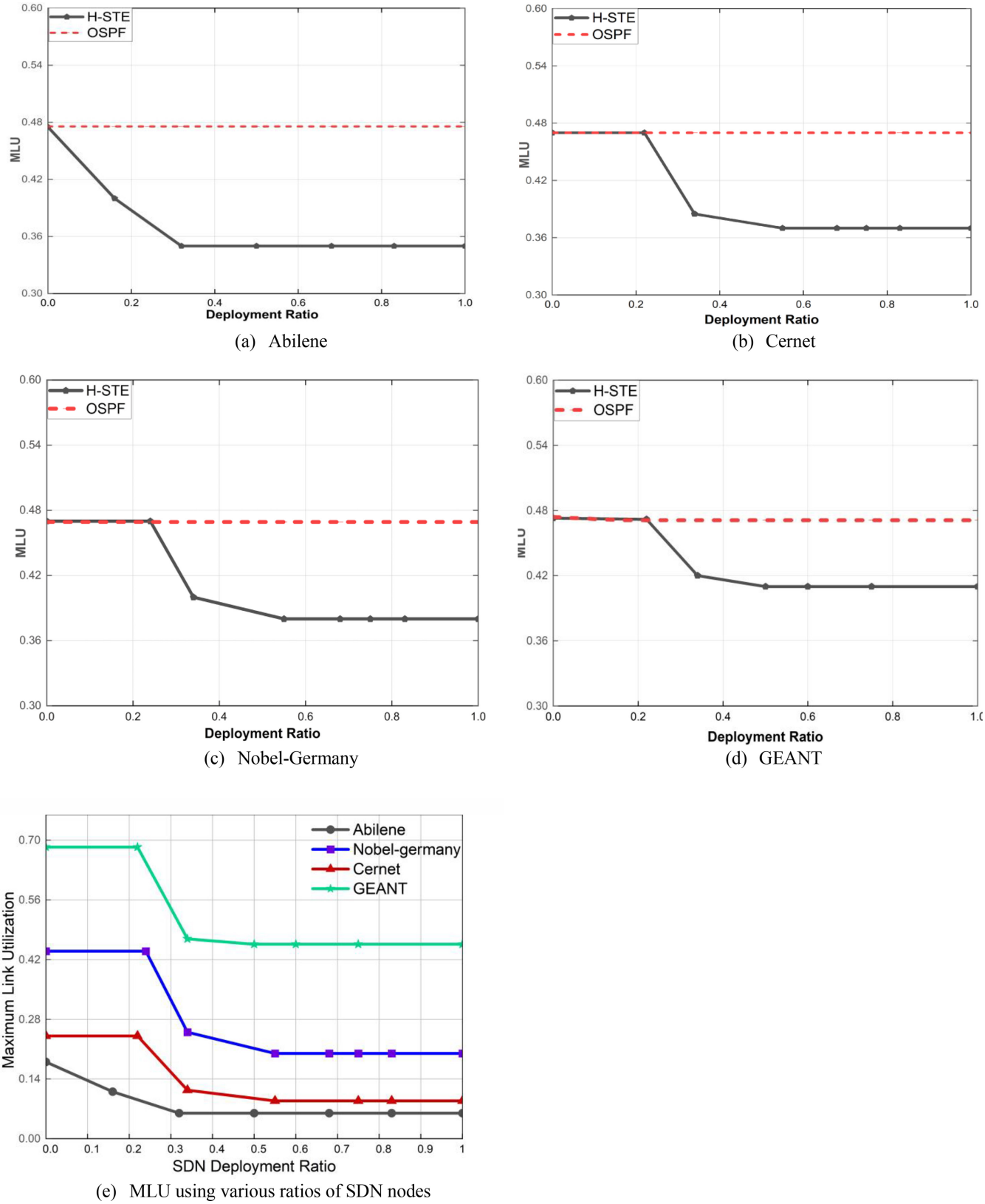



**FIGURE 3** | MLU with various SDN deployment ratios under real network topologies.

However, the result may vary under different network conditions and configurations. The deployment of 30% of SDN nodes offers a scalable and cost-effective migration toward full SDN; its effectiveness depends on how the SDN nodes are deployed strategically. They can offer significant advantages regarding flexibility and TE when deployed in high-impact areas.



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

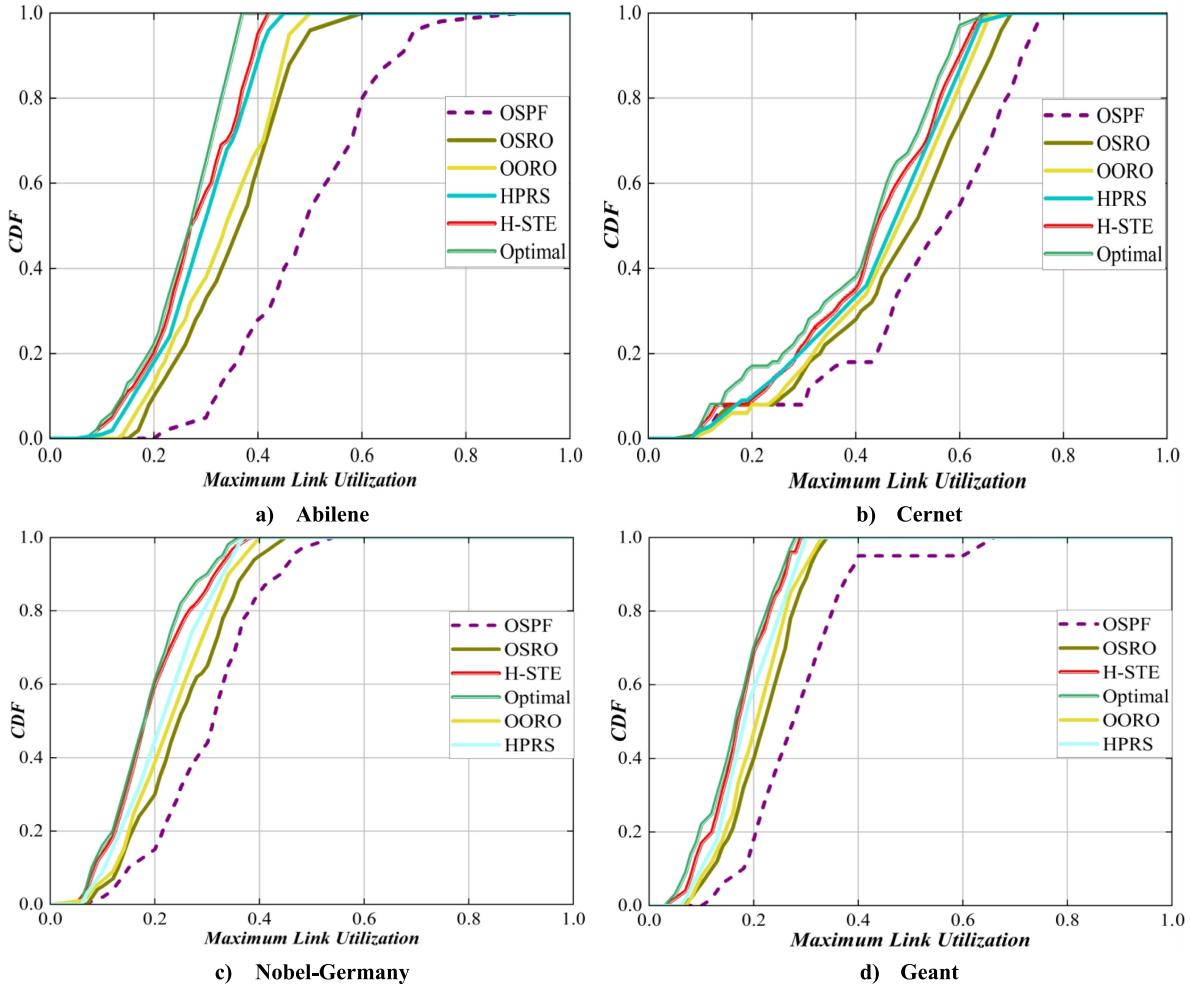



**FIGURE 4** | MLU CDF curve under real network topologies.

However, poor placement could lead to more complex with few performance advantages.

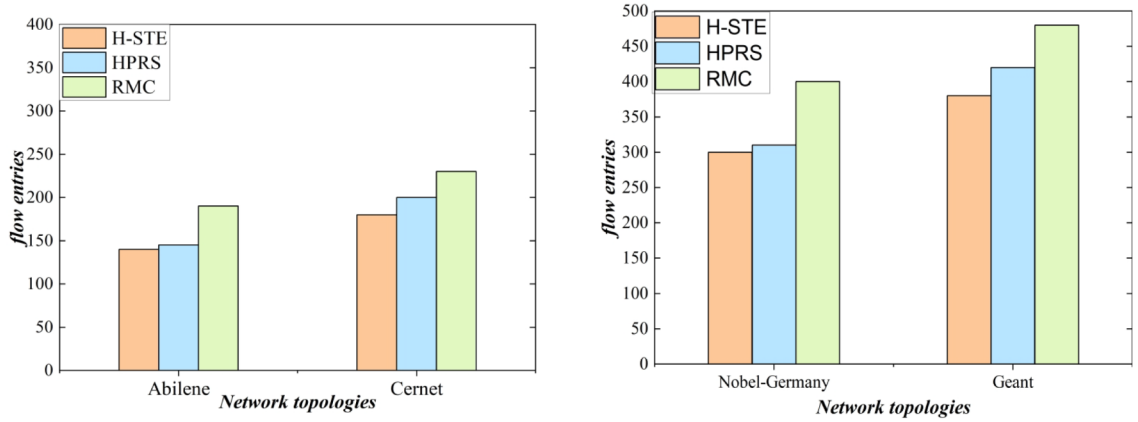
### 6.3.2 | Flow Entries

In the hybrid SDN environment, minimizing flow entries is crucial to maximize resource utilization and enhance network efficiency, particularly in situations with limited bandwidth and potential performance issues due to network congestion. Our method effectively controls flow entries, enhances the data transmission process, and decreases the costs of maintaining extensive routing tables. This capability is particularly helpful for network environments that require a flexible and dynamic management approach to accommodate different application demands and fluctuating traffic volumes. With digital resources and online platforms continuing to expand, minimizing flow entries can enhance network stability and dependability. It is essential to provide low latency and seamless connectivity. We demonstrate the stability and scalability of our method in adapting to different SDN deployment ratios while consistently outperforming alternative methods.

In this study, our objective is to minimize flow entries with the minimum deployment ratios. We have compared the proposed H-STE method with previous routing optimization approaches in the hybrid SDN environment. Our method demonstrates superior performance in minimizing the number of flow entries under different real network topologies. Figure 5 illustrates the total flow entries employed in different methods under various real network topologies. As demonstrated in Figure 5, the number of flow entries generated by H-STE increases as the topological scale expands. This trend arises because larger topological scales inherently provide greater opportunities for minimizing and optimizing flow entries; the network topology size directly influences the scalability of our method. Compared to the previous routing algorithm, on average, the flow entries utilized in H-STE decrease by 10.9% and 25.6% for given network topologies. This reduction can be attributed to H-STE's efficient flow management strategy, which reduces flow entry redundancy and optimizes the routing path, enabling networks to function effectively with fewer entries while maintaining performance. Furthermore, the scalability of our approach ensures that the method maintains its efficiency as the topology expands, demonstrating its resilience in controlling diverse network topologies. The comparisons indicate that our method outperforms others in minimizing flow entries under different network topologies. This is a significant improvement as it reduces resource utilization while enhancing network efficiency. Considering these factors, our method can optimize routing decisions more effectively, ensuring the network functions optimally within its capacity limitations while meeting performance requirements.



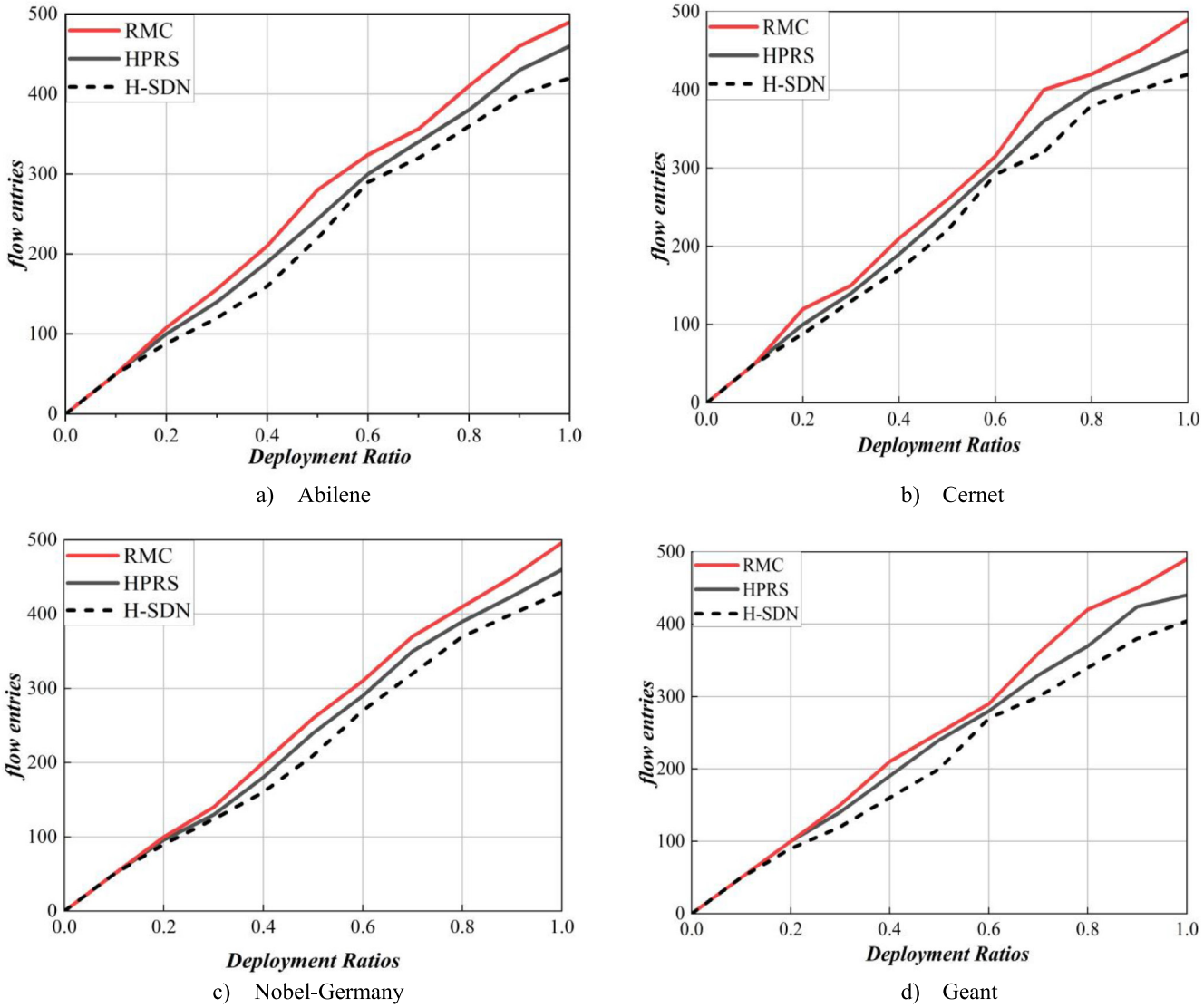
10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License




a) Abilene & Cernet topologies

b) Nobel & Geant topologies

**FIGURE 5** | Flow entries under various network topologies.




a) Abilene

b) Cernet

c) Nobel-Germany

d) Geant

**FIGURE 6** | Flow entries under various SDN deployment ratios.

The flow entries vary with different SDN deployment ratios under different network topologies. To better understand and evaluate the necessary flow entries under various deployment ratios, we generated curves demonstrating how the number of flow entries varies with different deployment ratios of SDN nodes. As illustrated in Figure 6, with changing SDN



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

deployment ratios, we can observe trends and patterns of flow entry utilization; this graphical representation can be utilized to determine the optimal deployment ratios for minimizing flow entries while maintaining efficient network performance. As the SDN deployment ratios increase, flow entries also increase for different topologies. This trend emphasizes the dynamic nature of network traffic and the associated requirement for flow entries, considering more SDN nodes in the network. For minimizing flow entries in the hybrid SDN, our method can manage flow entries more effectively than previous methods. For instance, considering the Abilene network topology and taking an average across different SDN deployment ratios, H-STE minimizes the flow entries by 6.8% compared to HPRS and 16.3% compared to the RMC method. This reduction demonstrates the efficacy of our method in optimizing resource utilization and minimizing unnecessary flow entries while balancing the network load. Similarly, in the Cernet topology, the H-STE method demonstrates a notable decrease in flow entries of 6.1% compared to HPRS and 15.3% compared to RMC, which highlights the benefits of this approach on different network topologies. The same phenomenon is observed in the Nobel-Germany topology; the H-STE approach demonstrates a significant decrease in the flow entries compared to HPRS and RMC by 6.5% and 17.1%, respectively. The ability of H-STE to modify and reduce flow entries demonstrates further potential for network scalability and efficiency. Lastly, the H-STE approach outperforms the other methods in the GEANT topology, with an average decrease in the flow entries of 10.4% compared to HPRS and a substantially larger decrease of 23.6% compared to RMC on average at various SDN deployment ratios. This significant reduction demonstrates the efficacy of H-STE in handling routing traffic in complex network environments, improving overall network performance and reducing overhead. The findings conclusively demonstrate that H-STE performs significantly better than other approaches in efficiently managing flow entries. Consequently, H-STE is a viable option for expediting SDN deployment ratios in different network topologies, ultimately enhancing resource management and network performance. The simulation results indicated that the proposed H-STE approach outperformed other approaches in modifying and decreasing the number of used flow entries across various SDN deployment ratios within different network topologies. Future research studies could explore comparing the performance impact of 30% vs. 50% vs. full SDN deployment in ISP and enterprise networks.

## 7 | Conclusion

Traffic engineering in hybrid SDN is one of the open areas for the researcher; with the combination of centralized SDN and decentralized traditional networks, it is much more challenging to optimize traffic flows, resource utilization, etc. This research contributes to the expanding discussion of TE and routing optimization in the hybrid SDN, necessitating innovative solutions to enhance network efficiency, performance, and resource utilization. This paper studied traffic engineering in a migrated hybrid SDN where SDN nodes are incrementally deployed to the conventional network. We introduced a heuristic technique, H-STE, to improve traffic engineering in the hybrid SDN; we concentrated on minimizing the maximum link utilization by optimizing two critical aspects: optimizing the OSPF weight settings across the entire network to balance the flows originating from conventional devices and optimizing the traffic splitting ratio of SDN nodes. Given the close interconnection of these devices, optimizing both aspects presents significant challenges for achieving efficient traffic management and maintaining a balanced load across the network. The H-STE method helps bridge the gap between the two by optimizing the whole network performance and adhering to the limitations imposed by legacy routing infrastructure. We have conducted several experiments on real network datasets; the results demonstrate that with a 30% deployment ratio of SDN nodes, we can achieve superior TE performance and maximum benefits from a hybrid SDN network. The study findings contribute to the progress and development of efficient, flexible, and robust communication infrastructures as we strive to enhance network management and design. In the Future, we will try to develop an AI-driven optimization model to determine which 30% of nodes should be SDN-enabled to gain maximum performance with minimal cost and try to compare the performance impact of 30% vs. 50% vs. full SDN deployment in ISP and enterprise networks.

### Data Availability Statement

Data are not shared.

### References


1. U. Cisco, 2022.Cisco Annual Internet Report (2018–2023) White Paper.

2. S. Mehraban and R. K. Yadav, ""Traffic Engineering and Quality of Service in Hybrid Software Defined Networks." China," *Communications* 21, no. 2 (2024): 96–121.

3. C.-Y. Hong, S. Kandula, R. Mahajan, et al., "Achieving high utilization with software-driven WAN," Proceedings of the ACM SIGCOMM, 2013 Conference on SIGCOMM. 2013.

4. S. Jain, A. Kumar, S. Mandal, et al., ""B4: Experience With a Globally-Deployed Software Defined WAN." ACM SIGCOMM," *Computer Communication Review* 43, no. 4 (2013): 3–14.

5. R. Amin, M. Reisslein, and N. Shah, "Hybrid SDN Networks: A Survey of Existing Approaches," *IEEE Communications Surveys & Tutorials* 20, no. 4 (2018): 3259–3306.

6. Y. Sinha and K. Haribabu, "A Survey: Hybrid SDN," *Journal of Network and Computer Applications* 100 (2017): 35–55.

7. S. Mehraban and R. K. Yadav, "Quality of Services in Hybrid SDN (HSDN): A Review," in *In 2022 7th International Conference on Communication and Electronics Systems (ICCES)* (IEEE, 2022), 652–658.

8. S. Agarwal, M. Kodialam, and T. V. Lakshman, "Traffic Engineering in Software Defined Networks," in *2013 Proceedings IEEE INFOCOM* (IEEE, 2013).

9. X. Jia, Y. Jiang, and Z. Guo, "Incremental Switch Deployment for Hybrid Software-Defined Networks," in *2016 IEEE 41st Conference on Local Computer Networks (LCN)* (IEEE, 2016).

10. H. Xu, X.-Y. Li, L. Huang, et al., "Incremental Deployment and Throughput Maximization Routing for a Hybrid SDN," *IEEE/ACM Transactions on Networking* 25, no. 3 (2017): 1861–1875.

11. Z. Yang and K. L. Yeung, "SDN Candidate Selection in Hybrid IP/SDN Networks for Single Link Failure Protection," *IEEE/ACM Transactions on Networking* 28, no. 1 (2020): 312–321.




10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

12. B. Kar, W. Eric Hsiao-Kuang, and Y.-D. Lin, "The Budgeted Maximum Coverage Problem in Partially Deployed Software Defined Networks," *IEEE Transactions on Network and Service Management* 13, no. 3 (2016): 394–406.

13. Y. Sun, Z. Guo, S. Dou, J. Zhang, C. Li, and X. Ouyang, "Poster: Enabling Fast Forwarding in Hybrid Software-Defined Networks," in *2021 IEEE 29th International Conference on Network Protocols (ICNP)* (IEEE, 2021).

14. Y. Guo, Z. Wang, Z. Liu, et al., "SOTE: Traffic Engineering in Hybrid Software Defined Networks," *Computer Networks* 154 (2019): 60–72.

15. Y. Guo, H. Luo, Z. Wang, X. Yin, and J. Wu, "Routing Optimization With Path Cardinality Constraints in a Hybrid SDN," *Computer Communications* 165 (2021): 112–121.

16. Y. Guo, Z. Wang, X. Yin, X. Shi, and J. Wu, "Traffic Engineering in Hybrid SDN Networks With Multiple Traffic Matrices," *Computer Networks* 126 (2017): 187–199.

17. Z. Xu, F. Y. Yan, R. Singh, J. T. Chiu, A. M. Rush, and M. Yu, "Teal: Learning-Accelerated Optimization of WAN Traffic Engineering," Proceedings of the ACM SIGCOMM 2023 Conference. 2023.

18. Y. Guo, W. Wang, H. Zhang, et al., "Traffic Engineering in Hybrid Software-Defined Network via Reinforcement Learning," *Journal of Network and Computer Applications* 189 (2021): 103116.

19. C. Liu, P. Wu, M. Xu, Y. Yang, and N. Geng, "Scalable Deep Reinforcement Learning-Based Online Routing for Multi-Type Service Requirements," *IEEE Transactions on Parallel and Distributed Systems* 34, no. 8 (2023): 2337–2351.

20. Y. Guo, M. Ding, W. Zhou, B. Lin, C. Chen, and H. Luo, "MATE: A Multi-Agent Reinforcement Learning Approach for Traffic Engineering in Hybrid Software Defined Networks," *Journal of Network and Computer Applications* 231 (2024): 103981.

21. B. Lin, Y. Guo, H. Luo, and M. Ding, "TITE: A Transformer-Based Deep Reinforcement Learning Approach for Traffic Engineering in Hybrid SDN With Dynamic Traffic," *Future Generation Computer Systems* 161 (2024): 95–105.

22. M. R. Garey and D. S. Johnson, *Computers and Intractability*, Vol. 29 (W. H. Freeman, 2002).

23. P. Phaal, S. Panchen, and N. McKee, *InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks* (Internet Eng. Task Force, 2001).

24. I. B. M. I. L. O. G. Cplex, "V12. 1: User's Manual for CPLEX," *International Business Machines Corporation* 46, no. 53 (2009): 157.

25. B. Lantz, B. Heller, and N. McKeown, "A Network in a Laptop: Rapid Prototyping for Software-Defined Networks," Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks. 2010.

26. https://opennetworking.org/onos/ accessed 13/04/2023.

27. B. Fortz and M. Thorup, "Internet Traffic Engineering by Optimizing OSPF Weights," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No. 00CH37064)*, vol. 2 (IEEE, 2000).

28. H. Huang, P. Li, S. Guo, and B. Ye, "The Joint Optimization of Rules Allocation and Traffic Engineering in Software Defined Network," in *2014 IEEE 22nd International Symposium of Quality of Service (IWQoS)* (IEEE, 2014).

29. S. Orlowski, R. Wessäly, M. Pióro, and A. Tomaszewski, "Sndlib 1.0—Survivable Network Design Library, Networks: An," *International Journal* 55, no. 3 (2010): 276–286.

30. S. Balon, and G. Monfort, "The Traffic Matrices and Topology of the Abilene Network," (http://totem.run.montefiore.ulg.ac.be/datatools.html).

31. B. Zhang, J. Bi, J. Wu, and F. Baker, ""Cte: Cost-Effective Intra-Domain Traffic Engineering." ACM SIGCOMM," *Computer Communication Review* 44, no. 4 (2014): 115–116.

32. S. Uhlig, B. Quoitin, J. Lepropre, and S. Balon, ""Providing Public Intradomain Traffic Matrices to the Research Community." ACM SIGCOMM," *Computer Communication Review* 36, no. 1 (2006): 83–86.



10991190, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nem.70017 by Delhi Technological University, Wiley Online Library on [06/05/2025]. 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

2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3315-2169-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/IATMSI64286.2025.10985119

# Twitter Sentiment Analysis of the Electric Vehicle Industry in India using NLP

Rishabh Sharma
*Electrical Engineering Department*
*Delhi Technological university*
Delhi, India
rsharma5752@gmail.com

Meha Joshi
*Department of Management*
*Indira Gandhi Delhi Technical University for Women*
Delhi, India
mehajoshi83@gmail.com

***Abstract*—This paper aims to examine public opinion regarding the introduction of electric vehicles in the Indian auto industry through the use of natural language processing and Nvivo software, which analyses tweets about electric vehicles from the past few years. To provide a comprehensive view of the electric vehicle market in India, the public perception of the leading EV firms has also been examined. For sentiment analysis, we have used Python's Tweepy library to examine the sentiment of a particular tweet.**

***Index Terms*—Electric Vehicle, EV India, EV Industry India, Sentiment Analysis, Natural Language Processing**

## I. Introduction

One of India's most promising and potentially lucrative industries is the electric car business. In addition to heavy government promotion, there is a growing popular awareness of electric cars among the general Indian population. One of the major benefits of using electric vehicles since its inception is cost effectiveness, and being environment friendly. But In Ind,ia's context the technology a has reached little late, and Industry is still catching up to make the EV Industry economically viable for the consumers [1]. The general attitude of the Indian public towards electric vehicles is one of the most important factors that all members of the decision-making body should consider since it is ultimately up to the Indian people to decide how to react to this significant development in their nation's automotive sector. By displacing gasoline and diesel fuels, electric cars and fleets offer significant public health benefits by reducing emissions from the transportation sector. However, public confidence in the reliability of charging infrastructure remains a fundamental barrier to adoption [2]. In the last couple of years, the increase of EVs in India has been for many reasons; this is because of what the government has been continuously doing to encourage people to adopt what is environmentally friendly. Most individuabelieveold the view that by opting to use electric vehicles (EVs), they can decrease the amount of carbon emissions produced, hence playing their part in helping the nation combat climate change.

## II. Current Industry

In the recent times, the Indian Government has rolled out wide policies and incentives to enhance the penetration of electric vehicles. These measures comprise the most important components for the adoption of electric mobility through the supply and demand side: subsidies for EV purchase, tax incentives, and the setting up of charging infrastructure. In India, EVs are viewed very positively because they are considered a very clean, efficient, and safe mode of transportation. After the execution of these new laws and initiatives by the government, there has been a huge snumbern the amount of electric vehicles (EVs) on Indian roads in the past few years in India saw a notable upsurge. In FY 2020-21, a total of 48,179 electric cars were sold. The number then jumped to 2,37,811 in FY 2021-22 and even further rose to 4,42,901 in FY 2022-23. The Indian Government has targeted to sell about 30% of electric automobiles by 2030, and the trend is likely to continue further. The availability of EV models in different price brackets has played a major role in making EVs acceptable and adoptable in India. Tata Motors and Mahindra, along with other major manufacturers, offer a wide variety of EV models in the Indian market. Several new entrants including MG Motors, Nexa, and Hyundai apart from the well-established players have announced their upcoming electric vehicles (EVs) for the Indian market. These steps forward notwithstanding, the EV industry in India still has some issues to deal with before it can realize its full potential. The lack of charging infrastructure or stations is one critical barrier. Though the government has made extensive efforts to address this matter, access to charge points remains limited, especially in many remote areas. Many potential buyers face a substantial economic hurdle as electric cars (EVs) are more expensive than their gas-powered counterparts.

The Indian market for electric vehicles (EV) offers a variety of models from Tata Motors, Mahindra and Hero Electric among other well-known manufacturers in India. These companies have had a long history in Indian auto sector and strong manufacturing and supply chain capabilities similar to those of other car makers. That has seen them grow their production lines for EVs rapidly to match growing demand. Also, few new entrants in the coming years have announced their plans to bring electric vehicles (Ein Vs) for sale India. The Indian market has seen many private companies entering the electric vehicle (EV) charging space. These include organisations that manage charging networks as well as power utility providers. Furthermore, apart from imposing fees on network operators and electricity utilities, there are also several start-ups that have seen rapid growth and continue to evolve at an astounding rate. In India, the construction of EV charging infrastructure is now in its initial phase, but there has been notable advancement in recent years. Public perception towards EVs and general willingness to use EV can play an important role in promoting these vehicles. Not only should the technical limitations of

979-8-3315-2169-1/25/$31.00 ©2025 IEEE
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:01 UTC from IEEE Xplore. Restrictions apply.

EVs, including battery capacity and weight, be improved, people's personal and social issues should be investigated to increase the commercial distributions [3]. Tata is one of the companies which is actively working in this area. Charging network operators and power utilities, which are private firms, are crucial in the development of charging infrastructure and are facilitating the widespread availability of charging stations around the nation. As with many new technologies, the entry of BEVs into the transportation sector requires a shift in consumer mindset.

## III. Methodology

The main goal of this research paper is to analyze the sentiment toward electric vehicles in the Indian customer segment using two approaches:-

### A. Using a Natural Language Processing Algorithm

By applying a Natural language processing algorithm Analyzing, the tweets tweeted by individuals It has become easier to understand how people believe in an issue. Sentiment Analysis using NLP (Natural Language Processing) - this is a much-loved application for the same. To collect tweets in real-time an API provided by Twitter, Tweepy was used.

*1) Data Collection:* To perform the analysis of Twitter data related to the electric vehicle (EV) industry in India we need tweets associated with the EV sector from the Twitter platform. To search tweets based on particular keywords and hashtags, we employ the Twitter API. For example, we might search for tweets with keywords such as "electric vehicle", "EV India", "EV", etc.

*2) Data Preprocessing:* Sentiment Analysis in twitter is quite difficult due to its short length. Presence of emoticons, slang words and misspellings in tweets forced to have a preprocessing step before feature extraction [4]. That is, data scrubbing will be removing unnecessary information like URLs or special characters. Furthermore, we tokenized the data by breaking down each text into different words and phrases and removed stop-words such as "and", "the" or "is" which provide little importance in text. Also, we continued by following the stemming or lemmatization of words to reduce them to their root form and avoid repeating unnecessary word times.

*3) Sentiment Analysis:* After preprocessing the data, we utilised NLP techniques, specifically sentiment analysis, to categorise tweets into positive, negative, or neutral sentiment categories. Sentiment analysis entails the utilisation of algorithms and machine learning models to discern the tone and emotion expressed in a document. For the EV industry in India, we employed sentiment analysis to ascertain public attitude towards electric vehicles, charging infrastructure, government legislation, and other relevant subjects. As with many emerging technologies, the introduction of BEVs in the transport fleet requires a mentality change of the consumers. This is due to the fact that BEVs have certain characteristics that differ from conventional vehicles [5].

TABLE I
Sentiment Analysis Results

| S. No | Keyword | No. of tweets analyzed | Positive sentiment (%) | Negative sentiment (%) | Neutral sentiment (%) |
|---|---|---|---|---|---|
| 1 | Electric vehicle India | 1000 | 32.05 | 5.77 | 62.18 |
| 2 | EV India | 1.2k | 47.72 | 8.50 | 43.78 |
| 3 | EV industry India | 110 | 58.97 | 11.54 | 29.49 |
| 4 | TATA EV | 1000 | 41.01 | 5.82 | 53.17 |
| 5 | OLA electric | 304 | 34.21 | 22.04 | 43.75 |
| 6 | TATA Nexon | 908 | 28.66 | 13.65 | 57.49 |
| 7 | TATA Tigor EV | 100 | 29.59 | 18.37 | 52.04 |
| 8 | Tesla India | 661 | 41.15 | 12.56 | 46.29 |
| 9 | TATA Tiago | 180 | 23.46 | 16.20 | 60.34 |
| 10 | Ola electric | 454 | 33.48 | 14.98 | 51.54 |
| 11 | Ather energy | 187 | 74.33 | 5.35 | 20.32 |
| 12 | Statiq | 982 | 52.12 | 17.58 | 30.3 |

*4) Analysis workflow:* In order to conduct the analysis, we provided several keywords as input to ensure comprehensive coverage of all keywords pertaining to the electric car market in India. Some examples of these keywords are: 'EV industry India', 'electric vehicles India', 'EV India', and so on. By employing these specific keywords, we retrieved approximately 7,000 tweets. In addition, we used keywords related to prominent electric vehicle (EV) manufacturers like TATA, Mahindra, Ola, etc., to obtain a comprehensive understanding of the entire sector, including both the general public's perspective and the viewpoints of EV manufacturing businesses.

The data clearly indicates that there is currently a significant lack of social media participation in the electric vehicle business in India. There are very less tweets done about the electric vehicle industry in India which shows the scarce penetration of EVs in the Indian automobile market. A further observation that can be inferred from the aforementioned readings is that the prevailing discourse on Twitter on the electric vehicle (EV) market in India exhibits a predominantly neutral attitude and lacks significant polarisation. However, the tweets exhibit a significant degree of polarization when it comes to Ather Energy and Statiq. The incosistency in tweets could be on of the reasons for this polarization. Sentiment analysis over Twitter data and other similar microblogs faces several new challenges due to the typical short length and irregular structure of such content. [6]

### B. Using Analysis Tools

In our research, we analyzed sentiments towards the electric vehicle industry using different sentiment analytical tools like Nvivo. In this article, we concentrated on the specific demographic groups in India that are of interest. The results exhibited a high similarity when compared to the tweet

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:01 UTC from IEEE Xplore. Restrictions apply.

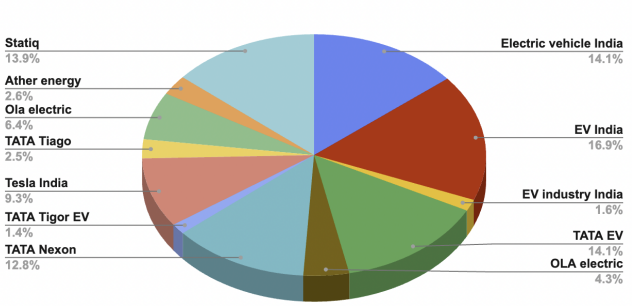



Fig. 1. Share of tweets for each keyword

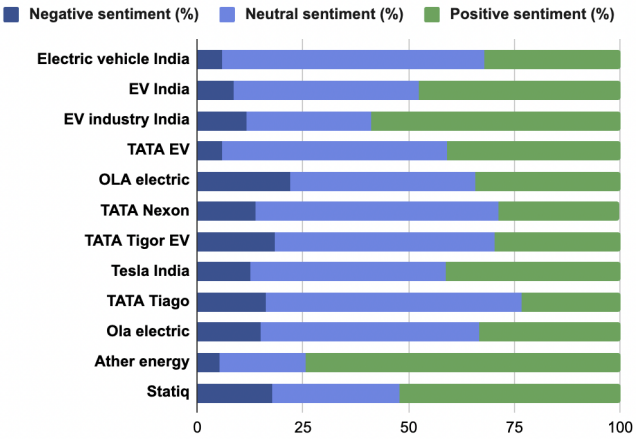



Fig. 2. Keyword wise division of sentiment of tweets

analysis conducted on the tweets utilising machine learning algorithms.

Currently, most automobiles in the EV market tend to have a significantly more negative attitude associated with them compared to the overall mood towards the industry. The recent incidents of fires in electric vehicles (EVs), along with a lack of knowledge and inadequate infrastructure, could perhaps explain the findings. The increasing negative attitude towards vehicles can be resolved by spreading accurate information without any fabrication and by promptly improving the whole charging infrastructure for electric vehicles in India.

*1) Positive Sentiment:* Twitter discussions exhibit a clear inclination towards the EV sector with a positive tone. Users frequently emphasise the environmental advantages of electric vehicles, including decreased pollutants and diminished reliance on fossil fuels. In addition, they commend the technological advancements and innovative features of electric vehicles (EVs), as well as the exceptional performance and driving experience they provide. The positive feeling is also influenced by the growing accessibility of charging infrastructure and government incentives aimed at encouraging the adoption of electric vehicles.

*2) Negative Sentiment:* Despite the positive comments concerning electric cars, there is also a wave of backlash about the electric vehicle (EV) industry among Twitter users. Some users are apprehensive about the financial implications of the electric cars, in particular the significant initial cost and the ongoing maintenance cost of electric vehicles. Some users are uncertain of the distances electric cars can travel and reliability of the electric vehicle, especially in severe weather conditions. There is also concerns of the environmental impacts of the battery production and the disposal of used batteries.

*3) Neutral Sentiment:* A significant amount of the tweets that have occurred about the electric vehicle (EV) industry with neutral sentiment, where not only we see a positive or negative sentiment. There are users who have alerted developments in the industry, sharing news but not expressing directly with a sentiment.

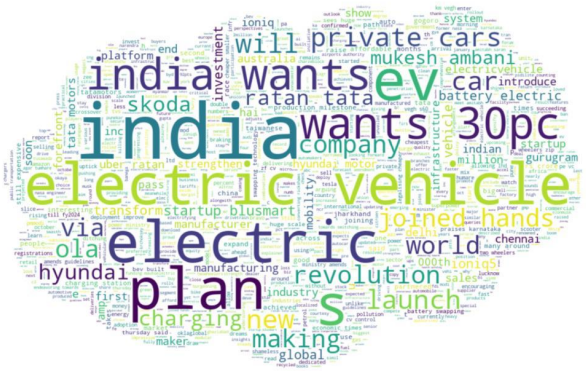



Fig. 3. Electric Vehicle India

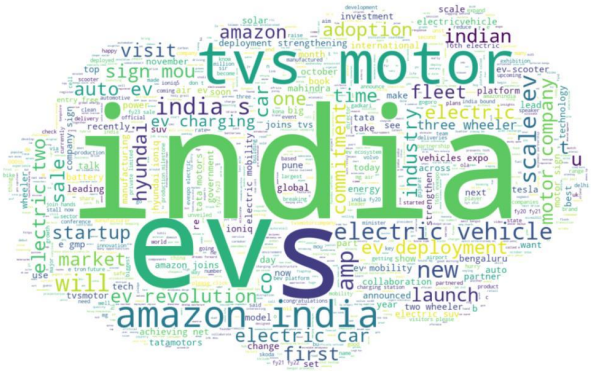



Fig. 4. EV India

## IV. Word Cloud Analysis

The preceding analysis is a word cloud generated from the keywords entered in the sentiment analysis section. The predominant terms observed in the search term 'electric vehicle industry India' primarily pertained to leading firms in the electric car sector, startups, and phrases associated with charging infrastructure. The word cloud analysis was

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:01 UTC from IEEE Xplore. Restrictions apply.

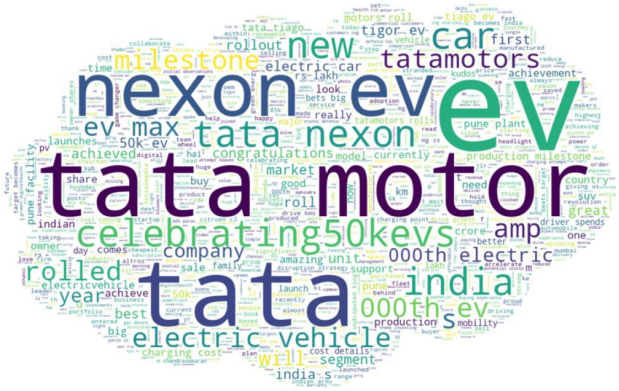



Fig. 5. Tata EV

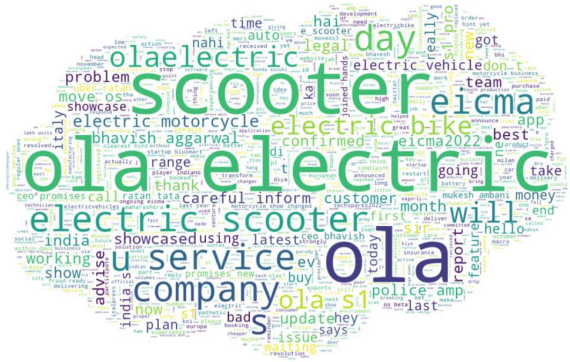



Fig. 6. Ola electric

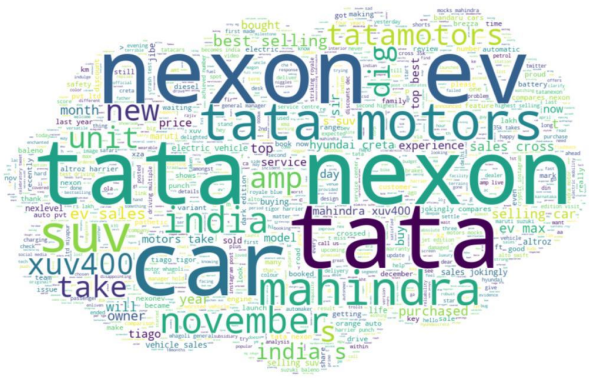



Fig. 7. Tata Nexon

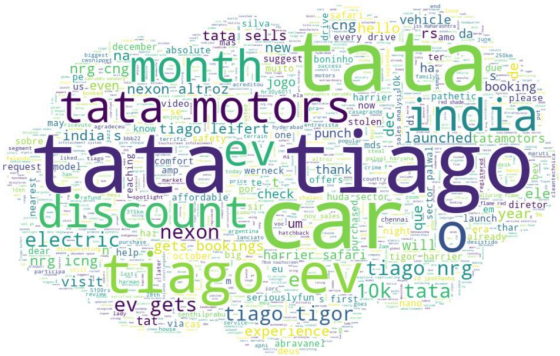



Fig. 8. Tata Tiago

conducted by segmenting the term search into distinct categories. Figures 3, 4, 5, 6, 7, and 8 show the word clouds for each keyword, highlighting the most frequently used words associated with each term.

*1) Positive Sentiment:* Twitter discussions reflect a mostly positive sentiment towards the electric vehicle (EV) industry. Users frequently support the environmental advantages of electric vehicles, such as less emissions and diminished reliance on fossil fuels. In addition, they acknowledge the science and creativity that underlies electric vehicles (EVs), as well as the amazing performance and driving experience they provide. The growing charging infrastructure and government incentives to encourage the use of electric automobiles is contributing to a positive outlook for the EV industry in India.

*2) Negative Sentiment:* Nevertheless, there is also a presence of adverse opinion against the electric vehicle (EV) industry on Twitter. Several users raise worries over the financial implications of electric cars, specifically highlighting the substantial initial expenses and ongoing maintenance costs. Some individuals express doubt over the distance and dependability of electric vehicles, especially in severe weather situations. Moreover, there are apprehensions over the ecological repercussions associated with battery production and the proper disposal of used batteries.

*3) Neutral Sentiment:* Our research revealed that the deviation in the neutral sentiment word cloud was independent of the keyword entered. The vast majority of the results were very similar, featuring keywords related to India's EV industry, leading companies and startups in the sector, and charging infrastructure. The positive sentiment word cloud analysis showed comparable patterns.

Nevertheless, Wordcloud's unfavourable feeling primarily stemmed from the reputation of the leading businesses in the Indian electric vehicle (EV) industry, as well as their service quality and customer feedback. Therefore, the negative sentiment word cloud exhibited a greater degree of polarity and suggestiveness compared to the word cloud including positive and neutral keywords.

## V. Conclusion

India aims to transition towards a more sustainable transport system, with electric vehicles (EVs) anticipated to be vital in achieving its climate and air quality objectives. Many governments worldwide are enacting regulations and providing incentives to promote the uptake of electric cars (EVs), such as offering subsidies for EVs and establishing charging

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:01 UTC from IEEE Xplore. Restrictions apply.

infrastructure. Moreover, through our study, analyzing public opinions is also very important to governments because it explains human activity and behavior and how they are influenced by the opinions of others [7-8]. Nevertheless, for the market to achieve its maximum capacity, it would be crucial for the government and private sector to tackle the obstacles related to charging infrastructure and substantial initial expenses. In all India, these points can take electric vehicles (EV) to the next level, as the Indian government is determined to implement and advocate green technology. Hence it is acceptable that demand for greener modes of transportation is likely to be flourishing lately. A weak point here is that e-vehicles are also facing several obstacles in the route to mass adoption in the nation. There has been great dilemma in the minds of e-vehicle buyers regarding the latest innovations like there are certain electric cars of the same size and power that can be used with less power. Is it worth questioning that we should pay for the most damaged goods?An additional and most important topic is that of the already discussed electric car charging infrastructure in the nation. Due to the limited number of charging stations, electric cars today run considerably differently. We cannot consider electric cars as a medium of travelling long distances. Furthermore, the high cost of batteries remains a significant barrier for several users, given that the initial cost of electric automobiles is considerably greater than that of petrol vehicles.

Considering these obstacles, the general opinion in India towards the electric car industry is predominantly supportive and enthusiastic. The level of awareness and comprehension of the advantages of electric vehicles is on the rise, and it is anticipated that government initiatives to encourage the use of electric transportation will persist in the foreseeable future.

## VI. Future Scope

The current EV industry needs a lot of changes in terms of policies, infrastructure, engineering, etc, but what we fail to realise is that the most important thing which is to be kept in the prime focus is the sentiment of common people in India towards electric vehicles. The change of sentiment, or the positive motivation towards the EV industry in India, should come from the internal motivation of common people in India rather than external. This can be done in many ways, such as a change in government policies, increased incentives for both manufacturers and consumers, infrastructural expansion, raising awareness of EVs in the masses, etc. On a technological front, techniques other than NLP could also be used to gather the sentiment of the people in India towards the industry. But, the most important thing is to analyze the change in the sentiment of people in India towards electric vehicles in further studies and to take action accordingly. This can be done by incorporating qualitative and quantitative studies together at regular intervals of time, so as to depict and predict the current and the future state of the EV industry in India respectively.

## References


[1] P. Sharma, P. K. Singh, A. Verma, "The Challenges faced by EV Industry in India: An Analysis of Consumer Perception," *Research Journal of Humanities and Social Sciences*, vol. 13, no. 3, pp. 162-168, March, 2023.

[2] O.I. Asensio, K. Alvarez, A. Dror, E. Wenzel, and C. Hollauer, "Real-time data from mobile platforms to evaluate sustainable transportation infrastructure," *Nature Sustainability*, vol.3, pp. 463–471, June, 2020.

[3] R. Jena, "An empirical case study on Indian consumers' sentiment towards electric vehicles: A big data analytics approach," *Industrial Marketing Management*, vol. 90, pp. 605-616, Oct. 2020.

[4] M. S. Neethu, and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques." *In 2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, pp. 1-5. July 2013.

[5] K. Lebeau, J. V. Mierlo, P. Lebeau, O. Mairesse, and C. Macharis, "Consumer attitudes towards battery electric vehicles: a large-scale survey," *International Journal of Electric and Hybrid Vehicles, Inderscience Publishers Ltd*, vol. 5, no. 1, pp. 28-41, 2013.

[6] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," *In The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Springer*, no. 11, pp. 508-524, November, 2012.

[7] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study" *Electronics*, vol. 9, no. 3, p. 483, 2020.

[8] S. Kumar et al., "Songs Recommendation using Context-Based Semantic Similarity between Lyrics", *in proc. IEEE INDISCON*, India, 2021.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 05:01:01 UTC from IEEE Xplore. Restrictions apply.

Journal of Applied Physics

RESEARCH ARTICLE | MAY 16 2025

# Ultrasensitive terahertz metasurface for biomolecule detection based on toroidal dipole resonance

Pooja Agarwal ; Kamal Kishor ; Ravindra Kumar Sinha

Check for updates

*J. Appl. Phys.* 137, 193103 (2025)
https://doi.org/10.1063/5.0267155

View Online Export Citation

Articles You May Be Interested In

Substrate-integrated microfluidics for sensitive biosensing with complementary THz metamaterials in water
*Appl. Phys. Lett.* (February 2022)

Dual-surface flexible THz Fano metasensor
*Appl. Phys. Lett.* (November 2017)

Terahertz sensing of 7 nm dielectric film with bound states in the continuum metasurfaces
*Appl. Phys. Lett.* (October 2019)

20 May 2025 07:24:49

Journal of Applied Physics
Special Topics
Open for Submissions
Learn More
AIP Publishing

AIP Publishing

Journal of Applied Physics ARTICLE pubs.aip.org/aip/jap

# Ultrasensitive terahertz metasurface for biomolecule detection based on toroidal dipole resonance

Cite as: J. Appl. Phys. **137**, 193103 (2025); doi: 10.1063/5.0267155
Submitted: 22 February 2025 · Accepted: 1 May 2025 ·
Published Online: 16 May 2025

View Online Export Citation CrossMark

Pooja Agarwal,[1,2] Kamal Kishor,[1] and Ravindra Kumar Sinha[1,a)]

## AFFILIATIONS

[1]TIFAC-Centre of Relevance & Excellence in Fiber Optics & Optical Communications, Applied Physics Department, Delhi Technological University (Formerly Delhi College of Engineering), Bawana Road, Delhi 110 042, India
[2]University Department of Physics, Lalit Narayan Mithila University, Kameshwaranagar, Darbhanga, Bihar 846004, India

[a)]**Author to whom correspondence should be addressed:** dr_rk_sinha@yahoo.com

## ABSTRACT

The presented study proposed a metasurface design with an I-shaped structure for detecting biomolecules in the terahertz (THz) region. The proposed metasurface exhibits three sharp resonance modes $f_1$, $f_2$, and $f_3$ appearing in the transmittance spectra at frequencies of 3.78, 4.32, and 5.00 THz, respectively. The modes $f_1$ and $f_2$ correspond to Fano resonance and mode $f_3$ represents a toroidal dipole resonance. The toroidal dipole resonance arises in the structure as the displacement currents form interconnected loops. An exceptionally high Q-factor of 14 443 is achieved for the toroidal dipole resonance. A maximum figure of merit of 838/RIU was attained when the sensing performance was examined for biomolecules with refractive indices in the range of 1.0–1.4, where RIU is refractive index unit. Additionally, the metasurface demonstrates outstanding sensitivity, with a maximum value of 2163 GHz/RIU, in terms of detecting malaria-infected red blood cells (RBCs) from healthy RBCs. The proposed biosensor holds significant potential for applications in biosensing, enabling the detection and analysis of trace amounts of biomolecules.

20 May 2025 07:24:49

*© 2025 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).* https://doi.org/10.1063/5.0267155

## INTRODUCTION

Biosensing is in high demand across a range of industries, including biochemical analysis, clinical diagnostics, and drug discovery.[1,2] Many biomolecules and biological tissues exhibit unique fingerprints within the terahertz (THz) frequency range. THz spectroscopy and imaging have, thus, emerged as powerful tools for biological investigations, particularly those leveraging resonance interactions with the fingerprint spectra.[3] One of the key advantages lies in their low photon energy, which prevents ionization of biological materials, making them safer for prolonged exposure and ensuring sample integrity during measurement. Additionally, their strong penetration capability enables deeper tissue imaging and material inspection without causing structural damage. THz sensing operates in a label-free and non-destructive manner, eliminating the need for fluorescent or chemical markers that could interfere with biological function.[4] As a result, THz technology is increasingly applied in characterizing cellular components, differentiating between healthy and diseased tissues, and detecting disease-specific biomarkers. Its unique ability to probe both intramolecular and intermolecular forces, such as hydrogen bonds, van der Waals interactions, and hydrophobic effects, enables detailed analysis of complex macromolecules, including amino acids, proteins, nucleic acids (DNA and RNA), and lipids. Since these molecular interactions govern the structural and functional behavior of biomolecules, THz radiation offers the capacity to detect subtle molecular changes, identify specific biomolecular signatures, and diagnose pathological alterations in tissues with high precision.[5] Overall, the non-toxic, non-invasive nature of THz waves positions them as a promising platform for real-time, cost-effective applications in pharmaceutical, biomedical, and therapeutic domains.[6–11]

Despite notable advancements in THz spectroscopy and imaging, practical implementations still face limitations such as low sensitivity and image contrast, restricting its broader adoption in high-performance sensing applications.[12,13] The inherently weak interaction of THz waves with many materials and the low

signal-to-noise ratio in conventional THz systems have hindered their utility in detecting minute changes in analytes. In this context, metamaterials offer the fascinating possibility to overcoming these challenges. Metamaterials are generally composed of a periodic array of meta-atoms, which are key to defining the properties and functionality of meta-photonic devices.[14–16] The electromagnetic behavior of metamaterials is primarily dictated by the shape, size, and orientation of their constituent meta-atoms.[17–19] Metasurface, the 2D counterparts of metamaterial enable more compact and cost-effective designs while maintaining functional versatility.[20–22] These metasurfaces are designed to excite a variety of resonance modes, such as electromagnetic-induced transparency (EIT), surface plasmon resonance (SPR), LC resonance, and Mie resonances.[23,24] These resonant features have been extensively employed in THz biosensing to detect biomolecular analytes through measurable frequency shifts in the resonant spectrum.[25] These resonances are also vital in amplifying the interaction between light and matter by enabling artificial manipulation of electromagnetic properties.[26,27] The resulting spectral shifts or amplitude changes at resonant frequencies offer robust mechanisms for both qualitative and quantitative analyses of trace analytes.[28] By harnessing the unique properties of terahertz radiation, metasurface sensors facilitate rapid, non-invasive, and highly sensitive trace detection of biomolecules.

According to the Mie theory, electromagnetic multipole resonances in metasurface structures can induce electric or magnetic responses. Among these, toroidal dipole (TD) moments provide a unique mechanism to enhance light–matter interactions due to their ability to tightly confine electromagnetic fields and support extremely sharp resonance line shapes.[29] These features make TDs valuable for applications in sensing, data processing and information storage, and nonlinear effects.[30] Initially identified in molecular structures and ferroelectric systems, the electromagnetic signatures of toroidal moments were frequently overshadowed by the more dominant effect of electric and magnetic multipoles, making their detection a significant challenge.[31] Their inherently weak interaction with incident waves further complicated experimental observation. However, the advent of metamaterials enabled the successful detection of TD responses. Toroidal dipole resonances (TDRs) were first directly observed in SRR metamaterials in 2010.[32] Notably, strong TD excitations have been observed in all-dielectric metasurfaces, further highlighting their potential in advanced sensing and photonic applications. TDRs in all-dielectric metamaterials were achieved by Basharin *et al.* in 2015.[33] Metallic materials possess inherent electromagnetic properties, making it challenging to dynamically control the desired function once a metallic metasurface is fabricated. Various materials, including traditional semiconductors, phase-change materials, liquid crystals, perovskites, and Dirac semimetals, have been investigated to enable multi-mode operation.[34] Among these, traditional semiconductors such as silicon, germanium, and gallium arsenide stand out due to their superior performance in practical applications. Their advantages include well-established fabrication techniques, cost-effective processing, and a large modulation depth, making them highly suitable for dynamic and reconfigurable metasurface designs.[35] The absence of Ohmic losses in high-refractive-index materials gives them a low dissipation loss advantage over their metallic counterparts, which makes them an appropriate substitute for toroidal excitations.[36] Numerous studies have been conducted on TD excitations, which focused on exploring their fascinating physical phenomena.[37–40] TDRs in all dielectric metasurfaces making them well-suited for achieving substantial field confinement, necessary for various applications.[41–43]

In this paper, we report the design of a metasurface biosensor consisting of an I-shaped resonator that enables TDR to be excited in the THz regime. The transmittance spectra reveal three prominent resonance modes $f_1$, $f_2$, and $f_3$ at frequencies 3.78, 4.32, and 5.00 THz, respectively. The mode $f_3$ exhibits TDR resulting in a high Q factor of 14 443. The transmittance spectrum can be understood through a detailed analysis of the electric field distribution. The highest figure of merit (FOM) of 838/RIU is reported as the refractive indices of the surrounding analyte vary from 1.0 to 1.4, where RIU is refractive index unit. The biosensor also shows ultrahigh sensing properties (maximum sensitivity 2163 GHz/RIU), when implemented in malaria detection. The proposed metasurface exhibits ultrasensitive detection capabilities, particularly for malaria diagnostics, and holds significant potential for various biomedical sensing applications.

## DESIGN AND MODELING

The proposed metasurface design for biomolecule detection is depicted in the schematic diagram shown in Fig. 1(a). Figure 1(b) shows the unit cell of the metasurface, which consists of an I-shaped resonator made of gallium arsenide (GaAs) on a quartz substrate. The geometric parameters of the unit cell structure are $p = p_x = p_y = 45\,\mu$m, $l_1 = 14\,\mu$m, $l_2 = 8\,\mu$m, $t_m = 4\,\mu m$, where the period $p$ is uniform in both the $x$- and $y$-directions. The resonator is characterized by its length $l_1$, width $l_2$, and thickness $t_m$. The refractive index value for GaAs is considered to be 3.65 and for quartz it is 2.13 and have been taken from the Palik refractive index database for the frequency range of interest.[44] Numerical simulations were performed using COMSOL Multiphysics to analyze the spectral response of the metasurface and an $x$-polarized plane wave

20 May 2025 07:24:49

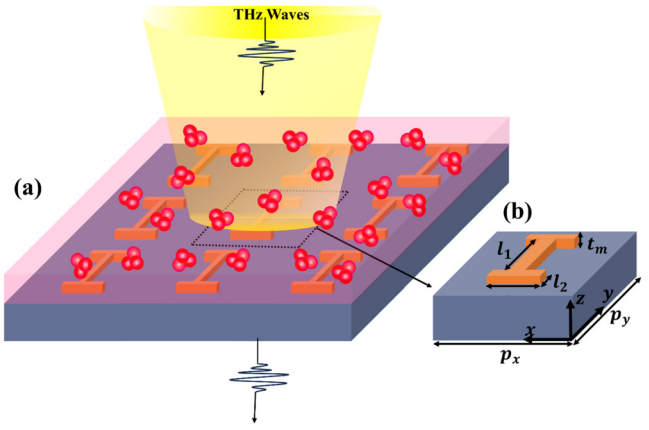



**FIG. 1.** (a) Schematic representation of the designed metasurface and (b) a detailed view of the unit cell structure of the proposed metasurface.


© Author(s) 2025

propagated in the $z$-axis, and periodic boundary conditions were applied in the $x$ and $y$ directions.

The fabrication feasibility of the proposed metasurface is carefully considered, encompassing a multi-step process. A GaAs wafer is first spin-coated with a photoresist and patterned using standard UV photolithography techniques. The patterned wafer is then bonded to a quartz substrate using a UV-curable adhesive, ensuring strong adhesion and optical compatibility. Subsequent etching is performed using an Inductively Coupled Plasma (ICP) etching system to transfer the pattern into the GaAs layer with high precision.[45]

## RESULTS AND DISCUSSION

The spectral response for transmittance is illustrated in Fig. 2(a), where three distinct Fano resonance modes $f_1$, $f_2$, and $f_3$ observed at frequencies 3.78, 4.32, and 5.00 THz, respectively. The transmittance spectra exhibit Fano-type resonances due to the interference between discrete Mie resonance modes and broad continuum modes. In high-index dielectric resonators, Mie resonances arise when incident electromagnetic waves excite localized oscillations of electric and magnetic dipoles as well as higher-order multipoles, resulting in strong scattering at specific frequencies. When these sharp, discrete resonances couple with a broad spectral background or continuum of modes, their interference results in the characteristic asymmetric line shape known as a Fano resonance.[46]

The Fano line shape formula described in Eq. (1) is used to fit the transmittance spectra,

$$T_{\text{Fano}} = \left| a_1 + ia_2 + \frac{b}{\omega - \omega_0 + i\gamma} \right|^2, \tag{1}$$

where $a_1$, $a_2$, and $b$ are constant real numbers; $\omega_0$ represents the central resonant frequency; and $\gamma$ denotes the damping rate.[47,48] Figures 2(b)–2(d) present the corresponding Fano fitting curves. The damping rates associated with various modes were determined through Fano fitting, and the corresponding Q-factors were subsequently calculated using the formula

$$Q = \frac{\omega_0}{2\gamma}. \tag{2}$$

Typically, a sharper resonance mode leads to a higher Q factor. The computed Q-factors for the mode $f_1$, $f_2$, and $f_3$ are 6037.95, 349.88, and 14 443, respectively.

The multipolar expansion method is employed to investigate the characteristics of the different resonance modes.[49,50] This approach is highly effective for precisely analyzing the far-field radiation pattern. As part of this study, the displacement current density, $\boldsymbol{J}_\omega(\boldsymbol{r})$, within the unit cell is defined as

$$\boldsymbol{J}_\omega(\boldsymbol{r}) = i\omega\varepsilon_0(n^2 - 1)\boldsymbol{E}_\omega(\boldsymbol{r}), \tag{3}$$

where $\boldsymbol{E}_\omega(\boldsymbol{r})$ is the electric field, $\varepsilon_0$ is the permittivity of free space, and $n$ is the effective refractive index. In the Cartesian coordinate system, the multipoles are defined by integrating the charge or current density, $\boldsymbol{J}_\omega(\boldsymbol{r})$ within the unit cell,

$$\text{ED: } \boldsymbol{P}_\alpha = -\frac{1}{i\omega}\int d^3\boldsymbol{r}\, \boldsymbol{J}_{\boldsymbol{\alpha}}^{\boldsymbol{\omega}}\, j_0(kr), \tag{4a}$$

$$\text{MD: } \boldsymbol{M}_\alpha = \frac{3}{2}\int d^3\boldsymbol{r}(\boldsymbol{r}\times\boldsymbol{J}_\omega)_{\boldsymbol{\alpha}}\frac{j_1(kr)}{kr}, \tag{4b}$$

$$\text{TD: } \boldsymbol{T}_\alpha = \frac{i}{\omega}\int [d^3\boldsymbol{r}\left[3r_\alpha(\boldsymbol{r}.\boldsymbol{J}_\omega) - r^2\boldsymbol{J}_{\boldsymbol{\alpha}}^{\boldsymbol{\omega}}\right]\frac{j_2(kr)}{(kr)^2}, \tag{4c}$$

$$\text{EQ: } \boldsymbol{Q}^e_{\alpha\beta} = -\frac{3}{i\omega}\int d^3\boldsymbol{r}\left[3\left(\boldsymbol{r}_{\boldsymbol{\beta}}\boldsymbol{J}_{\boldsymbol{\alpha}}^{\boldsymbol{\omega}} + \boldsymbol{r}_{\boldsymbol{\alpha}}\boldsymbol{J}_{\boldsymbol{\beta}}^{\boldsymbol{\omega}}\right) - 2(\boldsymbol{r}\cdot\boldsymbol{J}_\omega)\delta_{\alpha\beta}\right]\frac{j_1(kr)}{kr}, \tag{4d}$$

$$\text{MQ: } \boldsymbol{Q}^m_{\alpha\beta} = 5\int d^3\boldsymbol{r}[\boldsymbol{r}_{\boldsymbol{\alpha}}(\boldsymbol{r}\times\boldsymbol{J}_{\boldsymbol{\omega}})_\beta + \boldsymbol{r}_{\boldsymbol{\beta}}(\boldsymbol{r}\times\boldsymbol{J}_{\boldsymbol{\omega}})_\alpha]\frac{j_2(kr)}{(kr)^2}. \tag{4e}$$

The spherical Bessel function is represented by $j_q(kr)$, $k$ indicates the magnitude of the propagation vector, and $\boldsymbol{r}$ is the position vector from the origin to a point $(\alpha, \beta = x, y, z)$ in the Cartesian coordinate system. The representations for the electric dipole (ED), magnetic dipole (MD), toroidal dipole (TD), electric quadrupole (EQ), and magnetic quadrupole (MQ) moments are $\boldsymbol{P}_\alpha$, $\boldsymbol{M}_\alpha$, $\boldsymbol{T}_\alpha$, $\boldsymbol{Q}^e_{\alpha\beta}$, and $\boldsymbol{Q}^m_{\alpha\beta}$, respectively. Each multipole corresponds to a distinct scattering channel. Higher-order terms in the multipole expansion are excluded from the computations due to their negligible contributions. The following are the expression for the total scattered power:

$$I_{Scat} = I_P + I_M + I_T + I_{Q^e} + I_{Q^m}, \tag{5a}$$

where the scattered powers of these multipoles are determined from

$$I_P = \frac{2\omega^4}{3c^3}|\boldsymbol{P}|^2, \tag{5b}$$

$$I_M = \frac{2\omega^4}{3c^3}|\boldsymbol{M}|^2, \tag{5c}$$

$$I_T = \frac{2\omega^6}{3c^5}|\boldsymbol{T}|^2, \tag{5d}$$

$$I_{Q^e} = \frac{\omega^6}{5c^5}|\boldsymbol{Q}^e|^2, \tag{5e}$$

$$I_{Q^m} = \frac{\omega^6}{40c^5}|\boldsymbol{Q}^m|^2. \tag{5f}$$

Figures 3(a)–3(c) illustrate the multipole contributions to the far-field scattering for the resonance modes $f_1$, $f_2$, and $f_3$, respectively. The calculations reveal that each resonance mode is characterized by the dominance of a specific multipole component. In mode $f_1$, the ED is the predominant contribution, followed by the

20 May 2025 07:24:49


© Author(s) 2025

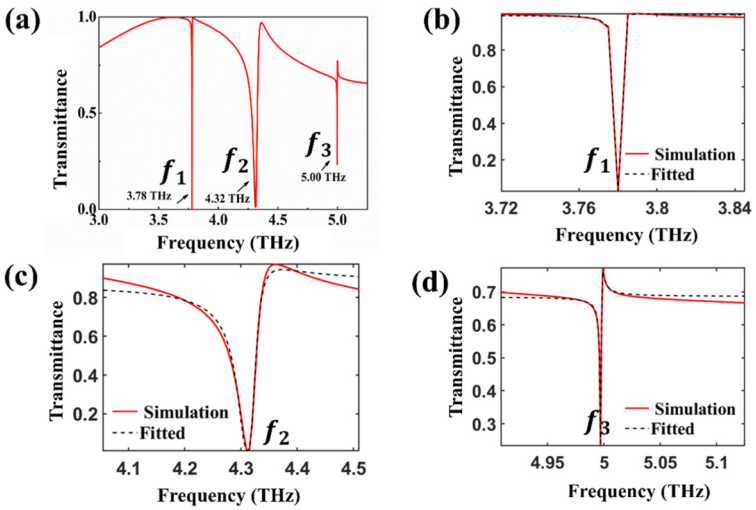



**FIG. 2.** (a) Transmittance curve for all three distinct modes labeled as modes $f_1$, $f_2$, and $f_3$, respectively. (b), (c), and (d) display the Fano fitting for modes $f_1$, $f_2$, and $f_3$. The solid curves represent the simulation results, while the dashed curves indicate the Fano fitting results.

MQ as shown in Fig. 3(a). The MD is dominating for the mode $f_2$, followed by the EQ as depicted in Fig. 3(b). In mode $f_3$, the TD is the dominant component, leading to extremely low radiative loss and weak coupling to the radiation continuum, as illustrated in Fig. 3(c).

TDRs are highly sensitive to environmental perturbations, enabling the precise detection of even minor frequency shifts. These shifts may arise from trace analytes binding to the sensor surface or slight variations in parameters such as the refractive index. The ability to detect these small shifts is critical for applications in biomedical diagnostics, where sensitivity and accuracy are paramount.

The electromagnetic field distributions are presented in Fig. 4 to provide a better comprehension of the physical mechanisms underlying the different Fano resonance modes $f_1$, $f_2$, and $f_3$. The electric field distributions for these modes are shown in Figs. 4(a)–4(c), respectively. The displacement currents, indicated by red arrows in electric field profiles, illustrate the mode-specific current flow. It is observed in Fig. 4(c) that displacement current forms loops, with each loop connecting to the next in a head to tail manner, confirming that $f_3$ corresponds to the TDR mode. The magnetic field distribution for the modes $f_1$, $f_2$, and $f_3$ are depicted in Figs. 4(d)–4(f), respectively.

The influence of geometric parameters $p$, $l_1$, $l_2$, and $t_m$ on the transmittance spectrum is analyzed and illustrated in Figs. 5(a)–5(d), respectively. As the geometric parameters increase, a pronounced red shift in the resonance frequencies of all modes is observed. This shift in resonance can be attributed to the Mie theory, as explained by the Lewin model,[51]

$$f = \theta c/2\pi V(x, y, z)\sqrt{\varepsilon_1 \mu_1}, \tag{6}$$

where $\theta$ remains a constant for the particular resonance. The parameters $\varepsilon_1$ and $\mu_1$ are the permittivity and permeability of the dielectric material, respectively, while $V(x, y, z)$ denotes a function associated with the size of the dielectric resonator. According to

20 May 2025 07:24:49

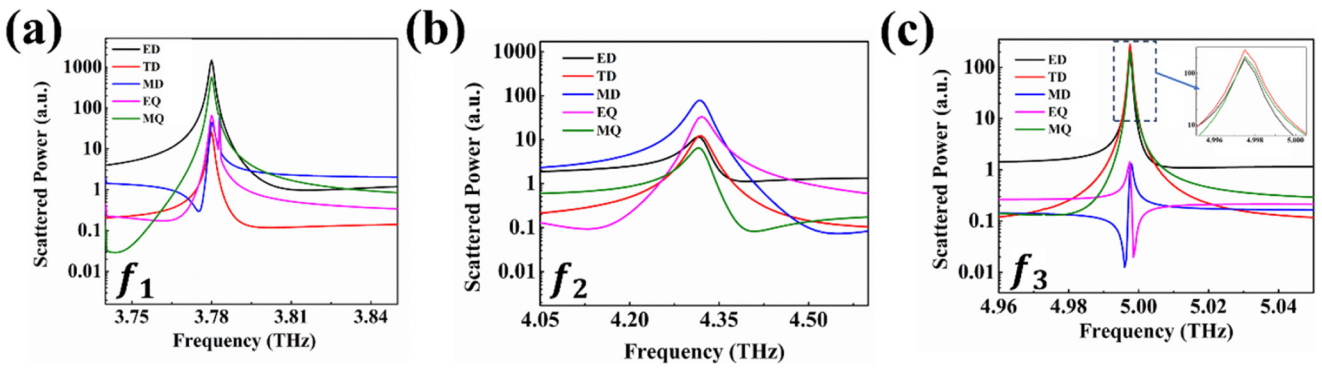



**FIG. 3.** (a)–(c) Calculated scattered power for various multipoles for modes$f_1$, $f_2$, and $f_3$. The curves for ED, MD, TD, EQ, and MQ represent their individual contributions to the scattering power.


© Author(s) 2025

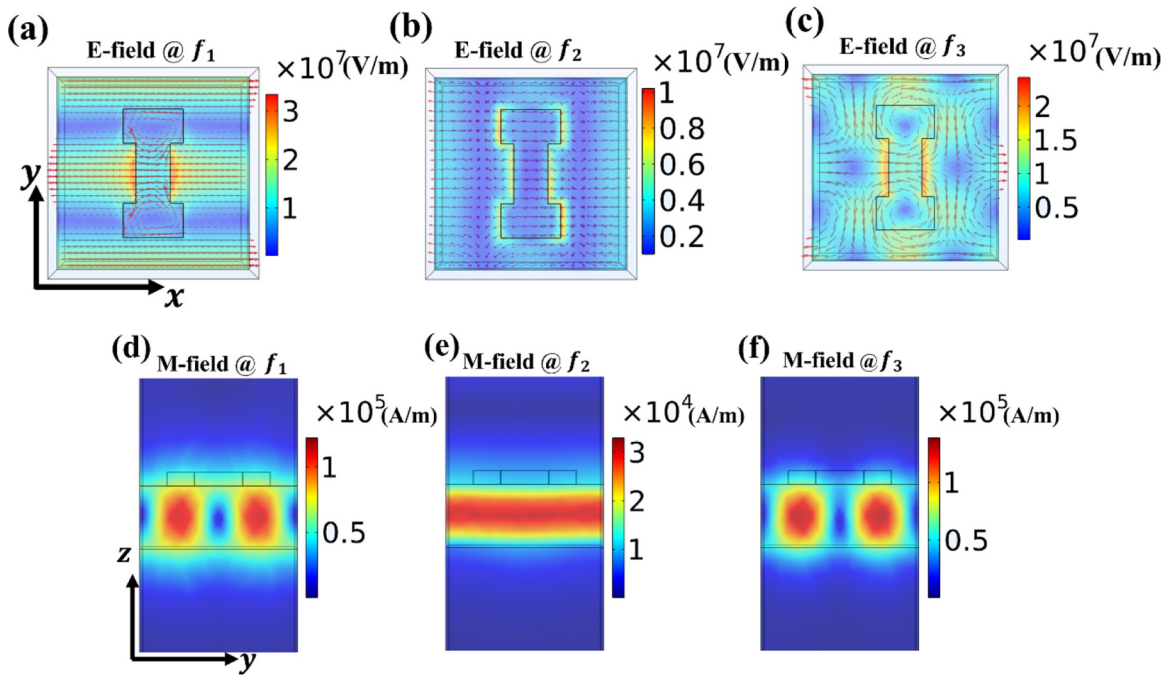



**FIG. 4.** Distribution of electromagnetic fields and displacement current corresponding to the resonant frequencies of modes $f_1$, $f_2$, and $f_3$, respectively. (a)–(c) Electric field distributions and displacement current vector of modes $f_1$, $f_2$, and $f_3$, respectively. (d)–(f) Magnetic field distributions. Red arrows in the electric field profile represent the displacement current.

20 May 2025 07:24:49

Eq. (6), the resonance frequency is governed by both the physical dimensions and the electromagnetic properties of the dielectric resonator. An increase in the values of $p$, $l_2$, $t_m$, and $l_1$ results in a larger volume of the medium oscillator $V(x, y, z)$. As a consequence, the resonance frequency shifts to a lower value, leading to a redshift, as illustrated in the corresponding figure.

The proposed metasurface structure, featuring an ultranarrow linewidth enabled by the TDR, holds significant potential for sensing applications. To validate its performance, we numerically evaluated the sensing capabilities of the designed biosensor. Most biomolecules including proteins, lipids, nucleic acids, and other cellular components exhibit refractive indices within the range of 1.0–1.4.[52,53] The variations in refractive index within this range are crucial in sensing techniques, as they influence light–matter interactions, enabling precise detection and characterization of biomolecules in biomedical and biochemical applications. As shown in Fig. 6(a), an increase in the refractive index of the surrounding medium from 1.0 to 1.4 causes all three resonance dips to undergo a significant red shift in frequency. These results clearly indicate a strong correlation between the sensor's spectral response and the refractive index of the analyte. The variation in the intensity of the resonance modes with changes in the refractive index of the surrounding medium is significantly observed. This behavior arises from the frequency-dependent effective refractive index of the metasurface. At a particular resonance frequency, when the refractive index of the surrounding medium approaches the effective refractive index of the metasurface, the transmission is nearly zero. In contrast, when the refractive index of the surrounding medium is lower than the effective refractive index of the metasurface, it results in a different, often higher transmission.[54] Figure 6(b) illustrates the relationship between the resonant frequency and the refractive index. The fitting results reveal a linear relationship between the resonant frequency shift and the refractive index change for all three dips.

The sensitivity of the biosensor can be evaluated using the relation $S = \frac{\Delta f}{\Delta n}$, where $\Delta f$ is the shift in the resonance frequency in the response of the change in the refractive index of the surrounding medium $\Delta n$.[55] The evaluated sensitivity is ∼307, 960, and 290 GHz/RIU for the modes $f_1$, $f_2$, and $f_3$, respectively, where RIU represents the refractive index unit. The figure of merit is defined as the ratio of sensitivity (S) to the full width at half maximum (FWHM) of the resonance dip, expressed as FOM = S/FWHM. The evaluated FOM is 490.41/RIU, 77.85/RIU, and 838/RIU for the modes $f_1$, $f_2$, and $f_3$, respectively. The high sensing performance makes the biosensor highly suitable for practical applications, especially in scenarios requiring precise detection of refractive index changes, such as in biomedical domains.

Furthermore, we investigate the application of the proposed biosensor for identifying malaria infection. Rapid screening for malaria plays an essential role in global healthcare efforts. Plasmodium parasites that infiltrate red blood cells (RBCs) are the root cause of the disease, with the severity of infection being closely


© Author(s) 2025

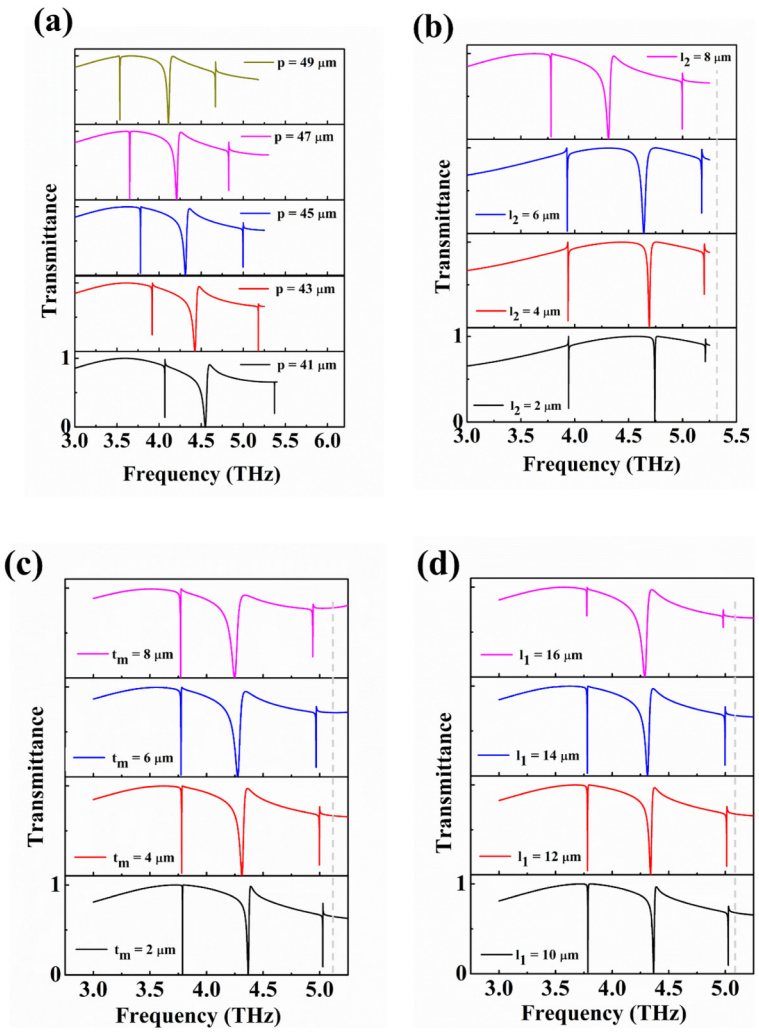



**FIG. 5.** Transmittance spectra of the metasurface for different structural parameters: (a) period $p$, (b) width $l_2$, (c) thickness $t_m$, and (d) length $l_1$.

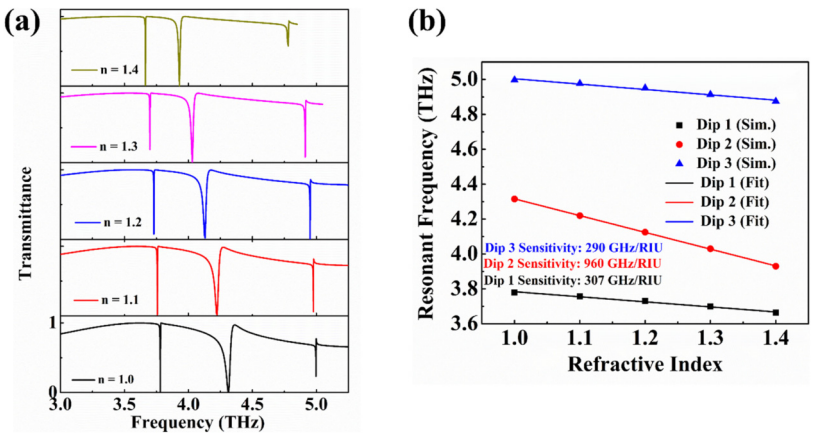



**FIG. 6.** (a) Transmittance spectra and (b) resonance frequency variation with respect to varying refractive indices of the surrounding analyte of the proposed metasurface biosensor.

20 May 2025 07:24:49


© Author(s) 2025

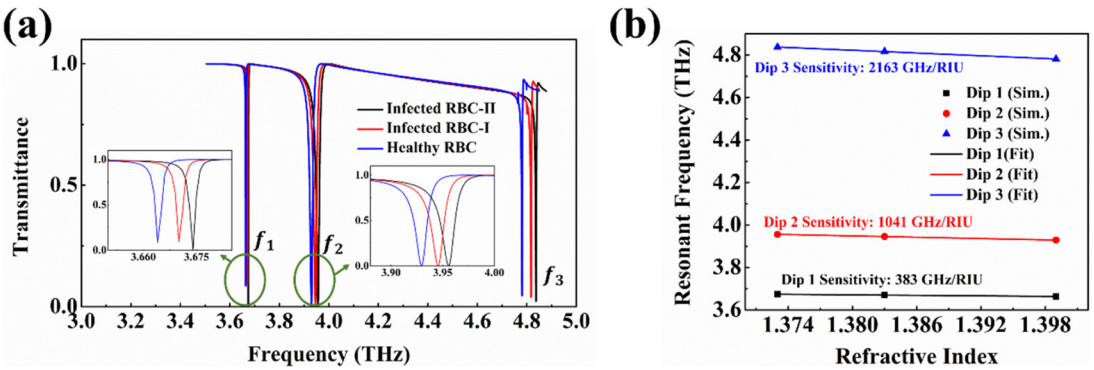



**FIG. 7.** (a) Transmittance spectra of the proposed biosensor for detection of malaria with respect to refractive index values 1.399 (healthy RBC), and 1.373 and 1.383 (infected RBCs). (b) The change in resonance frequency with varying refractive indices of the surrounding analyte, along with linear fitting (solid line).

linked to changes in the properties of these RBCs. Healthy RBCs have a refractive index of 1.399, a key property that can be leveraged to detect malaria infection. As the infection progresses, the optical properties of RBCs alter significantly, allowing for the potential to detect these changes and thus monitor the stages of malaria. Specifically, infected RBCs exhibit a refractive index of 1.373 in the schizont stage and 1.383 in the trophozoite stage, indicating measurable variations that can serve as potential biomarkers for malaria detection.[56]

As shown in Fig. 7(a), a frequency shift is observed in the transmittance spectra when the refractive index changes from 1.399 to 1.383 (trophozoite stage) to 1.373 (schizont stage). In the case of malaria detection, the sensor has demonstrated high sensitivity values of 383, 1041, and 2163 GHz/RIU for the modes $f_1$, $f_2$, and $f_3$, respectively, as depicted in Fig. 7(b). These sensitivity values indicate that the sensor is capable of detecting even the slightest variations in the refractive index of infected red blood cells, making it a promising tool for real-time, *in situ* malaria diagnosis. This success highlights the potential of the proposed sensor to accurately identify the different stages of malaria, offering a non-invasive, cost-effective, and rapid diagnostic tool for the global fight against the disease. The proposed biosensor can not only detect malaria at various stages but also show potential for adaptation to other infectious diseases, particularly those associated with refractive index changes in biological samples. Its robust performance across a broad refractive index range makes it highly suitable for diverse biomedical applications, including the early-stage detection of diseases where biomarker concentrations are extremely low.

The complementary features of ultrahigh Q-factor, simple design, and non-invasive detection establish the proposed biosensor as a versatile and high-performance platform for next-generation terahertz biosensing. The continued development of metasurface-based biosensors could revolutionize diagnostic approaches, particularly in regions with limited access to traditional healthcare facilities, thereby contributing to more effective disease control and prevention strategies worldwide.

Finally, a comparative analysis of the sensing performance between the proposed metasurface and previously reported THz metasurface designs is presented in Table I. Our proposed structure demonstrates an exceptionally high Q factor and a notable FOM, outperforming several recent THz metasurface sensors. These enhancements offer clear advantages for sensing applications.

20 May 2025 07:24:49

## CONCLUSION

To summarize, we have proposed a metasurface sensor to detect biomolecules, featuring an I-shaped resonator design. The metasurface demonstrates three distinct sharp dips $f_1$, $f_2$, and $f_3$ in the transmittance spectrum at frequencies 3.78, 4.32, and 5.00 THz, respectively. Multipolar analysis reveals that mode $f_1$ is primarily governed by an ED, mode $f_2$ by a MD, and mode $f_3$ by a TD. The highest Q-factor of 14 443 is achieved for the TDR. The sensing performance has been evaluated for refractive index variations in the surrounding medium, ranging from 1.0 to 1.4, achieving a maximum FOM of 838/RIU. When applied to malaria detection, the metasurface demonstrates ultrahigh sensitivity, reaching up to

**TABLE I.** Performance comparison of the metasurface with previously reported sensors.

| Reference | Resonant mode | Sensitivity | FOM | Q-factor | Year |
|---|---|---|---|---|---|
| 57 | TD resonance | 521.6 GHz/RIU | 244/RIU | 1104 | 2019 |
| 58 | TD resonance | … | 515/RIU | 3189 | 2019 |
| 48 | Fano resonance | 2000 GHz/RIU | 514.40 | 920 | 2024 |
| 43 | TD resonance | 59.64 GHz/RIU<br>290 GHz/RIU | … | 272.84 | 2024 |
| This work | TD resonance | 2163 GHz/RIU (malaria detection) | 838/RIU | 14 443 | |


© Author(s) 2025

2163 GHz/RIU. These results underscore the exceptional sensitivity of the proposed sensor, particularly for malaria diagnostics, and highlight its potential for diverse biomedical sensing applications.

## ACKNOWLEDGMENTS

This work was supported by the TIFAC-Centre of Relevance and Excellence in Fiber Optics and Optical Communication at Delhi College of Engineering, now DTU, Delhi, through the Mission REACH program of Technology Vision-2020, Government of India. Pooja Agarwal would like to thank Dr. Yogita Kalra, DTU Delhi for valuable discussions on this research work.

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

### Author Contributions

**Pooja Agarwal:** Conceptualization (lead); Formal analysis (lead); Resources (equal); Software (equal); Validation (equal); Writing – original draft (lead). **Kamal Kishor:** Resources (equal); Software (equal); Supervision (equal); Validation (equal); Writing – review & editing (equal). **Ravindra Kumar Sinha:** Resources (equal); Software (equal); Supervision (lead); Validation (equal); Writing – review & editing (equal).

## DATA AVAILABILITY

The data that support the findings of this study are available within the article.

## REFERENCES

[1]S. Menon, M. R. Mathew, S. Sam, K. Keerthi, and K. G. Kumar, "Recent advances and challenges in electrochemical biosensors for emerging and re-emerging infectious diseases," J. Electroanal. Chem. **878**, 114596 (2020).
[2]L. Ahlawat, K. Kishor, and R. K. Sinha, "Photonic spin Hall effect-based ultra-sensitive refractive index sensor for haemoglobin sensing applications," Opt. Laser Technol. **170**(June 2023), 110183 (2024).
[3]J. Lou, Y. Jiao, R. Yang, Y. Huang, X. Xu, L. Zhang, Z. Ma, Y. Yu, W. Peng, Y. Yuan, Y. Zhong, S. Li, Y. Yan, F. Zhang, J. Liang, X. Du, C. Chang, and C. W. Qiu, "Calibration-free, high-precision, and robust terahertz ultrafast meta-surfaces for monitoring gastric cancers," Proc. Natl. Acad. Sci. U.S.A. **119**(43), 1–9 (2022).
[4]L. Ma, F. Fan, W. Shi, Y. Ji, X. Wang, and S. Chang, "Retime-mapping terahertz vernier biosensor for boosting sensitivity based on self-reference waveguide inter-ferometers," Fundam. Res. **5**(2), 593-601 (2025).
[5]Z. Zhang, Z. Wang, C. Zhang, Z. Yao, S. Zhang, R. Wang, Z. Tian, J. Han, C. Chang, J. Lou, X. Yan, and C. Qiu, "Advanced terahertz refractive sensing and fingerprint recognition through metasurface-excited surface waves," Adv. Mater. **36**(14), 1–9 (2024).
[6]L. Sun, L. Zhao, and R. Y. Peng, "Research progress in the effects of terahertz waves on biomacromolecules," Mil. Med. Res. **8**(1), 1–8 (2021).
[7]L. Ma, F. Fan, J. Feng, P. Shen, H. Yin, Y. Ji, X. Wang, and S. Chang, "Lag-compensated hyperfine terahertz dual-comb interferometer beyond intrinsic resolu-tion and sensitivity," Laser Photonics Rev. **2401784**, 1–9 (published online 2025).
[8]X. Yang, X. Zhao, K. Yang, Y. Liu, Y. Liu, W. Fu, and Y. Luo, "Biomedical appli-cations of terahertz spectroscopy and imaging," Trends Biotechnol. **34**(10), 810–824 (2016).
[9]Y. Roh, S. H. Lee, J. Kwak, H. S. Song, S. Shin, Y. K. Kim, J. W. Wu, B. K. Ju, B. Kang, and M. Seo, "Terahertz imaging with metamaterials for biological appli-cations," Sens. Actuators B: Chem. **352**(P1), 130993 (2022).
[10]L. Engineering, "Cutting-edge terahertz technology," Nat. Photonics **1**, 97–105 (2002).
[11]L. Ma, F. Fan, W. Shi, Y. Ji, X. Wang, and S. Chang, "Terahertz on-chip sensor based on Mach-Zehnder waveguide interferometer for selective recognition of reducing drug," Sens. Actuators A: Phys. **370**(October 2023), 115282 (2024).
[12]L. Yang, T. Guo, X. Zhang, S. Cao, and X. Ding, "Toxic chemical compound detection by terahertz spectroscopy: A review," Rev. Anal. Chem. **37**(3), 20170021 (2018).
[13]E. MacPherson, G. P. Gallerano, G. S. Park, H. Hintzsche, and G. J. Wilmink, "Guest editorial: Terahertz imaging and spectroscopy for biology and biomedi-cine," IEEE J. Biomed. Health Inform. **17**(4), 765–767 (2013).
[14]N. I. Zheludev and Y. S. Kivshar, "From metamaterials to metadevices," Nat. Mater. **11**(11), 917–924 (2012).
[15]N. I. Landy, S. Sajuyigbe, J. J. Mock, D. R. Smith, and W. J. Padilla, "Perfect metamaterial absorber," Phys. Rev. Lett. **100**(20), 207402 (2008).
[16]Ankit, M. N. Baitha, K. Kishor, and R. K. Sinha, "Quadrupole mode plasmon resonance enabled dual-band metamaterial for refractive index sensing applica-tion," J. Appl. Phys. **136**(2), 1–14 (2024).
[17]K. Kishor, R. K. Sinha, M. N. Baitha, and B. Lahiri, "Tunable negative refrac-tive index metamaterial from V-shaped SRR structure: Fabrication and character-ization," J. Opt. Soc. Am. B **31**(7), 1410–1414 (2014).
[18]Ankit, K. Kishor, R. K. Sinha, and R. K. Sinha, "Design and analysis of a far-infrared metamaterial perfect absorber with sensing applications," Appl. Opt. **63**(35), 8994–9001 (2024).
[19]K. Kishor, M. N. Baitha, and R. K. Sinha, "Optik design and simulation of 'I' shaped split ring resonator metamaterial at optical communication window around 1.55 $\mu$m," Optik **126**(23), 4708–4711 (2015).
[20]Ankit, K. Kishor, and R. K. Sinha, "Design, fabrication, and characterization of epsilon negative and near-zero index metasurface," Appl. Phys. A **130**(2), 1–15 (2024).
[21]S. Zhang, C. L. Wong, S. Zeng, R. Bi, K. Tai, K. Dholakia, and M. Olivo, "Metasurfaces for biomedical applications: Imaging and sensing from a nano-photonics perspective," Nanophotonics **10**(1), 259–293 (2021).
[22]J. Zhu, C. Li, B. Fang, J. Lu, Z. Hong, and X. Jing, "Advancements in biosens-ing detection based on terahertz metasurfaces," Opt. Laser Technol. **176**(March), 110978 (2024).
[23]M. Beruete and I. Jáuregui-López, "Terahertz sensing based on metasurfaces," Adv. Opt. Mater. **8**(3), 1–26 (2020).
[24]D. Hao, J. Liu, P. Zou, Y. Zhang, R. Moro, and L. Ma, "All-dielectric metasur-faces and their applications in the terahertz range," Laser Photonics Rev. **18**(9), 2301210 (2024).
[25]P. Shen, Y. Ji, X. Yue, Y. Li, M. Han, Y. Ma, M. Meng, F. Fan, and S. Chang, "Ultrasensitive terahertz microfluidic biosensor integrated with tetrahedral DNA nanostructure for specific detection of live cancer cells," Sens. Actuators B: Chem. **428**(January), 137252 (2025).
[26]Q. He, S. Sun, S. Xiao, and L. Zhou, "High-efficiency metasurfaces: Principles, realizations, and applications," Adv. Opt. Mater. **6**(19), 1–23 (2018).
[27]Q. Wang, Y. Chen, J. Mao, F. Yang, and N. Wang, "Metasurface-assisted tera-hertz sensing," Sensors **23**(13), 1–17 (2023).
[28]J. Qi, C. Li, Y. Xia, B. Fang, Z. Hong, L. Chen, and X. Jing, "Research progress of biomolecular detection based on metasurfaces," Infrared Phys. Technol. **136**-(December 2023), 105111 (2024).
[29]V. Savinov, V. A. Fedotov, and N. I. Zheludev, "Toroidal dipolar excitation and macroscopic electromagnetic properties of metamaterials," Phys. Rev. B **89**, 205112 (2014).
[30]N. Talebi, S. Guo, and P. A. Van Aken, "Theory and applications of toroidal moments in electrodynamics: Their emergence, characteristics, and technological relevance," Nanophotonics **7**(1), 93–110 (2018).
[31]V. M. Dubovik and V. V. Tugushev, "Toroid moments in electrodynamics and solid-state physics," Phys. Rep. **187**(4), 145–202 (1990).

20 May 2025 07:24:49


© Author(s) 2025

[32]T. Kaelberer, V. A. Fedotov, N. Papasimakis, D. P. Tsai, and N. I. Zheludev, "Toroidal dipolar response in a metamaterial," Science **330**(6010), 1510–1512 (2010).
[33]A. A. Basharin, M. Kafesaki, E. N. Economou, C. M. Soukoulis, V. A. Fedotov, V. Savinov, and N. I. Zheludev, "Dielectric metamaterials with toroidal dipolar response," Phys. Rev. X **5**(1), 011036 (2015).
[34]Y. Jiao, J. Lou, Z. Ma, L. Cong, X. Xu, B. Zhang, D. Li, Y. Yu, W. Sun, Y. Yan, S. Hu, B. Liu, Y. Huang, L. Sun, R. Wang, R. Singh, Y. Fan, C. Chang, and X. Du, "Photoactive terahertz metasurfaces for ultrafast switchable sensing of colorectal cells," Mater. Horiz. **9**(12), 2984–2992 (2022).
[35]J. Zhang, J. Lou, Z. Wang, J. Liang, X. Zhao, Y. Huang, C. Chang, and G. Hu, "Photon-induced ultrafast multitemporal programming of terahertz metadevices," Adv. Mater. **37**(7), 2410671 (2024).
[36]V. R. Tuz, V. V. Khardikov, and Y. S. Kivshar, "All-dielectric resonant metasurfaces with a strong toroidal response," ACS Photonics **5**(5), 1871–1876 (2018).
[37]L. Cong, Y. K. Srivastava, and R. Singh, "Tailoring the multipoles in THz toroidal metamaterials," Appl. Phys. Lett. **111**(8), 081106 (2017).
[38]C. Zhou, S. Li, Y. Wang, and M. Zhan, "Multiple toroidal dipole Fano resonances of asymmetric dielectric nanohole arrays," Phys. Rev. B **100**(19), 195306 (2019).
[39]M. Gupta and R. Singh, "Toroidal versus Fano resonances in high Q planar THz metamaterials," Adv. Opt. Mater. **4**(12), 2119–2125 (2016).
[40]S. Xu, A. Sayanskiy, A. S. Kupriianov, V. R. Tuz, P. Kapitanova, H. B. Sun, W. Han, and Y. S. Kivshar, "Experimental observation of toroidal dipole modes in all-dielectric metasurfaces," Adv. Opt. Mater. **7**(4), 1–9 (2019).
[41]M. Gupta, Y. K. Srivastava, M. Manjappa, and R. Singh, "Sensing with toroidal metamaterial," Appl. Phys. Lett. **110**(12), 121108 (2017).
[42]C. Zhang, T. Xue, J. Zhang, L. Liu, J. Xie, G. Wang, J. Yao, W. Zhu, and X. Ye, "Terahertz toroidal metasurface biosensor for sensitive distinction of lung cancer cells," Nanophotonics **11**(1), 101–109 (2022).
[43]F. Cui, X. Huang, Q. Zhou, Y. Tong, F. Liu, Y. Tang, X. Meng, C. Li, B. Fang, and X. Jing, "Magnetic toroidal dipole resonance terahertz wave biosensor based on all-silicon metasurface," Opt. Lasers Eng. **177**(October 2023), 108128 (2024).
[44]E. D. Palik, *Handbook of Optical Constants of Solids* (Academic Press, 2012), Vol. 1, pp. 1–804.
[45]S. Han, L. Cong, Y. K. Srivastava, B. Qiang, M. V. Rybin, A. Kumar, R. Jain, W. X. Lim, V. G. Achanta, S. S. Prabhu, Q. J. Wang, Y. S. Kivshar, and R. Singh, "All-dielectric active terahertz photonics driven by bound states in the continuum," Adv. Mater. **31**(37), 1901921 (2019).
[46]W. Liu and Y. S. Kivshar, "Multipolar interference effects in nanophotonics," Philos. Trans. R. Soc. A **375**(2090), 20160317 (2017).
[47]L. Cong, and R. Singh, "Symmetry-protected dual bound states in the continuum in metamaterials," Adv. Opt. Mater. **7**(13), 1–7 (2019).
[48]P. Agarwal, K. Kishor, and R. K. Sinha, "All dielectric high Q-factor metasurface refractive index sensor based on dual Fano resonances in THz region," IEEE Sensors Lett. **8**(5), 1–4 (2024).
[49]Z. Sadrieva, K. Frizyuk, M. Petrov, Y. Kivshar, and A. Bogdanov, "Multipolar origin of bound states in the continuum," Phys. Rev. B **100**(11), 1–12 (2019).
[50]A. B. Evlyukhin and B. N. Chichkov, "Multipole decompositions for directional light scattering," Phys. Rev. B **100**(12), 1–9 (2019).
[51]X. Liu, C. Lan, K. Bi, B. Li, Q. Zhao, and J. Zhou, "Dual band metamaterial perfect absorber based on Mie resonances," Appl. Phys. Lett. **109**(6), 062902 (2016).
[52]P. Upender and A. Kumar, "THz dielectric metamaterial sensor with high Q for biosensing applications," IEEE Sens. J. **23**(6), 5737–5744 (2023).
[53]S. Banerjee, P. Dutta, A. V. Jha, B. Appasani, and M. S. Khan, "A biomedical sensor for detection of cancer cells based on terahertz metamaterial absorber," IEEE Sensors Lett. **6**(6), 1–4 (2022).
[54]E. Sinchenko, W. E. K. Gibbs, A. P. Mazzolini, and P. R. Stoddart, "The effect of the cladding refractive index on an optical fiber evanescent-wave sensor," J. Light. Technol. **31**(20), 3251–3257 (2013).
[55]P. Agarwal, K. Kishor, and R. K. Sinha, "Ultrasensitive dual-band terahertz metasurface sensor based on all InSb resonator," Opt. Commun. **522**, 128667 (2022).
[56]P. Y. Liu, L. K. Chin, W. Ser, H. F. Chen, C. M. Hsieh, C. H. Lee, K. B. Sung, T. C. Ayi, P. H. Yap, B. Liedberg, K. Wang, T. Bourouina, and Y. Leprince-Wang, "Cell refractive index for cell biology and disease diagnosis: Past, present and future," Lab Chip **16**(4), 634–644 (2016).
[57]X. Chen and W. Fan, "Toroidal metasurfaces integrated with microfluidic for terahertz refractive index sensing," J. Phys. D: Appl. Phys. **52**(48), 485104 (2019).
[58]X. Chen and W. Fan, "Ultrahigh-Q toroidal dipole resonance in all-dielectric metamaterials for terahertz sensing," Opt. Lett. **44**(23), 5876 (2019).

20 May 2025 07:24:49


© Author(s) 2025

2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3315-2169-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/IATMSI64286.2025.10984921

# Unpaved Road Damage Detection via Semantic Segmentation: Enabling Smarter ADAS

Ruchi Tripathi
*Department of ECE*
*Delhi Technological University*
Delhi
India
ruchi 2k20phdec514@dtu.ac.in

S.Indu
*Department of ECE*
*Delhi Technological University*
Delhi
India
s.indu@dtu.ac.in

Rohit Kumar
*Department of ECE*
*Delhi Technological University*
Delhi
India
rohit.kumar@dtu.ac.in

***Abstract*—The visual assessment of damages on unpaved roads, which lack the structured features of paved roads, is one of the primary issues facing the Advanced Driver Assistance System (ADAS). Most existing solutions focus on well-maintained roads with high-quality data and neglect the vehicles operating on unpaved or unstructured roads. This paper proposes a novel deep-learning approach for damage detection on unpaved roads using a UNet architecture with a ResNet-18 backbone. The encoder is ResNet-18, a lightweight network that uses residual connections to capture hierarchical information and alleviate the vanishing gradient issue effectively. Experiments are conducted on the Road Transverse Knowledge (RTK) dataset, which contains low-resolution images and challenging road conditions. The experiments are performed using three architectures—U-Net, LinkNet, and Feature Pyramid Network (FPN)—each tested with ResNet-18, InceptionV3, and VGG16 as the backbone. The results demonstrate that the UNet with ResNet 18 backbone achieves an IoU value of 73.92(%), and F1 Score value of 82.98(%). The U-Net with ResNet-18 backbone model outperforms other state-of-the-art (SOTA) models in segmenting damages on unpaved roads.**

***Index Terms*—Deep learning, UNet, Unpaved Road Segmentation, and Advanced Driver Assistant System.**

## I. Introduction

In the context of Advanced Driver Assistance Systems (ADAS), segmenting damages on unpaved roads is crucial since it has a direct effect on driving efficiency and vehicle safety. It is more difficult to detect and segment damages on unpaved roads due to their uneven surfaces, varied textures, and fluctuating environmental conditions. Accurate segmentation of potholes, fractures, and erosions is necessary for adaptive driving reactions, real-time notifications, and navigational adjustments on unpaved roads. Traditional manual inspection methods are labor-intensive and prone to human error [1]. In recent years, automatic damage detection on unpaved roads has received far too much attention. With the introduction of autonomous driving and ADAS, it is now essential to promptly assess road conditions in order to guarantee the safety of both vehicles and other road users. By combining powerful machine learning and deep learning models designed specifically for uneven road conditions, ADAS can enhance overall driving comfort, save vehicle wear and tear, and promote road safety in rural and underdeveloped areas. Automated damage detection systems, leveraging advanced technologies such as deep learning and computer vision, offer a practical alternative to manual methods. These technologies are particularly pertinent to Advanced Driver Assistance technologies (ADAS), which rely on precise and up-to-date road condition information to make crucial decisions like modifying vehicle behavior or alerting drivers to possible dangers [2, 3]. Even though road damage identification has advanced significantly, the majority of research has concentrated on paved roads because of their predictable surface conditions and the availability of high-quality data [4, 5]. Therefore, there is still a large research gap in the creation of reliable models for unpaved roads, which frequently have imperfections, different surface textures, and intricate environmental conditions. [6].

Recent developments in deep learning have shown promising results in addressing these issues, especially in semantic segmentation. Although models like U-Net and its variations have demonstrated cutting-edge performance in identifying road imperfections, such as cracks and potholes, their use on unpaved roads continues to be challenging because of the intricacy and unpredictability of these surfaces [7, 8]. Furthermore, many current approaches ignore the larger context of unpaved road conditions in favor of concentrating on specific damage types [9].

To overcome these constraints, this study suggests a strong deep learning-based method for identifying surface deterioration on unpaved roads. The proposed approach uses a U-Net architecture with a ResNet-18 backbone to recognize a variety of surface defects, such as erosion, cracks, and ruts, by utilizing semantic segmentation capabilities. The RTK dataset is used to refine this architecture, which yields highly annotated images that are specific to unpaved road conditions [3]. The model provides a workable approach for incorporation into ADAS applications by taking into consideration the particular difficulties presented by unpaved road surfaces, resulting in high accuracy and generalization.

## II. Related Work

Semantic segmentation was first achieved by the fully convolutional network (FCN) [10], which has subsequently accelerated the development of the convolution-based approach. Because it was an early development, it lacked an appropriate method for utilizing the global scene category information. A pyramid

979-8-3315-2169-1/25/$31.00 © 2025 IEEE
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 04:59:35 UTC from IEEE Xplore. Restrictions apply.

pooling module was presented by Zhao et al.[11] to aggregate various regional contexts and enhance the capacity of global context information. Chen et al.[12] introduced atrous spatial pyramid pooling, which improved learning capacity for numerous scale characteristics by varying receptive fields and atrous rates. A reliable CNN-based technique for detecting road damage on highways was published by Wang et al. [13], whereas Xie et al. [14], who focused on model optimization for mobile applications, provided an EfficientNet-based method for pavement crack segmentation.

Using machine learning approaches, Dhiman and Klette [5] have presented two computer vision-based algorithms for damage detection. Similar to this, Fan *et al.* [7] added graph attention layers to enhance pothole detection semantic segmentation performance. With deep learning, Huang *et al.* [6] suggested a system that uses modest RGB and depth sensors to automatically evaluate road conditions. They demonstrate the potential of sensor fusion approaches by utilizing heterogeneous data fusion to detect and measure potholes in an efficient manner. Ruchi *et al.* [15] proposed a modified YOLOv4 model for detecting potholes on a publicly available dataset.A 3D imaging method was proposed by Bosurgi *et al.* [3] for automatic pothole detection. To identify road damage, they have provided high-quality data. An Enhanced Residual Convolutional U-Net-based model for pothole segmentation was proposed by Tripathi et al. [16] for Advanced Driver-Assistance Systems (ADAS) to improve driving safety. In order to increase segmentation accuracy, they implemented a region-based IoU loss and new residual convolutional blocks with dropout to lessen overfitting.

Ma *et al.* [9] reviewed the difficulty of detecting damages on unpaved roads in previous research, they emphasized the state-of-the-art techniques for road imaging and damage identification. Their effort focuses primarily on existing unpaved damage-detection techniques. The RoadFormer model, developed by Li *et al.* [8], uses a transformer-based architecture for RGB-normal semantic segmentation in road scene parsing and performs well under a variety of road conditions. Despite these developments, most approaches are applied to developed areas with well-maintained roadways, ignoring the difficulties presented by unpaved roads.

### A. Limitations in Existing Studies

Traditional techniques are unable to detect damages on unpaved roads, while many deep learning-based techniques need high-quality data in order to detect damages. Furthermore, a lot of current models are unable to generalize their performance for low-resolution images and the worst road conditions, which emphasizes the necessity of methods designed especially for unpaved areas.

## III. Proposed Methodology and Model Architecture

The reliable detection of road damages, particularly on unpaved roads, is a critical task for enhancing Advanced Driver Assistance Systems (ADAS). Unpaved roads pose unique challenges such as varying textures, irregular lighting, and a wide range of surface damage types, including cracks and potholes. To address these challenges, we evaluated three state-of-the-art semantic segmentation architectures: U-Net, LinkNet, and Feature Pyramid Network (FPN), each with ResNet-18, InceptionV3, and VGG16 backbones. Based on the experimental results, U-Net with ResNet-18 was identified as the most effective model for the task.

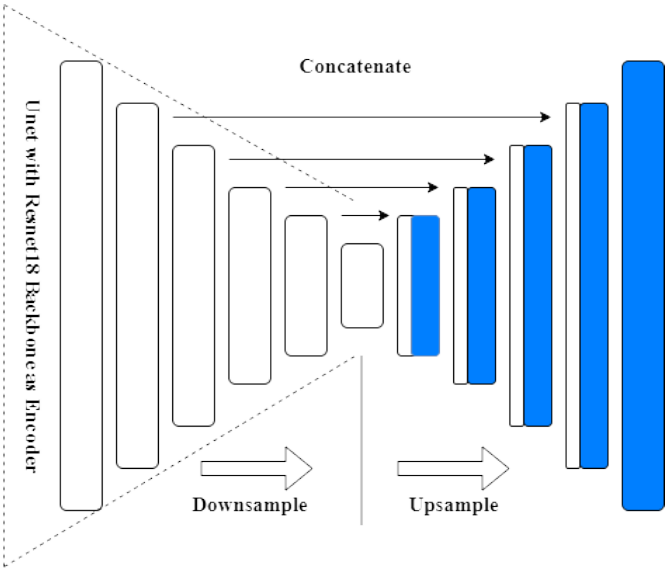



Fig. 1: Overall architecture of our proposed network

### A. Proposed Model: U-Net with ResNet-18

The U-Net architecture was selected as the backbone of our proposed model due to its ability to effectively capture both global context and fine s patial d etails t hrough its encoder-decoder structure with skip connections. By integrating ResNet-18 as the encoder, the network benefits from residual connections, which alleviate the vanishing gradient problem and enable deeper feature extraction [17, 18]. The decoder complements the encoder by progressively upsampling the feature maps while concatenating features from corresponding encoder layers. This combination allows the network to maintain spatial resolution while learning complex patterns necessary for accurate segmentation of road damages, such as cracks and potholes. Fig. 1 represents the overall architecture of the proposed model, comprising UNet with ResNet18 backbone as an encoder part.

The overall methodology of our work is shown in Fig. 2. The input RGB images with corresponding ground truth images are first preprocessed according to the backbone. After preprocessing, extensive experiments have been done to finalize the best-suited backbone for the unpaved road segmentation, and then the model is evaluated based on segmentation metrics: IoU and F1 score. The model is trained for 150 epochs for each segmentation architecture with different backbones.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 04:59:35 UTC from IEEE Xplore. Restrictions apply.

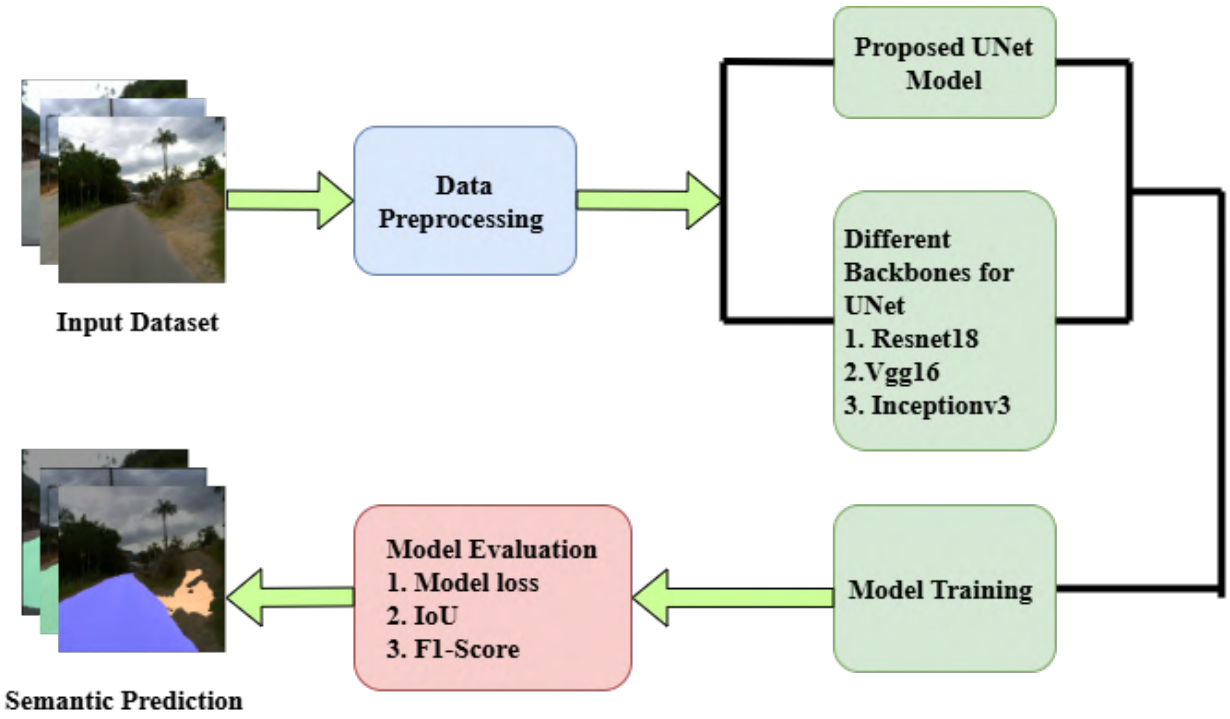



Fig. 2: Proposed Methodology

## IV. Experimental Framework

This section ensures the reproducibility of experiments and provides clarity on the settings, configurations, and resources used during the proposed work.

### A. Loss Function and Evaluation Metrics

All the models are trained using hybrid loss, which combines focal loss and dice loss, to maximize the training outcomes. This hybrid loss function promotes the shape consistency of the segmented regions while guaranteeing pixel-by-pixel precision. Intersection over Union (IoU) and F1 Score metrices are used to assess the model's performance over 150 epochs.

### B. Data Processing and Training Setup

The experiments are conducted using the Road Transverse Knowledge (RTK) dataset [19], which comprises 701 images with ground truth images. 561 images are used for training, while 140 images are used for validation. The dataset includes images of unpaved roads taken in a range of environmental circumstances and capturing different kinds of degradation. We scaled the original images to 256 x 256 resolution for training purposes, even though they had a resolution of 352 x 288. The models are trained using the Adam optimizer with a learning rate of $1 \times 10^{-3}$. The proposed architecture is implemented using the Keras framework. Training and evaluation are conducted on Google Colab, utilizing a T4 GPU accelerator with 12 GB of RAM and 12 GB of GDDR5 VRAM.

## V. Result and Analysis

This section evaluates U-Net, LinkNet, and FPN architectures using ResNet-18, InceptionV3, and VGG16 as backbones for unpaved road damage segmentation. The models are evaluated based on IoU and F1 Score, which highlight their segmentation quality and optimization performance.

### A. Qualitative Analysis

The qualitative analysis evaluates the segmentation outputs by visually inspecting predictions overlaid on input images. The U-Net with ResNet-18 backbone demonstrated superior performance, producing accurate and smooth segmentation boundaries for various road damages (e.g., potholes, cracks) on unpaved road.
Overall, U-Net with ResNet-18 offered consistent and reliable results, making it well-suited for real-world applications like Advanced Driver Assistance Systems (ADAS). Fig. 3 provides a visual comparison of segmentation results on the RTK test dataset, demonstrating how different models perform on RTK test Dataset.

### B. Quantitative Analysis

The performance of the proposed models is assessed through quantitative analysis using standard evaluation metrics such as Intersection over Union (IoU) and F1 Score. The experiments are performed using three architectures—U-Net, LinkNet, and Feature Pyramid Network (FPN)—each tested with ResNet-18, InceptionV3, and VGG16 as backbones. Fig. 4 presents the IOU of different architectures, and Fig. 5 presents the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 04:59:35 UTC from IEEE Xplore. Restrictions apply.

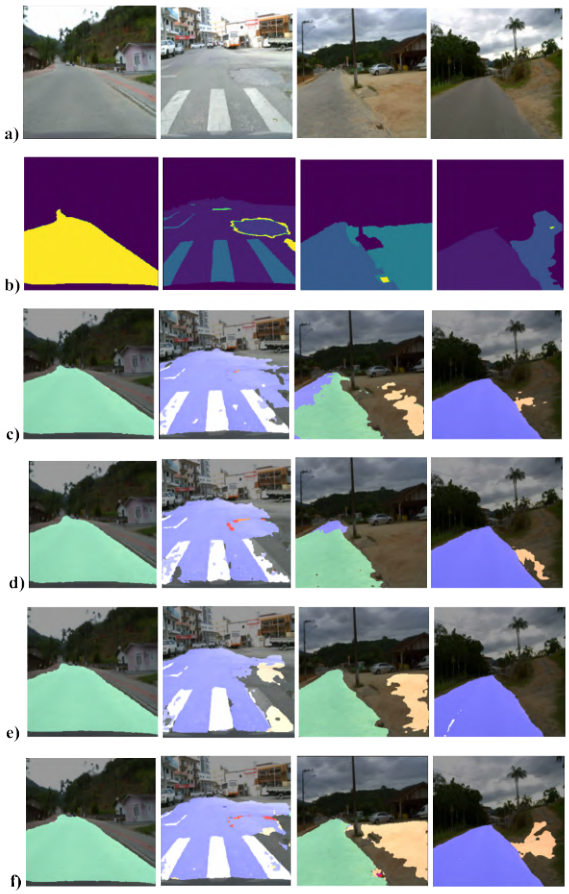



Fig. 3: Sample qualitative demonstrations for the networks on the RTK test dataset. a) RGB Images b)Ground Truth images c) FPN d)LinkNet e)UNet f)UNet-Resnet18

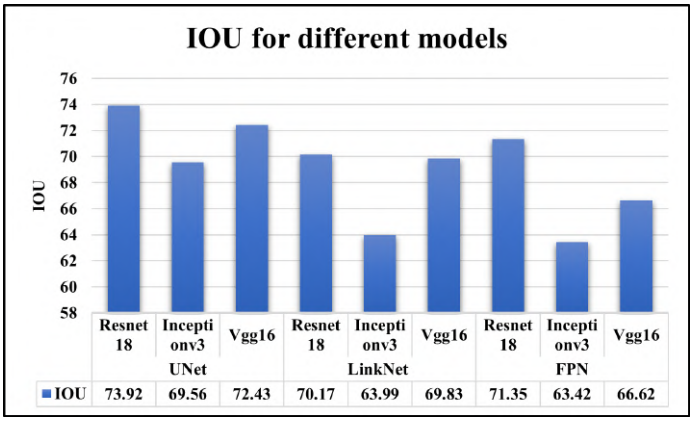



Fig. 4: IoU for different model architectures

F1 scores for different network architectures on the RTK test dataset.

The proposed approach leverages the strengths of U-Net with ResNet-18 to address the specific challenges of unpaved road damage detection, providing a reliable solution for ADAS.

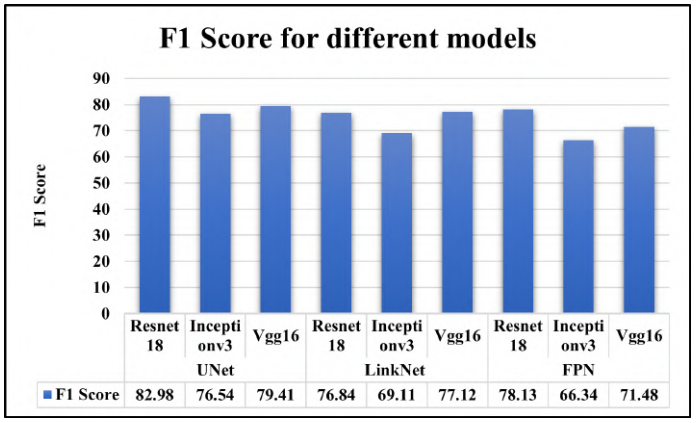



Fig. 5: F1 score for different model architectures

### C. Performance comparison and Suitability for ADAS

Our experimental results demonstrated that the U-Net with ResNet-18 outperformed the other architectures in terms of IOU and F1 Score. The lightweight nature of ResNet-18 ensures that the model is computationally efficient, making it well-suited for real-time deployment in ADAS. This enables robust damage detection on unpaved roads, contributing to safer navigation in challenging environments. Table I represents the performance metrics comparison of different segmentation models based on IOU and F1 Score. Bold values indicate the best performance in respective metrics.

TABLE I: Comparison of segmentation models based on IOU and F1 Score. Bold values indicate the best performance in respective metrics.

| Model | IOU (%) | F1 Score (%) |
|---|---|---|
| PAM-UNet [20] | 65.81 | 79.41 |
| DCWin-Attention [21] | 67.72 | - |
| Improved BiSeNet [22] | 72.63 | **83.21** |
| FPN with ResNet18 | 71.35 | 78.13 |
| LinkNet with ResNet18 | 70.17 | 76.84 |
| **UNet with ResNet18** | **73.92** | 82.98 |

## VI. Conclusion and Future Scope

This work proposes a semantic segmentation approach for unpaved road damage detection using U-Net with a ResNet-18 backbone. The model achieved superior performance in detecting damages on unpaved roads and outperforms LinkNet and FPN. The results highlight that the model is efficient enough to segment unpaved roads containing dirt or mud, which is helpful for ADAS driving on underdeveloped road. Future work can focus on the diversity of more datasets containing more challenging environments and improving real-time performance for ADAS deployment.

## References

[1] Y. Huan, X. Li, and R. Wang, "Challenges in unpaved road detection and the role of automated systems," *Journal of Transportation Engineering*, vol. 17, no. 2, pp. 53–68, 2023.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 04:59:35 UTC from IEEE Xplore. Restrictions apply.

[2] L. Wang and Y. Zhang, "Road damage detection using convolutional neural networks: A survey," *Computer Vision and Image Understanding*, vol. 200, p. 101945, 2022.
[3] G. Bosurgi, M. Modica, and O. Pellegrino, "An automatic pothole detection algorithm using pavement 3d data," *International Journal of Pavement Engineering*, vol. 24, no. 2, p. 2057978, 2023.
[4] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3d reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.
[5] A. Dhiman and R. Klette, "Pothole detection using computer vision and learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3536–3550, 2019.
[6] Y. Huang, M. Jahanshahi, F. Shen, and T. Mondal, "Deep learning–based autonomous road condition assessment leveraging inexpensive rgb and depth sensors and heterogeneous data fusion: Pothole detection and quantification," *Journal of Transportation Engineering, Part B: Pavements*, vol. 149, no. 2, p. 04023010, 2023.
[7] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE Transactions on Image Processing*, vol. 30, pp. 8144–8154, 2021.
[8] J. Li, Y. Zhan, P. Yun, G. Zhou, Q. Chen, and R. Fan, "Roadformer: Duplex transformer for rgb-normal semantic road scene parsing," *IEEE Transactions on Intelligent Vehicles*, 2024.
[9] N. Ma, J. Fan, W. Wang, J. Wu, Y. Jiang, L. Xie, and R. Fan, "Computer vision for road imaging and pothole detection: a state-of-the-art review of systems and algorithms," *Transportation Safety and Environment*, vol. 4, no. 4, p. tdac026, 2022.
[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.
[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
[13] L. Wang and Z. Chen, "Road damage segmentation using convolutional neural networks on urban highways," *Transportation Research Part C: Emerging Technologies*, vol. 134, p. 103485, 2022.
[14] Q. Xie and Y. Zhou, "Efficientnet-based pavement crack detection for real-time applications," *Automation in Construction*, vol. 113, p. 103125, 2020.
[15] S. Indu, R. Kumar *et al.*, "Detection of road potholes using deep learning based improved yolov4 network," in *2022 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*. IEEE, 2022, pp. 1–6.
[16] R. Tripathi, S. Indu, and R. Kumar, "Ercu-net: segmentation of road potholes using enhanced residual convolutional block based on u-net for adas," *Signal, Image and Video Processing*, pp. 1–10, 2024.
[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.04597*, 2015.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[19] T. Rateke and A. Von Wangenheim, "Road surface detection and differentiation considering surface damages," *Autonomous Robots*, vol. 45, no. 2, pp. 299–312, 2021.
[20] Y. Nie, K. An, X. Chen, L. Zhao, W. Liu, X. Wang, Y. Yu, W. Luo, K. Li, and Z. Zhang, "An improved u-net network for sandy road extraction from remote sensing imagery," *Remote Sensing*, vol. 15, no. 20, p. 4899, 2023.
[21] W. Li, J. Zhang, J. Li, and L. Zhuo, "Unpaved road segmentation of uav imagery via a global vision transformer with dilated cross window self-attention for dynamic map," *The Visual Computer*, pp. 1–19, 2024.
[22] C. Bai, L. Zhang, L. Gao, L. Peng, P. Li, and L. Yang, "Real-time segmentation algorithm of unstructured road scenes based on improved bisenet," *Journal of Real-Time Image Processing*, vol. 21, no. 3, p. 91, 2024.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on May 14,2025 at 04:59:35 UTC from IEEE Xplore. Restrictions apply.

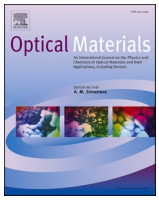
Optical Materials 165 (2025) 117158



Contents lists available at ScienceDirect

Optical Materials

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



Research Article

# Up-conversion luminescence study of thermally stable $Sr_2ZnSi_2O_7$: $Er^{3+}$ phosphors

Seema [a], A.S. Rao [a,b,*], Sarita Baghel [a]

[a] Department of Applied Physics, Delhi Technological University, Bawana Road, New Delhi 110 042, India
[b] Department of Physics, Vikrama Simhapuri University, Nellore 524 324, Andhra Pradesh, India

ARTICLE INFO

*Keywords:*
$Er^{3+}$
Up-conversion
Inokuti-hiryama plot
Temperature dependent PL

ABSTRACT

Current research on rare-earth-doped phosphors often faces challenges related to thermal stability, energy transfer efficiency, and color purity, which limit their practical application in lighting technologies. In this study, we address these issues by synthesizing a series of thermally stable $Sr_2ZnSi_2O_7$: $Er^{3+}$ (SZSi: $Er^{3+}$) phosphors via the high-temperature solid-state reaction method. X-ray diffraction (XRD) confirmed the formation of a tetragonal crystalline phase (space group: P-421m) however, scanning electron microscopy (SEM) revealed comprehensive surface morphology and particle size distribution of irregularly shaped particles. Diffuse reflectance spectroscopy (DRS) was used to calculate the optical band gap of the synthesized phosphors. The photoluminescence (PL) studies demonstrated efficient near-ultraviolet (n-UV) excitation ($\lambda_{ex}$ = 378 nm). Energy transfer analysis using the Dexter theory and the Inokuti-Hirayama (I–H) model indicated that dipole-dipole interactions dominate the energy transfer process between $Er^{3+}$-$Er^{3+}$ ions. The optimized SZSi: $Er^{3+}$ phosphor exhibited high color purity (96 %) with Commission Internationale de l'Eclairage (CIE) chromaticity coordinates of (0.3279, 0.6651) under $\lambda_{ex}$ = 378 nm. The upconversion luminescence intensity at 661 nm [$^4F_{9/2} \rightarrow {}^4I_{15/2}$ ($Er^{3+}$)], when excited at 980 nm, was enhanced with a higher magnitude when increasing the doping concentration from 1 to 10 mol% of $Er^{3+}$ ions. The dependence of the laser pump power concerning the upconversion luminescence intensity depicts that the emission at 661 nm is due to a two-photon absorption process. In addition, under 980 nm, visible upconversion emissions (in the green and red regions) were observed, attributed to a two-photon absorption mechanism involving intermediate energy levels of $Er^{3+}$ ions. Thermal quenching analysis demonstrated a moderate decrease in emission intensity (~24 % at 100 °C and ~36 % at 150 °C), suggesting reasonable thermal stability. These findings underscore the potential of $Er^{3+}$-activated $Sr_2ZnSi_2O_7$ phosphors in non-display photonic applications, such as bioimaging, anti-counterfeiting, or infrared-pumped display technologies.

## 1. Introduction

Phosphor materials play a crucial role in various advanced applications, including fluorescent lighting, plasma display panels, medical instrumentation, high-energy radiation detectors, and LEDs [1–4]. Among these, phosphor-converted w-LEDs (pc-w-LEDs) have garnered significant research interest due to their superior luminescent properties, characterized by extended operational lifespan, high energy efficiency, versatility, enhanced image resolution, and environmental sustainability [3,5]. Conventional pc-w-LEDs are typically fabricated by coupling blue-emitting semiconductor chips with commercially available yellow phosphors, such as $Y_3Al_5O_{12}$:$Ce^{3+}$ (YAG: $Ce^{3+}$) [6,7]. However, the absence of a red spectral component in these systems result in a high correlated color temperature (CCT) and a suboptimal color rendering index (CRI), thereby limiting their performance in high-fidelity lighting applications [8]. The characteristics of an ideal white light include low CCT values and high CRI [9,10]. Thus, combining RGB phosphors on a near-ultraviolet (n-UV) device can enable w-LED synthesis with a high CRI and a low CCT [11].

Among the many rare earth (RE) ions, $Er^{3+}$ is a remarkable prospect for a wide range of device applications because of its wide variety of electronic energy states. Accordingly, the electronic transitions $^4F_{9/2} \rightarrow {}^4I_{15/2}$, $^2H_{11/2} \rightarrow {}^4I_{15/2}$, and $^4S_{3/2} \rightarrow {}^4I_{15/2}$ enable $Er^{3+}$ ions to emit green light effectively [12,13]. Its strong photoluminescence (PL) and optical

* Corresponding author.
*E-mail addresses:* drsrallam@gmail.com, drsrallam@dec.ac.in (A.S. Rao).

https://doi.org/10.1016/j.optmat.2025.117158
Received 1 February 2025; Received in revised form 12 May 2025; Accepted 13 May 2025
Available online 14 May 2025
0925-3467/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

absorption in the near-UV (n-UV) range also contribute to its long lifespan. When paired with the specific host, these properties of RE ions can be used for high-sensitivity optical thermometry, solid-state illumination, solar cells, triple display, undersea communication, anti-counterfeiting, and fingerprint detection [14,15].

Furthermore, several luminescent materials with different phosphor hosts have attracted a lot of interest recently because of their special qualities [16–19]. Silicates represent a crucial class of host materials for the development of luminescent composites owing to their advantageous assets, including facile synthesis, structural versatility, superior physicochemical stability, resistance to moisture, inertness to atmospheric conditions, and typically strong absorption in the n-UV region [13,20]. Silicates are widely used due to their low cost of raw ingredients and simple manufacturing process [21].

Using the commonly employed solid-state reaction approach, the $Sr_2ZnSi_2O_7$: $Er^{3+}$ (SZSi: $Er^{3+}$) phosphors were synthesized in the present research work, and their phase purity and morphology were assessed. The energy transfer (ET) process and the host material's degradation patterns have been the subjects of several investigations. Based on the emission data, the CIE coordinates of SZSi: $Er^{3+}$ phosphors were also examined. When exposed to n-UV excitations, the SZSi: $Er^{3+}$ phosphors exhibit green emission, indicating that it is possible to use the as-synthesized phosphor as a solid-state green emitter to produce w-LEDs.

In the fields of solar cells, temperature probes, photocatalysis, medical diagnosis and display, and combating counterfeiting technologies [12], the upconversion (UC) luminescence of $RE^{3+}$ ions has drawn a lot of attention since Auzel discovered it in the mid-1960s [22]. These uses of UC materials are, nevertheless, extremely restricted. Numerous 4 $f^N$ electronic states in $RE^{3+}$ ions result in different photon transition paths and diverse color emissions [23]. As part of our research, we upconverted the host lattice and saw the effect of concentration on the PL intensity. Temperature and power-dependent PL investigations were also conducted to provide a good picture of PL intensity variations with temperature and power.

## 2. Synthesis method and instrumentation

To synthesize a series of SZSi: $Er^{3+}$ phosphors, the traditional solid-state reaction approach was employed. Commercially available sources provided the initial components, which included $SrCO_3$, ZnO, $SiO_2$, and $Er_2O_3$ with high color purity. After weighing the stoichiometric reactants on an electronic scale, the chemicals were thoroughly mixed and ground in an agate mortar with acetone acting as a wetting agent for 1 h. A programmable electric muffle furnace was then used to heat the homogeneous mixture for 5 h at 1250 °C at a rate of 5 °C per minute while placed within an alumina crucible. Additional grinding was done on the finished products for further analysis after they had cooled to room temperature. The procedures implemented during the synthesis process are illustrated in the flowchart presented in Fig. 1.

A comprehensive suite of characterization techniques was employed to investigate the structural and optical properties of the synthesized phosphor samples. The phase composition and crystallinity were analyzed using XRD patterns, recorded on a Bruker D8 Advance diffractometer with a precision of ±0.01, using Cu-$K_\alpha$ radiation (with λ = 1.54 Å) as a source with nickel monochromator. The surface morphology of the samples was examined using a JEOL JSM-6490 SEM

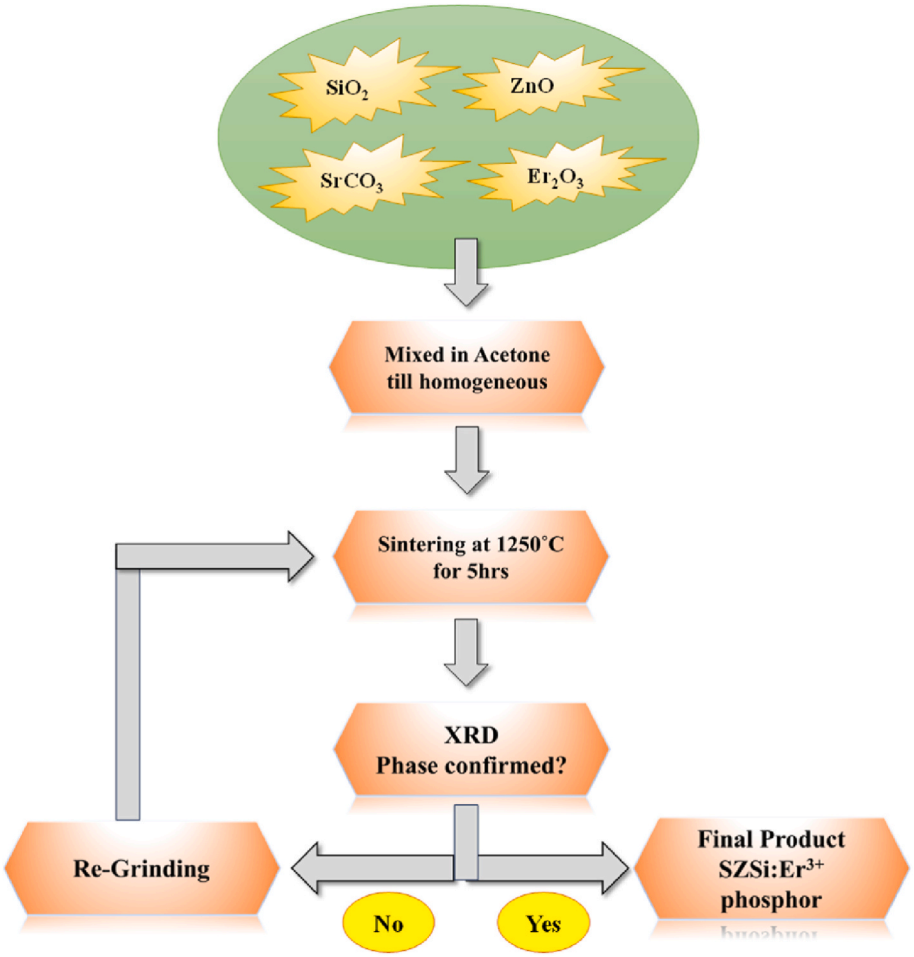



**Fig. 1.** Synthesis procedure for SZSi: $Er_x$ ($x$ = 1, 2, 3, 5, and 10 mol%) phosphors.

operated at an accelerating voltage of 10 kV. Using a JASCO V-770 spectrophotometer, which has a resolution of ±0.1 nm and works in the visible and ultraviolet spectrum with halogen and deuterium light sources, the DRS analysis was done. Down-conversion PL properties were assessed using a JASCO Spectrofluorometer FP-8300 (±1.0 nm resolution), and UC spectra were recorded using a 980 nm continuous-wave (CW) diode laser paired with a fast CCD spectrometer (ULS2048x64, Avantes, USA). Utilizing the Edinburgh FL 920 Fluorescence Spectrometer, the PL decay spectra were examined. The standard room temperature was used for all measurements.

For temperature-dependent UC emission investigations, a thermocouple placed near the laser spot on the sample surface was used, and a CCD spectrometer was used to record the spectra.

## 3. Results and discussion

### 3.1. X-ray diffraction (XRD): structural analysis

To evaluate the crystalline composition and phase purity of the as-synthesized samples, an XRD analysis was performed. The XRD patterns of SZSi: $Er_x$ ($x = 0,1, 2, 3, 5$, and 10 mol%) phosphors, presented in Fig. 2 (a), were compared with the reference data from the Joint Committee on Powder Diffraction Standards (JCPDS) database over the $2\theta$ range of 20°–70°. The lattice parameters for the tetragonal structure of SZSi were determined and found to be well matched with the standard JCPDS card no. 00-039-0235 with space group: P-421m. The absence of significant structural modifications or impurity phases upon $Er^{3+}$ doping suggests the successful substitution of $Sr^{2+}$ lattice sites by $Er^{3+}$ activator ions without disrupting the host crystal structure.

SZSi: $Er_x$ ($x = 0,1, 2, 3, 5$, and 10 mol%) phosphors' crystallite size ($D$) was calculated using the Williamson-Hall (W–H) plot formalism [24, 25]:

$$\beta \cos\theta = \frac{k\lambda}{D} + 4 \in \sin\theta \quad (1)$$

Here, $D$ indicates the mean crystallite size, and K is the structural component, which is equal to 0.94. '$2\theta$' is the diffraction angle, while $\beta$ is the full-width half maximum (FWHM) for the major diffraction peaks. The wavelength of the Cu-$K_\alpha$ source is denoted by '$\lambda$' while '$\in$' stands for the lattice micro-strain. The values of '$D$' for the un-doped and SZSi: $Er_x$ ($x = 1, 2, 3, 5$, and 10 mol%) phosphor samples may be determined by plotting a graph (presented as a straight line) between $\beta cos\theta$ and $4 sin\theta$, as seen in Fig. 2 (b). All of the samples' diffraction peaks are accurately indexed to those found in the standard JCPDS database, and the absence of any impurity-related peaks suggesting that RE ions have nearly no effect on the crystal structure. Strong and accurate diffraction peaks show that the prepared phosphors have a higher crystalline structure, which is required for comparatively superior luminosity.

The SZSi: $Er_x$ (x = 3 mol%) phosphor sample undergoes Rietveld refinement using the FullProf Suite programme to be more thoroughly investigated structurally, as shown in Fig. 2 (c). The backgrounds were improved using fifteen cycles of linear interpolation, and for each composition, the pseudo-Voigt function was employed to fit the experimental data. The black line in the refined map 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

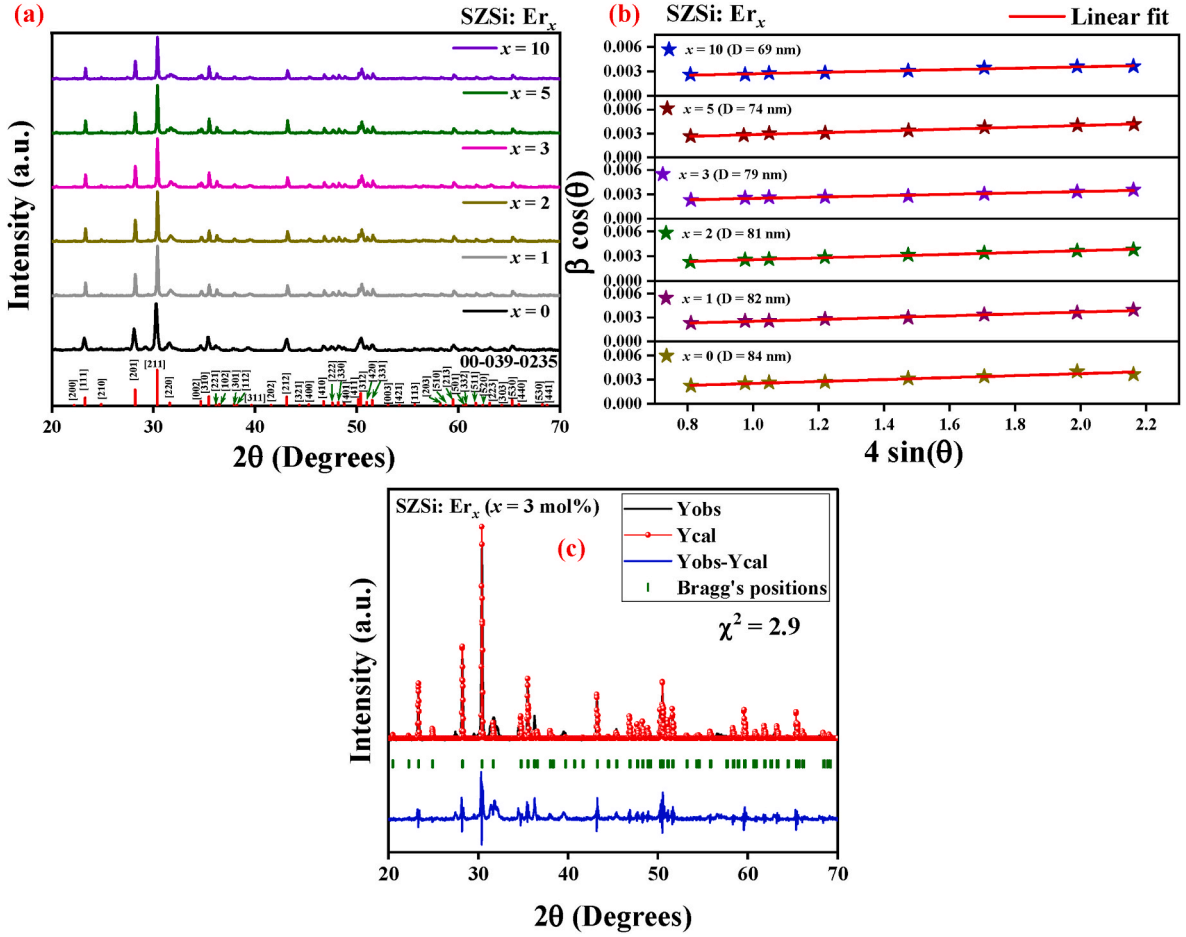



**Fig. 2.** **(a)** The XRD patterns, along with the standard JCPDS card no. 00-039-0235, **(b)** W–H plot for SZSi: $Er_x$ ($x = 0,1, 2, 3, 5$ and 10 mol%) phosphors and **(c)** Rietveld refinement for SZSi: $Er_x$ ($x = 3$ mol%) phosphor.

and the observed intensities, and the green perpendicular line symbolizes the positions of Bragg's peaks. The goodness value comes out to be 2.9. Table 1 lists all the related refined parameters for SZSi: $Er_x$ ($x = 3$ mol%) phosphor.

### 3.2. SEM and EDAX: morphological features

Phosphor particle shape affects both the decay kinetics and the PL emission intensity. SEM was used to further examine the prepared phosphor samples' microscopic structure and particle size. Fig. 3 (a) shows the SEM image and Fig. 3 (b) shows the energy-dispersive X-ray analysis (EDAX) plot of the as-prepared SZSi: $Er_x$ ($x = 3$ mol%) phosphor. Several irregular polyhedron particles, each with an average size of a few microns, make up the powder that lies underneath the microscopic morphology, which is consistent with the results obtained in literature [26,27].

The presence of each element in the as-synthesized phosphor is confirmed by the peaks in the EDAX spectrum that correspond to, say, Sr, Si, Zn, and Er. Additionally, some phosphor particles show very little agglomeration, which may be due to high-temperature reaction processes and manual grinding, and smaller particles may have a larger ratio of surface to volume than bulk particles [28]. SEM images revealed that the $Er^{3+}$-doped phosphor materials exhibit particle sizes in the micrometer (μm) range, making them highly suitable for application as coatings in LEDs and electronic display systems [29–31].

### 3.3. DRS and optical band gap ($E_g$) studies

The optical band gap ($E_g$) is a critical parameter for phosphors utilized in w-LEDs. DRS measurements were conducted on SZSi: $Er_x$ ($x = 1$, 2, 3, 5, and 10 mol%) phosphors over the wavelength range of 230–1200 nm, with the results presented in Fig. 4 (a). Six transitions were obtained with corresponding wavelengths 370, 492, 525, 655, 800, and 981 nm are from ground state $^4I_{15/2}$ to excited states $^4G_{11/2}$, $^4F_{7/2}$, $^2H_{11/2}$, $^4F_{9/2}$, $^4I_{9/2}$, and $^4I_{11/2}$, respectively [32]. By converting the DRS data into the Kubelka-Munk (K-M) function using the below equation, the optical band gap was evaluated [33,34]:

$$F(R) = \frac{(1 - R^2)}{2R} = \frac{\alpha}{S} \tag{2}$$

In this case, 'α' stands for the absorption coefficient, 'R' for the reflectance of the samples, and 'S' for the scattering coefficient. The link between '$E_g$' and 'α' is obtained using Tauc's equation [35]:

$$\alpha h\nu = C(h\nu - E_g)^{n/2} \tag{3}$$

In this context, $h\nu$ represents the photon energy, $C$ is an energy-dependent constant (here equal to 1), and $n$ represents the type of interaction (=1 for direct transitions and = 4 for indirect transitions) [7]. The direct proportionality between the $F(R)$ and $\alpha$ gives the relationship represented by the equation as follows [35,36]:

$$F(R)h\nu = C\,(h\nu - E_g)^{n/2} \tag{4}$$

The band gap for the SZSi: $Er^{3+}$ phosphors was found by graphing $[F(R)h\nu]^2$ versus $h\nu$ to create a Tauc's plot. The linear data is used to extrapolate the line, which yields the optical band gap ($E_g$). Fig. 4 (b) represents the Tauc's plot for all the prepared SZSi: $Er_x$ ($x = 1$, 2, 3, 5, and 10 mol%) phosphors with their bandgap values. The band gap values ($E_g$) decrease with increasing $Er^{3+}$ ion concentration due to the formation of the intermediate energy levels of the $Er^{3+}$ ions in the SZSi host matrix [33].

**Table 1**
Refined parameters for SZSi: $Er_x$ ($x = 3$ mol%) phosphor.

| Sample | JCPDS | SZSi: $Er_x$ ($x = 3$ mol%) |
|---|---|---|
| a = b (Å) | 8.097 | 7.991 |
| c (Å) | 5.240 | 5.162 |
| V (Å$^3$) | 343.619 | 341.470 |
| χ2 | – | 2.9 |
| $R_p$ (%) | – | 32.4 |
| $R_{wp}$ (%) | – | 37.6 |
| $R_{exp}$ (%) | – | 12.4 |

### 3.4. Down-converting optical studies for SZSi: $Er^{3+}$ phosphors

#### 3.4.1. PL spectral analysis

An analysis of the room temperature PL spectrum features reveal information about the photoluminescent properties of SZSi phosphors doped with $Er^{3+}$ ions. The PL excitation spectrum of SZSi: $Er_x$ ($x = 3$ mol %) phosphor in the 340–460 nm wavelength range, captured at 545 nm emission wavelength ($\lambda_{em}$) and is shown in Fig. 5 (a). The excitation spectra include five excitation bands with corresponding wavelengths of 364, 378, 406, 443, and 450 nm that correspond to the transitions from $^4I_{15/2}$ to $^4G_{9/2}$, $^4G_{11/2}$, $^2H_{9/2}$, $^4F_{3/2}$, and $^4F_{5/2}$, respectively [13,37]. The transition $^4I_{15/2} \rightarrow {}^4G_{11/2}$ is represented by the strongest peak among all of these peaks, which is located at 378 nm. This wavelength was thought to be ideal for recording the PL emission spectra [13]. The as-prepared phosphor's PL emission spectra, recorded at an excitation wavelength ($\lambda_{ex}$) of 378 nm in the 520–640 nm range, display green emissions that correspond to the distinctive emissions of $^2H_{11/2} \rightarrow {}^4I_{15/2}$ (524 nm) and $^4S_{3/2} \rightarrow {}^4I_{15/2}$ (545 nm) [13]. In recognition of the $^4S_{3/2} \rightarrow {}^4I_{15/2}$ (545 nm) transition, the green portion of these PL emission spectra showed a strong and prominent emission band. Additionally, the study examined the impact of $Er^{3+}$ doping concentration on luminescence intensity. SZSi: $Er_x$ ($x = 1$, 2, 3, 5, and 10 mol %) phosphors with varying doping concentrations are shown in Fig. 5 (b) with their PL emission spectra. As the $Er^{3+}$ ions concentration increases in SZSi, the emission intensity of the $^4S_{3/2} \rightarrow {}^4I_{15/2}$ transition progressively increases, reaching a maximum at 3 mol% of $Er^{3+}$ ions concentration.

However, as the concentration of $Er^{3+}$ ions further increased, the emission intensity dropped due to the concentration quenching effect, which was caused by energy transfer between $Er^{3+}$ ions, specifically through $Er^{3+}$-$Er^{3+}$ dynamics. The proximity between the nearest $Er^{3+}$ ions gradually decreases as the concentration of $Er^{3+}$ ions rises, hence enhancing the interaction between nearby $Er^{3+}$ ions. As a result, the increased proximity of ions leads to non-radiative energy transfer. Concentration quenching occurs due to interactions such as exchange coupling, radiative reabsorption, and multipole-multipole interactions among the activator ions. According to Van Uitert, these energy transfer processes are directly linked to the aforementioned interactions [38].

The as-synthesized SZSi: $Er^{3+}$ phosphors' energy level diagram is displayed in Fig. 6. It includes mechanisms for energy transmission, emission, excitation, and non-radiative transitions. The goal is to excite the $Er^{3+}$ ions to higher energy levels.

Through multi-phonon relaxation, it relaxes to $^2H_{11/2}$ and $^4S_{3/2}$ levels and emits green light emissions from these states to return to the ground state $^4I_{15/2}$. Using Blasse's equation, researchers calculate the critical distance ($R_c$) to comprehend the interaction between $Er^{3+}$ ions [25,39,40]:

$$R_C = 2\left(\frac{3V}{4\pi X_C N}\right)^{1/3} \tag{5}$$

Here, $X_c$ stands for the ideal concentration, N for the number of dopant sites, and V for the volume of the unit cell. It is found that the critical distance for $Er^{3+}$-doped SZSi phosphors is 21.92 Å. Multipolar interactions will be more vital than exchange contacts between the $Er^{3+}$ ions if the $R_c$ value is greater than 5 Å [36,39]. It is possible to examine the multipolar interactions using the Dexter theory [35,41]:

$$\log\left(\frac{I}{x}\right) = A - s\frac{\log(x)}{d} \tag{6}$$

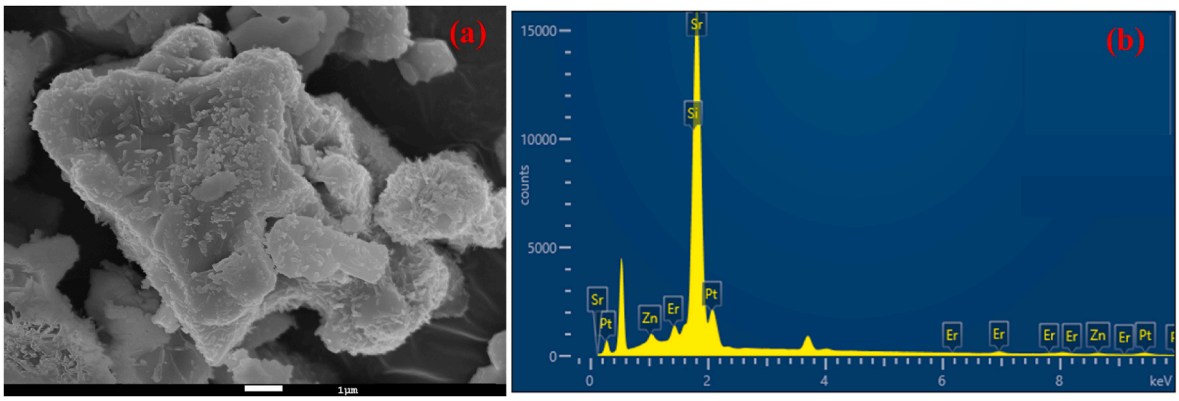



**Fig. 3.** **(a)** SEM image and **(b)** EDAX plot for SZSi: $Er_x$ ($x$ = 3 mol %) phosphor.

here, in the formula, A is a constant, d is dimensionality (=3), $x$ is the concentration of $Er^{3+}$ ions, and $I$ is the emission intensity. Types of interactions are defined by the parameter $s$, where $s = 6$ denotes dipole-dipole interactions, $s = 8$ denotes dipole-quadrupole interactions, and $s = 10$ denotes quadrupole-quadrupole interactions [42]. The correlation between the log(I/x) versus log(x) under $\lambda_{ex} = 378$ nm is plotted in Fig. 7. The '$s$' value is calculated using the slope of the above linear fitted graph and found to be equal to 5.61, which is very near to 6. Therefore, dipole-dipole interactions are thought to be the source of the energy transfer between $Er^{3+}$- $Er^{3+}$ ions in the as-prepared SZSi: $Er^{3+}$ phosphors.

### 3.4.2. Colorimetric analysis

The CIE chromaticity Coordinates can be used to understand the colorimetric performance of SZSi: $Er^{3+}$ phosphors. The coordinates and the PL emission spectra of the SZSi: $Er_x$ ($x$ = 1, 2, 3, 5, and 10 mol %) phosphor samples at $\lambda_{ex} = 378$ nm were computed using the formula given in the literature [43]. These estimated CIE color coordinates fall inside the green region, as shown in Fig. 8. This suggests that the $Er^{3+}$ ion-doped SZSi phosphor is a promising choice for photonic devices functioning in the green range.

The evaluation of phosphor performance also heavily relies on Color Purity (CP), another crucial metric. We can use the following formula to find Color Purity [44]:

$$CP\ \% = \sqrt{\frac{(x - x_{ee})^2 + (y - y_{ee})^2}{(x_d - x_{ee})^2 + (y_d - y_{ee})^2}} \times 100\% \tag{7}$$

The illumination point is represented by ($x_{ee}$, $y_{ee}$), the phosphors' color coordinate is (x, y), and the dominant wavelength position is ($x_d$, $y_d$). The as-prepared SZSi: $Er^{3+}$ phosphors have color purity greater than 93 % under $\lambda_{ex} = 378$ nm. The above calculated parameters, i.e., CIE and CP for each of the as-prepared SZSi: $Er^{3+}$ phosphors are shown in Table 2.

### 3.4.3. Time-resolved photoluminescence (TRPL) analysis

The SZSi: $Er_x$ ($x$ = 1, 2, 3, 5, and 10 mol%) phosphors' green luminescence decay under $\lambda_{ex} = 378$ nm and $\lambda_{em} = 545$ nm is displayed in Fig. 9 at ambient temperature. The decay curves were fitted with several exponential equations; however, the double exponential fitting is a useful evaluation technique. The luminous intensity can be represented using bi-exponential fitting and the following equation [45,46]:

$$I = I_0 + A_1 \exp\left(\frac{-t}{\tau_1}\right) + A_2 \exp\left(\frac{-t}{\tau_2}\right) \tag{8}$$

Here, $\tau_1$ and $\tau_2$ denote the decay times associated with the fast and slow components of luminescence, respectively, and are linked to the corresponding decay constants with amplitudes $A_1$ and $A_2$; $I_0$ and $I$ represent the intensities at time t = 0 and at any time t, respectively.

Utilizing the following formula, the values for the average lifetime ($\tau$) were calculated [24,39]:

$$\tau = \frac{A_1\,\tau_1^2 + A_2\,\tau_2^2}{A_1\,\tau_1 + A_2\,\tau_2} \tag{9}$$

Table 3 displays the measured variation of $\tau$ values, along with the fast and slow decay times for the as-synthesized phosphors with the concentration of $Er^{3+}$ ions. The energy transfer between $Er^{3+}$-$Er^{3+}$ ions that occurs over close proximity is probably the cause of this phenomenon.

Two methods exist by which an excited ion or molecule can return to its ground state [47]. First, energy migrates directly from the donor ion to the acceptor ion. Second, before the energy reaches the acceptor ions, it moves from donor to donor ions [48]. The ET mechanism is accurately represented by the I–H model when ET from donor to the acceptor ions is larger than that among the donor ions. The intensity of PL is provided by Ref. [35]:

$$I_t = I_0 \exp\left\{ -\frac{t}{\tau_0} - Q\left(\frac{t}{\tau_0}\right)^{3/s} \right\} \tag{10}$$

Here, $Q$ represents the energy transfer variable, $I_0$ represents the intensity at $\tau = 0$, and $\tau_0$ represents the intrinsic decay period of donors in the dearth of acceptor ions.

Fig. 10 displays the $Er^{3+}$ (3 mol%) ion's I–H fitted decay profile for the as-synthesized phosphors for $s = 6$, 8, and 10. The data for $s = 6$ is the most appropriate for comprehending the energy transfer process between $Er^{3+}$-$Er^{3+}$ ions and hence representing the dipole-dipole interaction. This analysis confirms that the predominant mechanism responsible for the luminescence quenching in the synthesized phosphors is energy transfer facilitated by dipole-dipole interactions between adjacent $Er^{3+}$ ions. Furthermore, theoretical modelling based on Dexter's formalism corroborates these findings. Additionally, the critical transfer distance ($R_s$) in the I–H model can be estimated using the relationship given below [49]:

$$R_s = \left(\frac{3Q}{4\pi CN}\right)^{1/3} \tag{11}$$

Here, $Q$ = energy transfer parameter (derived from fitting the decay curves to the I–H model), $C$ = concentration of $Er^{3+}$ ions, and $N$ = number density of luminescent centers in the host matrix. The calculated value of $R_s$ for the optimized SZSi: $Er_x$ ($x$ = 3 mol%) is 36.30 Å.

Further, the decrease in lifetime values with the rise in $Er^{3+}$ ions concentration was analyzed using Auzel's model with the equation provided below [36]:

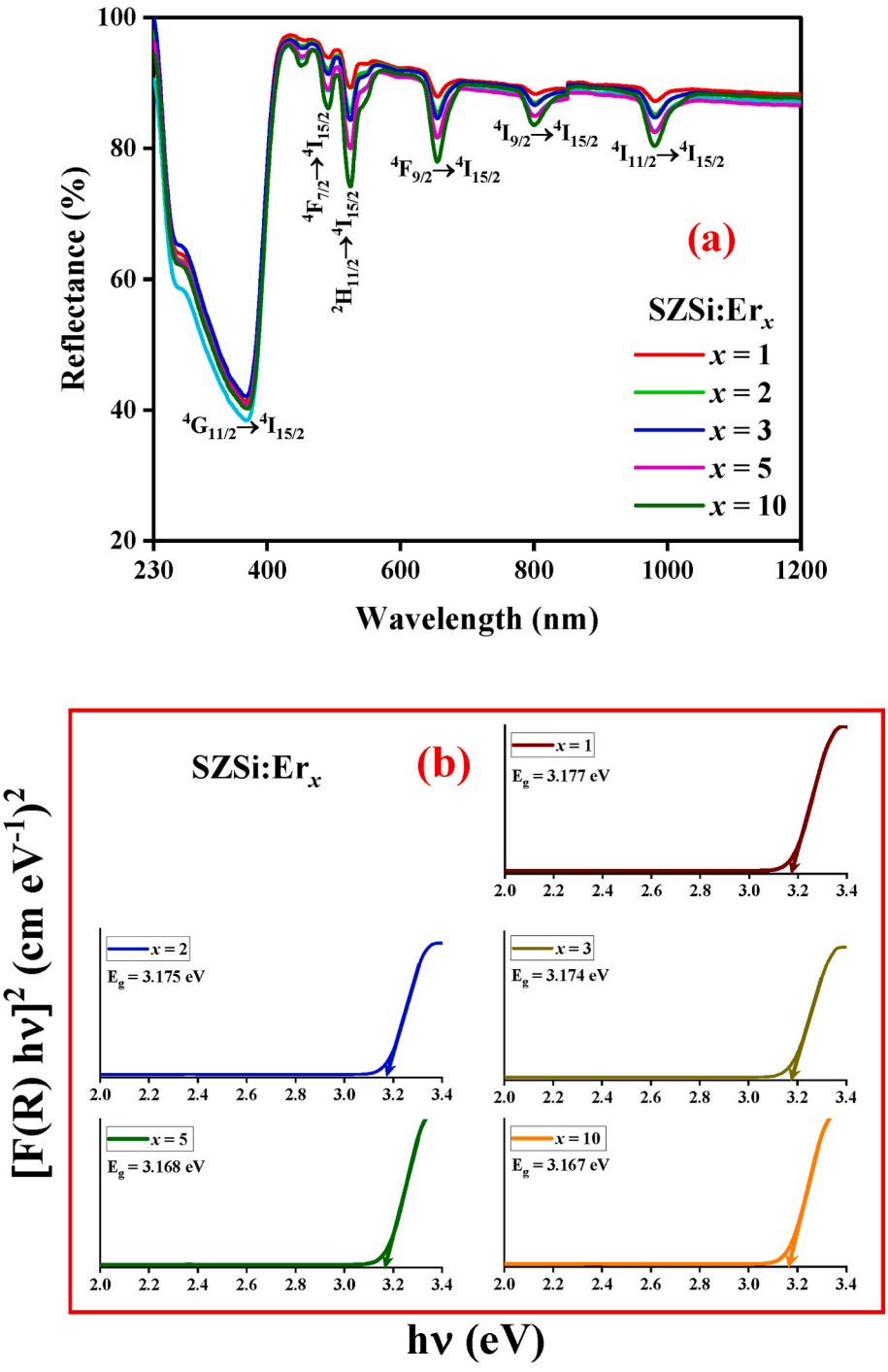



**Fig. 4.** **(a)** DRS and **(b)** Tauc's plot for SZSi: $Er_x$ ($x$ = 1, 2, 3, 5, and 10 mol%) phosphors.

$$\tau = \frac{\tau_0}{1 + \frac{c}{c_0 e^{-N/3}}} \quad (12)$$

Here, $\tau$ and $\tau_0$ are the calculated and intrinsic radiative lifetimes, $c$ refers to $Er^{3+}$ ions concentration, $c_0$ is a constant, and $N$ indicates the number of generated phonons during relaxation. The fitted data using Auzel's model give the value of $\tau_0$, shown in Fig. 10 (b) in SZSi: $Er^{3+}$ phosphors. The quantum efficiency ($\eta$) is calculated using the equation as follows [4] and tabulated in Table 3.

$$\eta = \frac{\tau}{\tau_0} \quad (13)$$

### 3.5. Up-conversion emission studies

#### 3.5.1. Concentration-dependent PL

The UC emission spectrum of phosphors doped with $Er^{3+}$ ions under stimulation at 980 nm has been examined and shown in Fig. 11 (a). The transitions from the energy states $^2H_{11/2}$, $^4S_{3/2}$, and $^4F_{9/2}$ to $^4I_{15/2}$ of $Er^{3+}$ ions are responsible for the weak green (located at 529 and 545 nm) and intense red (located at 661 nm) UC emissions for the as-prepared SZSi: $Er^{3+}$ phosphors [38]. It is clear that the three peaks' intensities rise with the concentration of $Er^{3+}$ ions in the host lattice. Fig. 11 (b) shows the CIE coordinates for SZSi: $Er_x$ ($x$ = 10 mol%) under $\lambda_{ex}$ = 980 nm.

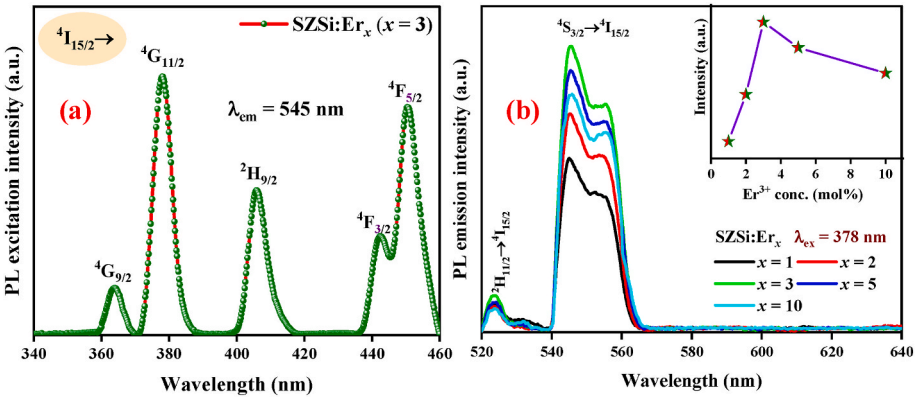



**Fig. 5.** **(a)** The PL excitation spectrum of SZSi: Er$_x$ ($x$ = 3 mol%) phosphor at $\lambda_{em}$ = 545 nm and **(b)** PL emission spectra for SZSi: Er$_x$ ($x$ = 1, 2, 3, 5, and 10 mol %) phosphors at $\lambda_{ex}$ = 378 nm [Inset shows the variation of PL emission intensity with Er$^{3+}$ ions concentration].

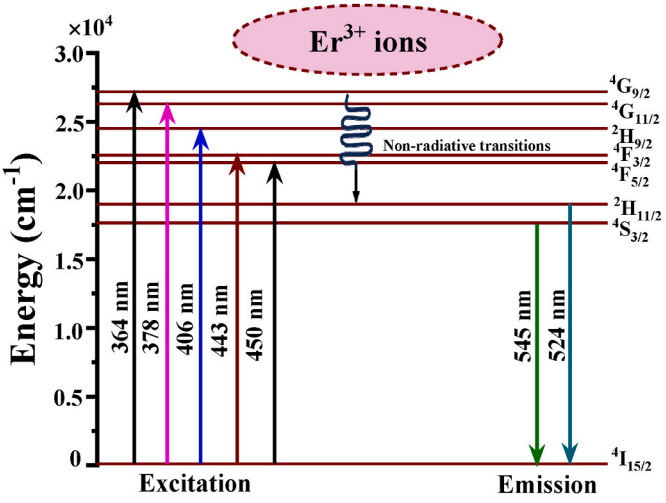



**Fig. 6.** Energy level diagram for Er$^{3+}$ ions in SZSi host lattice.

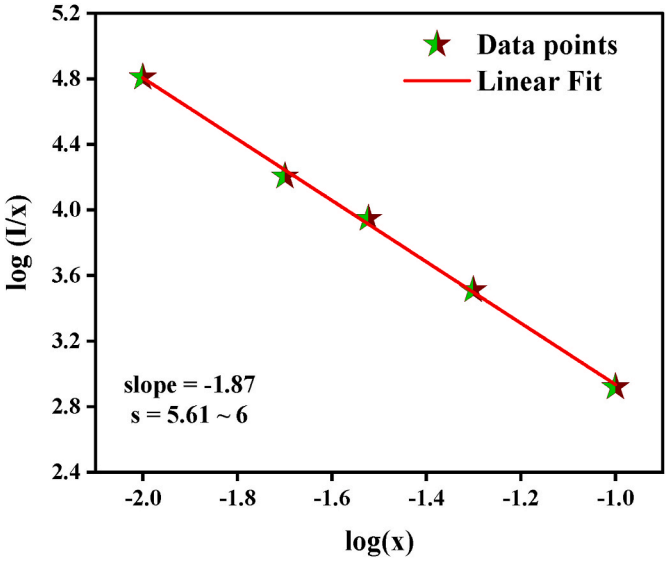



**Fig. 7.** A graph between log(I/x) versus log(x) under $\lambda_{ex}$ = 378 nm.

### 3.5.2. Temperature-dependent (TD) PL

As temperature rises, UV/blue LED chips deteriorate in optical performance. To examine the temperature-sensitive properties of the as-synthesized phosphors, TDPL measurements have been conducted by altering the temperature from RT (27 °C) to 197 °C. Fig. 12 (a) shows the variation of PL emission intensity for SZSi: Er$_x$ ($x$ = 10 mol%) phosphor for the transitions from $^2H_{11/2}$, $^4S_{3/2}$, and $^4F_{9/2}$ to the ground state ($^4I_{15/2}$) corresponding to different temperatures under $\lambda_{ex}$ = 980 nm. Using the most significant integrated emission intensity as 100 %, the inset of Fig. 12 (a) illustrates the percentage changes of the relative intensity with temperature. Due to an increase in non-radiative transitions, the emission peak intensity seems to decrease as the temperature rises [50]. At 150 °C, the PL intensity dropped to 64 % while the emission band's shape stayed the same, suggesting that the as-synthesized phosphors have good thermal stability [29,40]. The following Arrhenius formula is used to calculate the activation energy from the TDPL spectra [35,51,

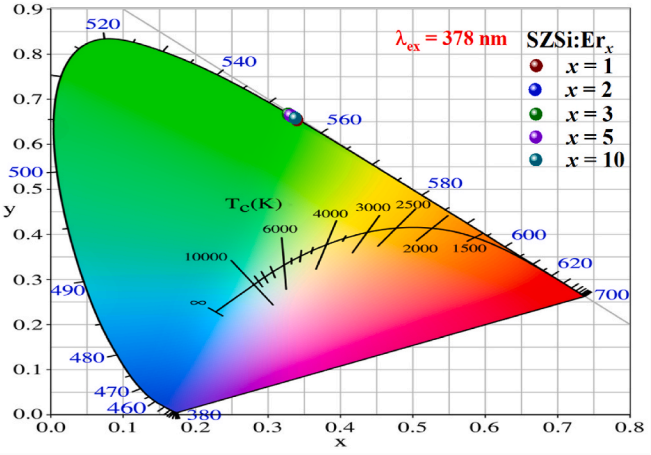



**Fig. 8.** CIE color coordinates for SZSi: Er$_x$ ($x$ = 1, 2, 3, 5, and 10 mol %) phosphor samples at $\lambda_{ex}$ = 378 nm. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

**Table 2**
CIE and CP (%) values for each of the as-prepared SZSi: Er$_x$ ($x$ = 1, 2, 3, 5, and 10 mol %) phosphor samples.

| SZSi: Er$_x$ | CIE coordinates | | CP (%) |
|---|---|---|---|
| | X | Y | |
| $x$ = 1 mol% | 0.3401 | 0.6525 | 93 |
| $x$ = 2 mol% | 0.3323 | 0.6611 | 95 |
| $x$ = 3 mol% | 0.3279 | 0.6651 | 96 |
| $x$ = 5 mol% | 0.3310 | 0.6622 | 95 |
| $x$ = 10 mol% | 0.3376 | 0.6561 | 94 |

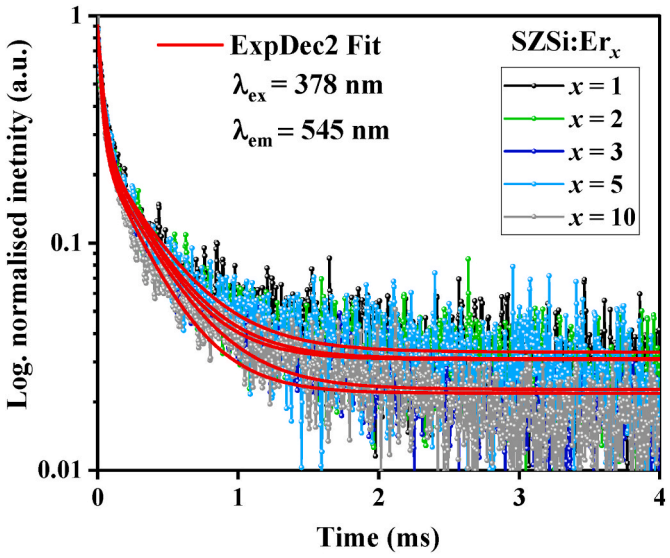



**Fig. 9.** The PL decay for SZSi: $Er_x$ ($x$ = 1, 2, 3, 5, and 10 mol%) phosphors at $\lambda_{ex}$ = 378 nm and $\lambda_{em}$ = 545 nm at ambient temperature.

**Table 3**
Fast & slow decay components along with the Average lifetime values for as-prepared SZSi: $Er_x$ ($x$ = 1, 2, 3, 5, and 10 mol %) phosphor samples.

| SZSi: $Er_x$ | Fast decay component | Slow decay component | Average lifetime (τ) (μs) | Quantum efficiency (η(%)) |
|---|---|---|---|---|
| $x$ = 1 mol% | 0.03801 | 0.36819 | 295 | 99 |
| $x$ = 2 mol% | 0.03547 | 0.34155 | 273 | 92 |
| $x$ = 3 mol% | 0.03514 | 0.32989 | 268 | 90 |
| $x$ = 5 mol% | 0.02739 | 0.30438 | 257 | 86 |
| $x$ = 10 mol% | 0.02680 | 0.29328 | 240 | 81 |

52]:

$$I_T = \frac{I_0}{\left[1 + c\exp\left(-\Delta E/KT\right)\right]} \quad (14)$$

here, K is the Boltzmann constant ($8.67 \times 10^{-5}$ eV/K), c is a constant, $\Delta E$ stands for activation energy, and $I_0$ & $I_T$ are the intensities at 27 °C and any temperature $T$. For as-synthesized phosphors, the value of the slope determined by fitting $\ln[(I_0/I_T)-1]$ vs 1/KT aids in evaluating $\Delta E$, which is equivalent to 0.196 eV (as illustrated in Fig. 12(b)). The as-synthesized phosphors' comparatively high $\Delta E$ value indicates thermal stability, supporting their prospective use in solid-state lighting devices [53,54]. The crossover process causes the emission intensity to decrease as the temperature rises, increasing the likelihood of a non-radiative transition.

### 3.5.3. Power dependent PL studies

To examine the power vulnerability of the as-synthesized phosphors, power dependent PL tests have been conducted for the SZSi: $Er_x$ ($x$ = 10 mol%) phosphor under $\lambda_{ex}$ = 980 nm by varying the power from 331 to 520 m W. The variations in the emission intensity of $^2H_{11/2}$ & $^4S_{3/2} \rightarrow {}^4I_{15/2}$ and $^4F_{9/2} \rightarrow {}^4I_{15/2}$ transitions of $Er^{3+}$ ions corresponding to the varying power are plotted in Fig. 13 (a). It is obvious from the above-mentioned figure that the intensities of the three peaks increase with increasing power. The value of the slope evaluated by the fitting of *log [Intensity]* versus *log[Power]* for all three transitions occurring at 529, 545, and 661 nm is positive, indicating the increase in intensity with the rise in power, as shown in Fig. 13 (b).

The relation between pump power and UC intensity is given by the formula [55]:

$$I_{UC} = P_{in}^{n} \quad (15)$$

Here, n shows the number of photons involved in any UC transition. It helps describe the UC process as a nonlinear phenomenon. The value of the slope '*n*' in Fig. 13 (b) is found using the *log[Intensity (a.u.)]* versus *log[P(W)]*. For all three emissions, the plots are linear with slopes being equal to 0.47, 0.72, and 1.83 for the green (529 & 545 nm) and red (661 nm) emission peaks, respectively, which represents the number of photons (1, 1 and 2, respectively) emitted for the respective transitions.

## 4. Conclusions

In conclusion, the present study focused on the synthesis of $Er^{3+}$-doped SZSi phosphors through a high-temperature solid-state reaction method, examining their structural, morphological, and luminescent properties. XRD analysis confirmed the crystallinity and purity of the as-prepared phosphor samples. The average crystallite size of the SZSi: $Er^{3+}$ phosphors was approximately 84–69 nm, depending on the concentration of $Er^{3+}$ ions. The SEM images revealed that the particles are in the micro-range, exhibiting asymmetric polyhedral shapes with minimal agglomeration. The optical band gap ($E_g$) of 3.17 eV further supports the potential of SZSi: $Er^{3+}$ phosphors in optical applications. The excitation spectra indicate efficient excitation of these phosphors under n-UV light. According to PL emission spectra, the transitions from $^2H_{11/2}$ and $^4S_{3/2}$ to $^4I_{15/2}$ were represented by distinctive peaks with centers at 524 and 545 nm, respectively. The threshold distance for energy transmission, as

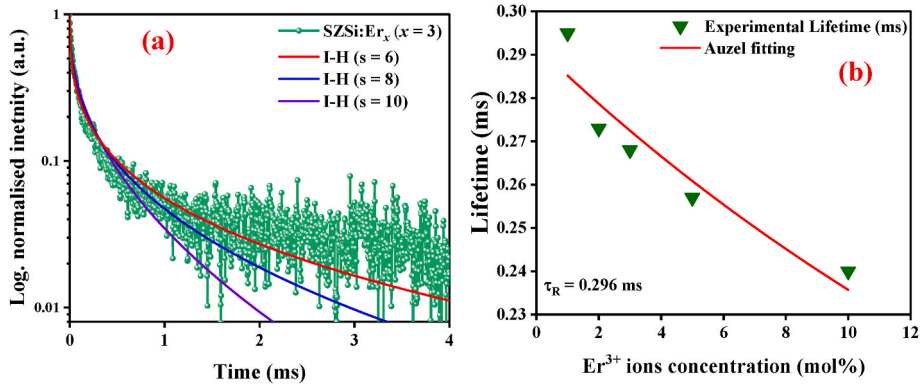



**Fig. 10. (a)** I–H fitted decay profile for the as-synthesized SZSi: $Er_x$ ($x$ = 3 mol%) phosphor for s = 6, 8, and 10 and **(b)** Auzel model's fitted curve.

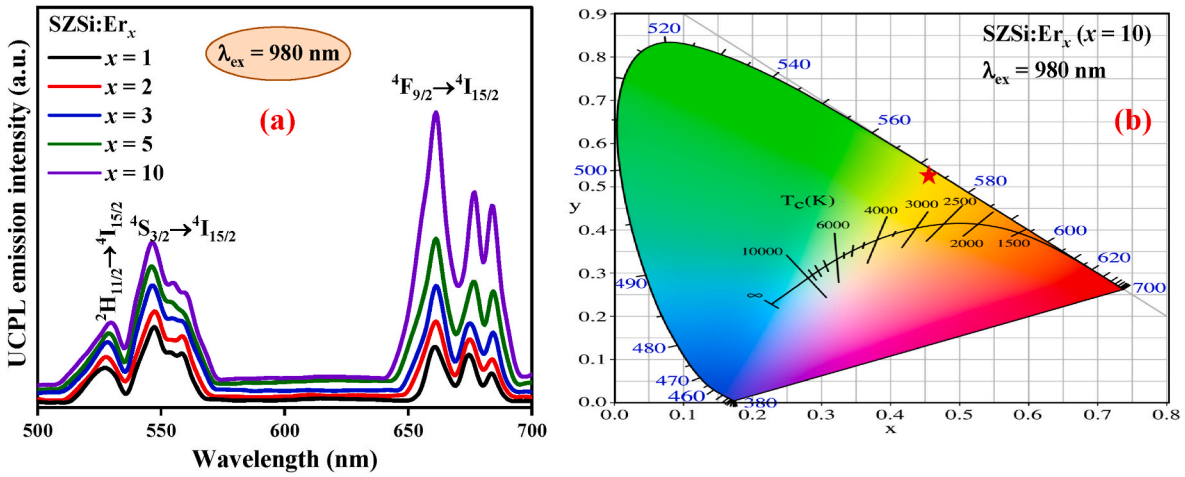



**Fig. 11.** **(a)** The UC emission spectra for SZSi: Er$_x$ ($x$ = 1, 2, 3, 5, and 10 mol%) phosphors under $\lambda_{ex}$ = 980 nm and **(b)** CIE coordinates for SZSi: Er$_x$ (x = 10 mol%) under $\lambda_{ex}$ = 980 nm.

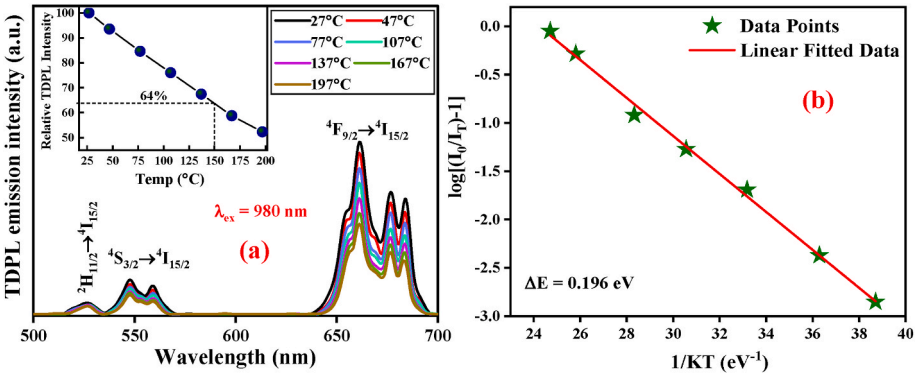



**Fig. 12.** **(a)** PL emission intensity for SZSi: Er$_x$ ($x$ = 10 mol%) at different temperatures under $\lambda_{ex}$ = 980 nm and **(b)** A graph between ln[($I_0$/$I_T$)-1] vs 1/KT.

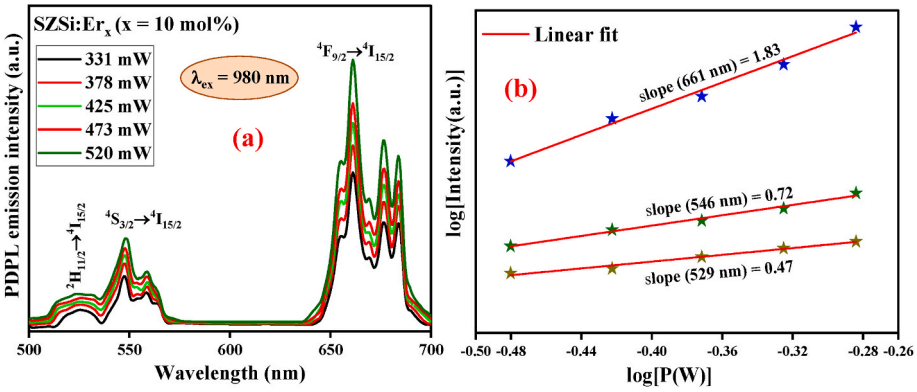



**Fig. 13.** **(a)** Power dependent PL spectra for SZSi: Er$_x$ ($x$ = 10 mol%) phosphor and **(b)** A graph between log [Intensity] versus log [Power].

established by Dexter theory, was 21.92 Å. Dipole-dipole interactions were shown to be the predominant mechanism for energy transfer which was further confirmed using the I–H plot. The decay lifetime (τ) of $Er^{3+}$ ions exhibits a consistent decrease with increasing doping ion concentration. Furthermore, the computed CIE coordinates for the $Er^{3+}$-doped phosphors lie within the green region. The UC luminescence intensity at 661 nm linearly increased when the doping concentration was increased from 1 to 10 mol% of $Er^{3+}$ ions. Under $\lambda_{ex}$ = 980 nm, the evaluated CIE coordinates for SZSi: Er$_x$ ($x$ = 10 mol%) phosphor lie in the yellow-reddish region. Additionally, the phosphors displayed remarkable thermal stability over a temperature range from 27 °C to 197 °C. These findings underscore the potential of $Er^{3+}$-activated $Sr_2ZnSi_2O_7$ phosphors in non-display photonic applications, such as bioimaging, anti-counterfeiting, or infrared-pumped display technologies.

## CRediT authorship contribution statement

**Seema:** Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **A.S. Rao:** Writing – review & editing, Supervision, Resources. **Sarita Baghel:** Writing –

review & editing, Formal analysis.

## Declaration of competing interest

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

## Acknowledgement

The authors are also thankful to the authorities of Delhi Technological University, New Delhi, INDIA, for providing the necessary infrastructure and facilities to carry out the present research work. We are very much thankful to the Department of Chemistry, Dayalbagh Educational Institute, Deemed University, Agra, for availing their SEM facility.

## Data availability

Data will be made available on request.

## References

[1] S. Shu, Y. Wang, Y. Ke, B. Deng, R. Liu, Q. Song, J. Wang, R. Yu, $NaCaTiTaO_6:Sm^{3+}$: a novel orange-red-emitting tantalate phosphor with excellent thermal stability and high color purity for white LEDs, J. Alloys Compd. 848 (2020) 156359, https://doi.org/10.1016/j.jallcom.2020.156359.

[2] L. Zhang, J. Che, Y. Ma, J. Wang, R. Kang, B. Deng, R. Yu, H. Geng, Luminescent and thermal properties of novel orange-red emitting $Ca_2MgTeO_6:Sm^{3+}$ phosphors for white LEDs, J. Lumin. 225 (2020) 117374, https://doi.org/10.1016/j.jlumin.2020.117374.

[3] G. Annadurai, S.M.M. Kennedy, Synthesis and photoluminescence properties of $Ba_2CaZn_2Si_6O_{17}:Eu^{3+}$ red phosphors for white LED applications, J. Lumin. 169 (2016) 690–694, https://doi.org/10.1016/j.jlumin.2015.03.007.

[4] Y. Anantha Lakshmi, K. Swapna, K. Siva Rama Krishna Reddy, M. Venkateswarlu, S. Mahamuda, A.S. Rao, Structural, optical and NIR studies of $Er^{3+}$ ions doped bismuth boro tellurite glasses for luminescence materials applications, J. Lumin. 211 (2019) 39–47, https://doi.org/10.1016/j.jlumin.2019.03.022.

[5] J. Wang, J. Guo, Q. Lv, Z. Ma, X. Feng, Y. Lu, J. Gao, W. Chen, B. Deng, R. Yu, Spectroscopic investigation of the novel orange-red phosphor $Ca_3La_2W_2O_{12}:Sm^{3+}$ with the high color purity for w-LED applications, J. Lumin. 241 (2022) 118459, https://doi.org/10.1016/j.jlumin.2021.118459.

[6] X. Hu, A. Zhang, R. Tang, X. Jin, J. Che, X. Ouyang, C. Zhou, L. Xie, B. Deng, R. Yu, Novel double perovskite reddish-orange emitting $Ca_2InNbO_6$: $Sm^{3+}$ phosphors with abnormal thermal quenching for white LEDs, J. Alloys Compd. 938 (2023) 168540, https://doi.org/10.1016/j.jallcom.2022.168540.

[7] P. Kumar, S. Singh, I. Gupta, A. Hooda, V. Kumar, D. Singh, Reddish-orange color tunable $Sm^{3+}$ activated $Gd_3Al_5O_{12}$ phosphors: crystallographic and photophysical investigation for lighting applications, J. Mol. Struct. 1271 (2023) 134074, https://doi.org/10.1016/j.molstruc.2022.134074.

[8] Y. Liang, H.M. Noh, W. Ran, S.H. Park, B.C. Choi, J.H. Jeong, K.H. Kim, The design and synthesis of new double perovskite (Na,Li)YMg(W,Mo)$O_6$ :$eu^{3+}$ red phosphors for white light-emitting diodes, J. Alloys Compd. 716 (2017) 56–64, https://doi.org/10.1016/j.jallcom.2017.05.027.

[9] Y. Xie, X. Geng, J. Guo, W. Shi, Q. Lv, J. Kong, Y. Li, B. Deng, R. Yu, Luminescence of a novel double-perovskite $Sr_2InSbO_6:Eu^{3+}$ orange-red-emitting phosphor for white LEDs and visualization of latent fingerprints, Mater. Res. Bull. 146 (2022) 111574, https://doi.org/10.1016/j.materresbull.2021.111574.

[10] X. Jin, Y. Xie, R. Tang, X. Geng, J. Che, L. Lou, W. Chen, B. Deng, R. Yu, Novel double perovskite $Sr_3WO_6:Sm^{3+}$, $na^+$ orange-red emitting phosphors with high thermal stability for white LEDs, J. Alloys Compd. 899 (2022) 162739, https://doi.org/10.1016/j.jallcom.2021.162739.

[11] J. Hou, X. Yin, F. Huang, W. Jiang, Synthesis and photoluminescence properties of $NaLaMgWO_6:RE^{3+}$ (RE=Eu, Sm, tb) phosphor for white LED application, Mater. Res. Bull. 47 (2012) 1295–1300, https://doi.org/10.1016/j.materresbull.2012.03.023.

[12] J. Sun, F. Lai, L. Ke, J. Chen, X. Shi, Y. Qiang, W. You, Upconversion luminescence of $La4Ti_9O_{24}$: er–yb phosphor with high green color purity, Opt. Mater. 138 (2023) 113656, https://doi.org/10.1016/j.optmat.2023.113656.

[13] H. Patnam, S.K. Hussain, J.S. Yu, Tunable luminescence and energy transfer behavior of $Ba_3La_6(SiO_4)_6$: $er^{3+}$/$eu^{3+}$ phosphors for solid-state lighting, J. Lumin. 223 (2020) 117204, https://doi.org/10.1016/j.jlumin.2020.117204.

[14] Y. Hua, J.S. Yu, Strong green emission of Erbium(III)-Activated $La_2MgTiO_6$ phosphors for solid-state lighting and optical temperature sensors, ACS Sustain. Chem. Eng. 9 (2021) 5105–5115, https://doi.org/10.1021/acssuschemeng.0c09375.

[15] A.V.S. Yeswanth, I.S. Deo, Anu, A.S. Rao, G.V. Prakash, Down-conversion and up-conversion photoluminescence studies of $Er^{3+}$ doped calcium zinc borosilicate glass for optical fiber amplifiers, J. Non-Cryst. Solids 609 (2023) 122280, https://doi.org/10.1016/j.jnoncrysol.2023.122280.

[16] J. Liao, M. Wang, F. Lin, Z. Han, B. Fu, D. Tu, X. Chen, B. Qiu, H.-R. Wen, Thermally boosted upconversion and downshifting luminescence in $Sc_2(MoO_4)_3$: Yb/Er with two-dimensional negative thermal expansion, Nat. Commun. 13 (2022) 2090, https://doi.org/10.1038/s41467-022-29784-6.

[17] B. Fu, H. Yan, R. Li, L. Feng, Y. Yu, G. Gong, H. Huang, J. Liao, Heterovalent $Zr^{4+}$/$Nb^{5+}$ -cosubstituted negative thermal expansion luminescent materials with anti-thermal quenching luminescence, Laser Photon. Rev. 18 (2024), https://doi.org/10.1002/lpor.202400739.

[18] X. Liu, A. Skripka, Y. Lai, C. Jiang, J. Liu, F. Vetrone, J. Liang, Fast wide-field upconversion luminescence lifetime thermometry enabled by single-shot compressed ultrahigh-speed imaging, Nat. Commun. 12 (2021) 6401, https://doi.org/10.1038/s41467-021-26701-1.

[19] K. R, P. A, S.M. Moses Kennedy, M.I. Sayyed, S.D. Kamath, Optical, thermal, and structural properties of $Er^{3+}$ incorporated double perovskite $Ca_2MgWO_6$ phosphors, J. Lumin. 276 (2024) 120853, https://doi.org/10.1016/j.jlumin.2024.120853.

[20] X. Xu, J. Zhao, H. Feng, Z. Gao, L. Guan, Y. Yang, F. Wang, D. Wang, X. Li, Preparation, luminescence properties and electronic structure of RE (RE = $Eu^{3+}$, $Sm^{3+}$, $Dy^{3+}$)-doped multi-color silico-phosphate $Ca_{15}(PO_4)_2(SiO_4)_6$ phosphors, J. Lumin. 226 (2020) 117356, https://doi.org/10.1016/j.jlumin.2020.117356.

[21] X. Zhang, Y. Chen, L. Zhou, Q. Pang, M. Gong, Synthesis of a broad-band excited and multicolor tunable phosphor $Gd_2SiO_5$ :$ce^{3+}$, $Tb^{3+}$, $Eu^{3+}$ for near-ultraviolet light-emitting diodes, Ind. Eng. Chem. Res. 53 (2014) 6694–6698, https://doi.org/10.1021/ie404312n.

[22] F. Auzel, Upconversion processes in coupled ion systems, J. Lumin. 45 (1990) 341–345, https://doi.org/10.1016/0022-2313(90)90189-I.

[23] P. Du, L. Luo, J.S. Yu, Upconversion emission, cathodoluminescence and temperature sensing behaviors of $Yb^{3+}$ ions sensitized $NaY(WO_4)_2:Er^{3+}$phosphors, Ceram. Int. 42 (2016) 5635–5641, https://doi.org/10.1016/j.ceramint.2015.12.083.

[24] Seema Anu, A. Kumar, N. Deopa, M.K. Sahu, A. Prasad, A.S. Rao, A single phase $Li_2Ba_5W_3O_{15}:Dy^{3+}$/$Eu^{3+}$ phosphor for color tunable devices and non-contact optical thermometry, J. Lumin. 269 (2024) 120444, https://doi.org/10.1016/j.jlumin.2024.120444.

[25] S. Jana, A. Mondal, J. Manam, S. Das, $Pr^{3+}$ doped $BaNb_2O_6$ reddish Orange emitting phosphor for solid state lighting and optical thermometry applications, J. Alloys Compd. 821 (2020) 153342, https://doi.org/10.1016/j.jallcom.2019.153342.

[26] Y. Niu, F. Wu, Q. Zhang, Y. Teng, Y. Huang, Z. Yang, Z. Mu, Luminescence and thermometry sensing of $Sr_2InTaO_6$: $eu^{3+}$, $Mn^{4+}$ phosphors in a wide temperature range, J. Lumin. 275 (2024) 120748, https://doi.org/10.1016/j.jlumin.2024.120748.

[27] S. Zhao, Z. Mu, L. Lou, S. Yuan, M. Liao, Q. Lin, D. Zhu, F. Wu, Broadening and enhancing emission of $Cr^{3+}$ simultaneously by co-doping Yb3+ in $Ga_{1.4}In_{0.6}SnO_5$, J. Rare Earths 41 (2023) 1895–1903, https://doi.org/10.1016/j.jre.2022.10.012.

[28] R. Lohan, A. Kumar, M.K. Sahu, A. Mor, V. Kumar, N. Deopa, 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. 34 (2023) 694, https://doi.org/10.1007/s10854-023-10055-z.

[29] Y. Shi, R. Cui, X. Liu, L. Yang, Y. Li, J. Zhang, C. Deng, $Ca_2YNbO_6$: $dy^{3+}$, $Sm^{3+}$/$Eu^{3+}$ warm white phosphors with tunable light color and FIR optical temperature measurement characteristics, Opt. Mater. 136 (2023), https://doi.org/10.1016/j.optmat.2023.113495.

[30] L. Shi, Y. Han, H. Wang, D. Shi, X. Geng, Z. Zhang, High-efficiency and thermally stable far-red emission of $Mn^{4+}$ in double cubic perovskite $Sr_9Y_2W_4O_{24}$ for plant cultivation, J. Lumin. 208 (2019) 307–312, https://doi.org/10.1016/j.jlumin.2018.12.065.

[31] Anu, A.S. Rao, Temperature sensing materials based on the fluorescence intensity ratio in $Li_2Ba_5W_3O_{15}:Dy^{3+}$ phosphors, Sensors Actuators A Phys 372 (2024) 115336, https://doi.org/10.1016/j.sna.2024.115336.

[32] W.T. Carnall, P.R. Fields, 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. 49 (1968) 4424–4442, https://doi.org/10.1063/1.1669893.

[33] Anu, A.S. Rao, Synthesis, structural and fluorescence investigations of novel - $Li_2Ba_5W_3O_{15}$ : $Sm^{3+}$ phosphors for photonic device applications, J. Fluoresc. (2023), https://doi.org/10.1007/s10895-023-03449-z.

[34] Q. Lin, Q. Wang, M. Liao, M. Xiong, X. Feng, X. Zhang, H. Dong, D. Zhu, F. Wu, Z. Mu, Trivalent chromium ions doped fluorides with both broad emission bandwidth and excellent luminescence thermal stability, ACS Appl. Mater. Interfaces 13 (2021) 18274–18282, https://doi.org/10.1021/acsami.1c01417.

[35] N. Deopa Anu, A.S. Rao, Structural and luminescence characteristics of thermally stable $Dy^{3+}$ doped oxyfluoride strontium zinc borosilicate glasses for photonic device applications, Opt. Laser Technol. 154 (2022) 108328, https://doi.org/10.1016/j.optlastec.2022.108328.

[36] A.S. Seema, Rao,Photoluminescence and energy transfer studies in the $Sm^{3+}$ and $Eu^{3+}$ co-doped $Sr_2ZnSi_2O_7$ red-emitting phosphors, J. Lumin. 275 (2024) 120742, https://doi.org/10.1016/j.jlumin.2024.120742.

[37] G. Seeta Rama Raju, J.Y. Park, H.C. Jung, E. Pavitra, B.K. Moon, J.H. Jeong, J. S. Yu, J.H. Kim, H. Choi, Blue and green emissions with high color purity from nanocrystalline $Ca_2Gd_8Si_6O_{26}$:Ln (ln = Tm or er) phosphors, J. Alloys Compd. 509 (2011) 7537–7542, https://doi.org/10.1016/j.jallcom.2011.04.122.

[38] L.G. Van Uitert, Characterization of energy transfer interactions between rare Earth ions, J. Electrochem. Soc. 114 (1967) 1048, https://doi.org/10.1149/1.2424184.

[39] A.S. Anu, $Pr^{3+}$ phosphors, Opt. Mater. 145 (2023) 114476, https://doi.org/10.1016/j.optmat.2023.114476.
[40] L. Yang, Z. Mu, S. Zhang, Q. Wang, D. Zhu, Y. Zhao, D. Luo, Q. Zhang, F. Wu, $Dy^{3+}$ doped $Ca_9Gd(PO_4)_7$: a novel single-phase full-color emitting phosphor, J. Mater. Sci. Mater. Electron. 29 (2018) 6548–6555, https://doi.org/10.1007/s10854-018-8638-7.
[41] C.B.A. Devi, K. Swapna, S. Mahamuda, M. Venkateswarlu, M.V.V.K.S. Prasad, K. Siva Rama Krishna Reddy, N. Deopa, A.S. Rao, Spectroscopic studies and lasing potentialities of $Sm^{3+}$ ions doped single alkali and mixed alkali fluoro tungstentellurite glasses, Opt. Laser Technol. 111 (2019) 176–183, https://doi.org/10.1016/j.optlastec.2018.09.051.
[42] H. Li, R. Zhao, Y. Jia, W. Sun, J. Fu, L. Jiang, S. Zhang, R. Pang, C. Li, $Sr_{1.7}Zn_{0.3}CeO_4$ : $eu^{3+}$ novel red-emitting phosphors: synthesis and photoluminescence properties, ACS Appl. Mater. Interfaces 6 (2014) 3163–3169, https://doi.org/10.1021/am4041493.
[43] A. Kumar Anu, Ravina, N. Deopa Ravina, A.S. Rao, Efficient tunable photoluminescence of $Dy^{3+}/Eu^{3+}$ co-doped OFSZBS glasses for warm white LEDs, J. Non-Cryst. Solids 616 (2023) 122421, https://doi.org/10.1016/j.jnoncrysol.2023.122421.
[44] G. Yadav, J. Parashar, Anu, A. Prasad, A.S. Rao, Synthesis, structural and photoluminescence characteristics of $Dy^{3+}/Eu^{3+}$ co-doped $CaGdSbWO_8$ phosphor for warm white LEDs, Emergent Mater (2024), https://doi.org/10.1007/s42247-024-00960-2.
[45] A.S. Anu, Rao,Novel orange-red-emitting $Li_2Ba_5(WO_5)_3:Eu^{3+}$ phosphor for w-LEDs and non-contact thermometry applications, J. Mater. Sci. Mater. Electron. 34 (2023) 2191, https://doi.org/10.1007/s10854-023-11602-4.
[46] P.R. Rani, M. Venkateswarlu, K. Swapna, S. Mahamuda, M.V.V.K.S. Prasad, A. S. Rao, Spectroscopic and luminescence properties of $Ho^{3+}$ ions doped barium lead alumino fluoro borate glasses for green laser applications, Solid State Sci. 102 (2020) 106175, https://doi.org/10.1016/j.solidstatesciences.2020.106175.
[47] S. Mahamuda, K. Swapna, M. Venkateswarlu, A. Srinivasa Rao, S. Shakya, G. Vijaya Prakash, Spectral characterisation of $Sm^{3+}$ ions doped Oxy-fluoroborate glasses for visible Orange luminescent applications, J. Lumin. 154 (2014) 410–424, https://doi.org/10.1016/j.jlumin.2014.05.017.
[48] B.C. Jamalaiah, M.V. Vijaya Kumar, K. Rama Gopal, Fluorescence properties and energy transfer mechanism of $Sm^{3+}$ ion in lead telluroborate glasses, Opt. Mater. 33 (2011) 1643–1647, https://doi.org/10.1016/j.optmat.2011.04.030.
[49] M. Inokuti, F. Hirayama, Influence of energy transfer by the exchange mechanism on donor luminescence, J. Chem. Phys. 43 (1965) 1978–1989, https://doi.org/10.1063/1.1697063.
[50] J.A. Jiménez, Temperature dependent spectroscopic properties of $cu^{+}$ and $Dy^{3+}$ co-doped phosphate glass: band gap analysis and Cu nanocluster-enhanced $Dy^{3+}$ luminescence, ChemPhysChem 20 (2019) 399–404, https://doi.org/10.1002/cphc.201800971.
[51] S. Loos, F. Steudel, B. Ahrens, S. Schweizer, Temperature-dependent luminescence and energy transfer properties of $Tb^{3+}$ and $Eu^{3+}$ doped barium borate glasses, J. Lumin. 181 (2017) 31–35, https://doi.org/10.1016/j.jlumin.2016.08.066.
[52] Y. Zhuo, F. Wu, Y. Niu, Y. Wang, Q. Zhang, Y. Teng, H. Dong, Z. Mu, Super broadband emission across NIR-I and NIR-II under blue light excitation of $Cr^{3+}$, $Ni^{2+}$ co-doped $Sr_2GaTaO_6$ phosphor achieved by two-site occupation and effective energy transfer, Laser Photon. Rev. 18 (2024), https://doi.org/10.1002/lpor.202400105.
[53] P. Kumar, D. Singh, I. Gupta, H. Kumar, Synthesis, crystallographic structure, Down shifting luminescence of Er(III) activated $GdSr_2AlO_5$ nanophosphors: an efficient green emitter for solid state lighting, Mater. Sci. Semicond. Process. 167 (2023) 107765, https://doi.org/10.1016/j.mssp.2023.107765.
[54] J. Li, Q. Hu, Y. Su, Q. Li, Z. Li, J. Xie, B. Wang, H. Jiang, H. Zhang, J. Zhu, Microwave-hydrothermal synthesis, near ultraviolet/infrared-excited green fluorescence and ratiometric thermometry of $Er^{3+}$-activated iodate phosphors, Ceram. Int. 51 (2025) 1032–1041, https://doi.org/10.1016/j.ceramint.2024.11.086.
[55] A. Prasad, A.S. Rao, G.V. Prakash, A study on up-conversion and energy transfer kinetics of $KGdF_4$ : $Yb^{3+}/Er^{3+}$ nanophosphors, J. Mol. Struct. 1205 (2020) 127647, https://doi.org/10.1016/j.molstruc.2019.127647.